Seeing Failure In Advance: Pre-Generation Failure Diagnosis for Multi-Modal Retrieval-Augmented-Generation

Chan-Wei Hu[1], **Shuo Xing**[1], **Ryan A. Rossi**[2], **Zhengzhong Tu**[1]
[1]Texas A&M University
[2]Adobe Research
{huchanwei123, shuoxing, tzz}@tamu.edu
ryrossi@adobe.com

Abstract

Multimodal retrieval-augmented generation (MMRAG) systems can fail even when relevant evidence is retrieved, exposing a gap between retrieval and generation. We formulate MMRAG failure diagnosis as an online single-fire decision problem over prefill hidden states, which is distinct from static probing of model internals or post-hoc self-correction. Across InfoSeek, E-VQA, and VISA, we show that hidden states encode both failure occurrence and failure type, with failure detection reaching around 90% accuracy and recall on InfoSeek and E-VQA. We further introduce a Fire-Then-Type prober that separates failure detection from failure typing, improving typed diagnosis over direct classification. Finally, our method matches the response accuracy of final-layer repair on InfoSeek and E-VQA while diagnosing before answer decoding. The available prefill depth varies substantially across models, from 4.6% to 24.3%, and is limited for Qwen3-VL and InternVL on VISA.

1 Introduction

MMRAG has become a standard paradigm for grounding Multi-Modal Large Language Models (MLLMs) in external knowledge, particularly for knowledge-intensive visual question answering where parametric memory is insufficient or stale (Lewis et al., 2020; Lin and Byrne, 2022; Chen et al., 2022; Wu et al., 2026; Hu et al., 2025; Hu and Tu, 2026; Xing et al., 2025). By augmenting an image-question pair with retrieved textual or visual evidence, MMRAG systems aim to produce factual answers without retraining the underlying MLLMs.

Despite steady progress in retrievers and rerankers (Chen et al., 2022; Hu et al., 2023; Zhang et al., 2025c), a persistent puzzle remains. *Improvements in retrieval quality do not translate proportionally into generation gains.* Even when relevant evidence is successfully retrieved and placed in context, MLLMs frequently fail to leverage it (Liu et al., 2024; Jin et al., 2025; Cuconasu et al., 2024). This retrieval-generation gap has been variously attributed to positional bias toward early context (Liu et al., 2024; Yao et al., 2025), knowledge conflicts between parametric and retrieved evidence (Xu et al., 2024; Zhang et al., 2025b), underutilization of fine-grained details, and sensitivity to noise from irrelevant candidates (Jin et al., 2025; Cuconasu et al., 2024). Yet the failure *mechanism*, and crucially *when* during inference the failure becomes detectable, remains poorly characterized for multimodal settings.

Existing remedies for this gap are largely *reactive*. Self-correction frameworks such as Self-Refine (Madaan et al., 2023), Reflexion (Shinn et al., 2023), and Self-RAG (Asai et al., 2024), together with corrective retrieval methods including CRAG (Yan et al., 2024), OmniSearch (Li et al., 2025), and DeepMMSearch-R1 (Narayan et al., 2025), share a common architectural assumption. They wait for the model to emit a final answer before evaluating its quality and triggering correction. This post-hoc paradigm incurs the full cost of a discarded forward pass and applies a generic retry regardless of whether the underlying issue is a distracting hard negative, an unparseable evidence chunk, or simply missing evidence. Attention-level interventions such as ALFAR (An et al., 2025) act earlier but are constructed under an oracle assumption of top-1 correct evidence, a condition that rarely holds in deployed retrieval pipelines.

In this work, we ask a different question. ***Does the MLLM internally know when it is about to fail, before any output token is generated?*** As an initial motivation, Figure 2 compares hidden states produced under two controlled conditions, *oracle*, where only gold evidence is provided, and *mixed*, where gold evidence is combined with hard negatives. Across three MLLMs, successful cases remain highly similar between the two conditions, whereas other failures show a clear representational divergence at deeper layers. These patterns indicate that MMRAG failures are not only visible in final

Code is availabe at https://github.com/taco-group/MMRAG-failure-diagnose

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 12959–12977
October 24-29, 2026 ©2026 Association for Computational Linguistics

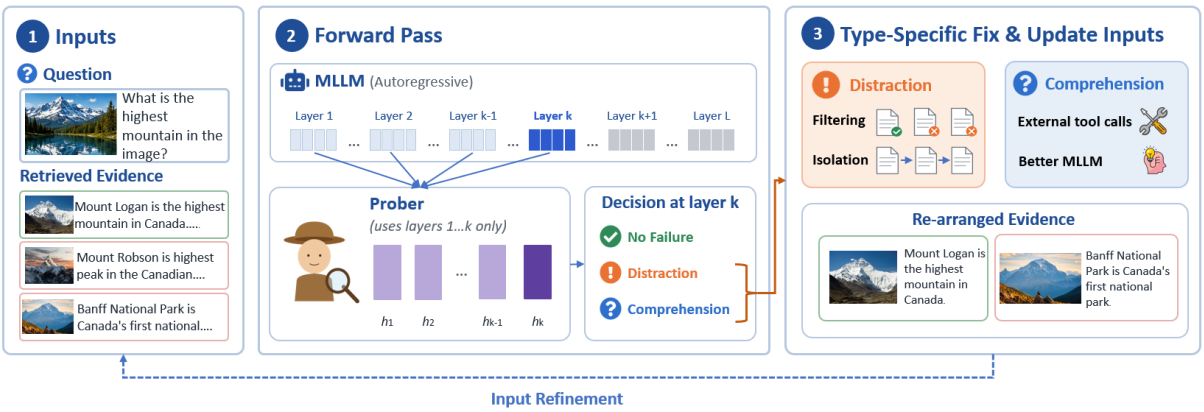


Figure 1: Overview of the proposed early failure routing framework for MMRAG. Given an image-question pair and retrieved evidence, the MLLM performs a forward pass while a lightweight prober monitors hidden-state prefixes at intermediate layers. When the prober fires a failure, it routes the example to a type-specific intervention and refines inputs.

answers, but are already reflected in the model's internal computation. Drawing on recent text-domain findings that LLM hidden representations cluster systematically by retrieval condition (Yeh and Li, 2026) and that lightweight probes can predict reasoning correctness mid-trajectory (Zhang et al., 2025a), we analyze the hidden states of contemporary MLLMs under controlled retrieval conditions and find that a usable failure signal emerges at intermediate layers, well before generation completes.

Building on this observation, we propose a lightweight framework, shown in Figure 1, that monitors hidden states during the MLLM forward pass, predicts failure before decoding, and selects an intervention based on the diagnosed failure type. The prober is a lightweight prefix Transformer with separate failure-detection and failure-type heads. Selective repair avoids changing examples that are not diagnosed as DISTRACTION while retaining the accuracy gain of context repair.

This work raises four research questions. We first ask whether a lightweight prober on MLLMs' hidden states can detect failures before generation completes. We then examine whether early failure routing improves end-to-end MMRAG accuracy over uniform retry. Next, we investigate at which pre-generation transformer depths failure signals emerge, and how the earliness-accuracy trade-off affects routing decisions. Finally, we examine how consistently failure signals appear across different MLLM architectures.

Contributions. Our contributions are threefold.

1. A novel formulation of MMRAG failure diagnosis as an online single-fire early-exit problem, distinct from static hidden-state probing and from post-hoc self-correction. A comparison of our work with others is shown in Table 8 in Appendix.

2. A probing-based early failure routing framework that distinguishes two intervention-relevant failure types and dispatches type-specific interventions, in contrast to the uniform-retry assumption of prior self-correction work.

3. An empirical study of where and how reliably failure signals emerge across layers and MLLM architectures.

2 Related Works

We summarize how our framework relates to the closest lines of research in Table 8 in Appendix.

Multimodal retrieval-augmented generation. Retrieval is especially important for knowledge-intensive visual question answering, where answers often require external knowledge beyond the image. Knowledge-based VQA benchmarks and models study how systems retrieve and use textual or structured knowledge for visual questions (Wang et al., 2018; Marino et al., 2019; Gui et al., 2022; Yang et al., 2022). More recent multimodal RAG methods retrieve outside knowledge for VQA (Lin and Byrne, 2022), improve fine-grained cross-modal retrieval (Lin et al., 2023), or retrieve from multimodal image-text memories (Chen et al., 2022; Hu et al., 2023). Dynamic MMRAG systems further incorporate multi-hop retrieval planning and web search tools (Li et al., 2025; Long et al., 2025; Narayan et al., 2025). Unlike these works, which primarily improve retrieval or retrieval planning, we diagnose whether the MLLM will successfully use the retrieved evidence after it has already been placed in context.

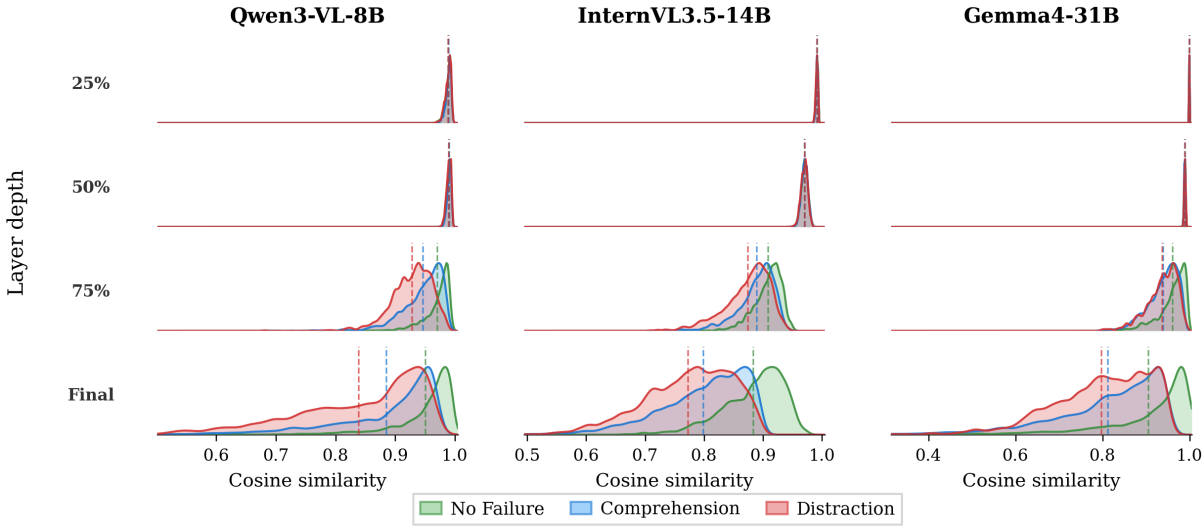


Figure 2: Distribution of cosine similarity between hidden states produced under two retrieval conditions, oracle (gold evidence only) and mixed (gold + hard-negative evidence), across layer depth for three MLLMs on InfoSeek training set. Samples are grouped into three types based on model response correctness. **No Failure**: the model answers correctly under both oracle and mixed conditions. **Comprehension**: the model remains incorrect even when gold evidence is provided. **Distraction**: the model succeeds with oracle retrieval but fails when hard negatives are added. Dashed lines indicate group means.

Corrective and self-reflective generation. Several methods improve generation through inference-time feedback or correction. Self-Refine iteratively critiques and revises an initial answer (Madaan et al., 2023), Reflexion uses verbal feedback from past failures (Shinn et al., 2023), Self-RAG trains models to retrieve and critique with reflection tokens (Asai et al., 2024), and CRAG estimates retrieval quality to trigger corrective retrieval (Yan et al., 2024). In multimodal settings, OmniSearch and DeepMMSearch-R1 use adaptive planning or tool-based search to improve complex visual question answering (Li et al., 2025; Narayan et al., 2025). These approaches are largely driven by retrieval-quality estimates, generated critiques, or post-hoc retry. In contrast, we detect impending failure from hidden states before answer generation completes and route the example to a type-specific intervention.

Knowledge utilization and hidden-state probing. Even with relevant evidence, MLLMs may fail due to weak evidence utilization, attention bias, or conflicts between parametric and retrieved knowledge. ALFAR addresses this issue by reallocating attention toward relevant context tokens and adaptively fusing parametric and contextual knowledge (An et al., 2025). More broadly, probing studies show that intermediate representations encode useful information about model behavior (Alain and Bengio, 2017; Belinkov, 2022), including uncertainty, latent knowledge, truthfulness, and reasoning correctness (Kadavath et al., 2022; Burns et al., 2023; Azaria and Mitchell, 2023; Zhang et al., 2025a). In the text-based RAG setting, Yeh and Li (2026) show that retrieved context systematically shapes LLM hidden representations. Our work extends this perspective to MMRAG by probing MLLM hidden states to predict both whether an execution will fail and what type of failure it reflects.

3 Analysis

3.1 Experiment Setup

We train the diagnostic prober using the training splits of InfoSeek (Chen et al., 2023), E-VQA (Mensink et al., 2023) and VISA-paper (Ma et al., 2025). For each example, we run three MLLMs, which are Qwen3-VL-8B (Bai et al., 2025), InternVL3.5-14B (Wang et al., 2025) and Gemma4-31B (Team et al., 2026), under two retrieved-context conditions. The first is oracle condition, where the model is given the target evidence document. The second is a mixed condition, where the relevant evidence is presented together with retrieved distractor documents. For both conditions, we record the model's pre-generation hidden states at every transformer layer.

For failure detection, each execution is labeled by whether the MLLM answer under that execution is correct (Sec. 3.3). For failure typing, we compare paired oracle and mixed executions. We label an example as *Comprehension* when the model is incorrect in both conditions and as *Distraction* when it is correct with oracle evidence but incorrect with mixed evidence. All paired executions

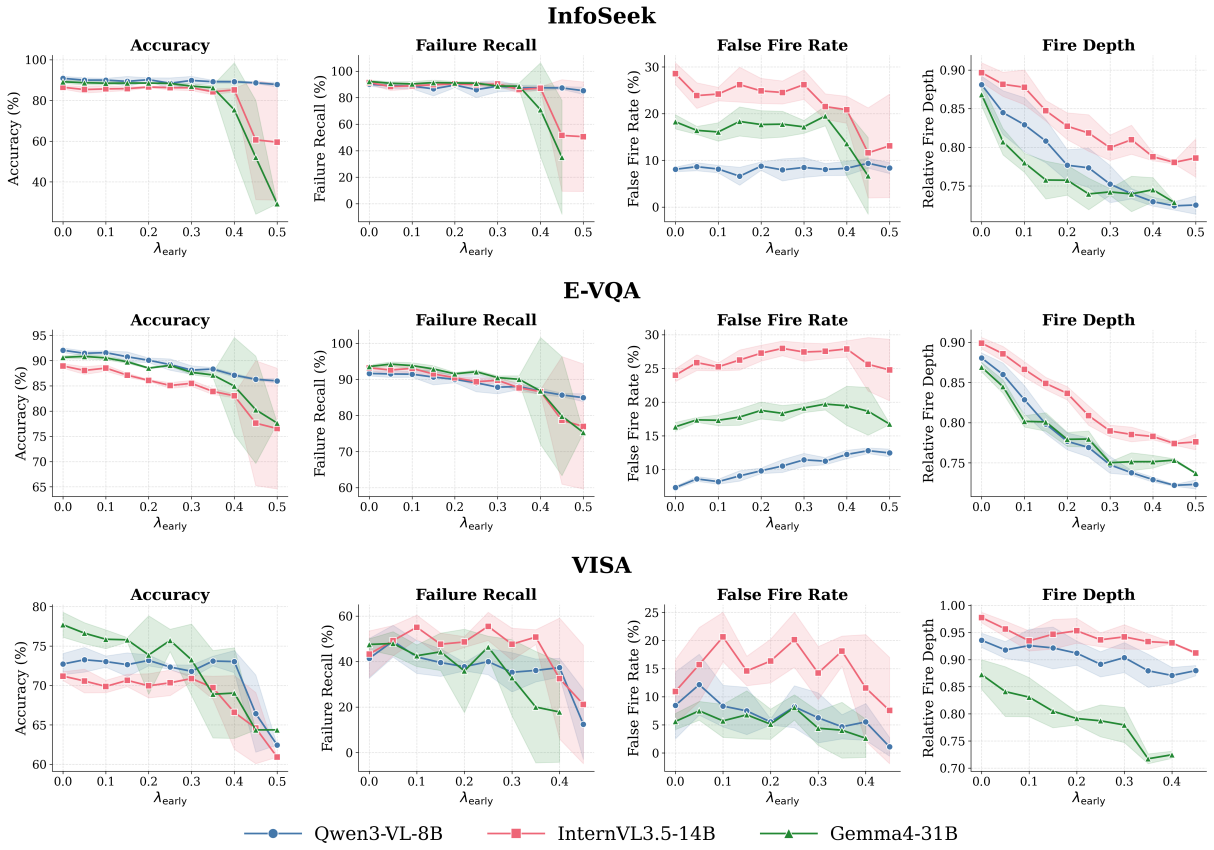


Figure 3: **Failure detection from hidden-state prefixes on three datasets.** We vary the earliness weight λ_{early} to study when failure signals become detectable and how early firing affects detection quality. Higher λ_{early} penalizes late firing more strongly, encouraging the prober to fire earlier when a reliable failure signal is present. Raw values are present in Table 9, 10, 11 in Appendix.

use deterministic greedy decoding at temperature 0. The oracle condition is used only to create training and evaluation labels. At inference, the prober sees only hidden states from the mixed context and receives no gold evidence. More experimental details are shown in Appendix B.

3.2 When does the failure start to emerge and become detectable?

Our first question is whether MMRAG failures can be detected from the model's internal states before answer generation is completed. We formulate failure detection as an online single-fire decision problem over hidden states, inspired by early-exit and adaptive-computation methods that make intermediate-layer decisions when sufficient evidence is available (Graves, 2016; Teerapittayanon et al., 2016). Given a retrieved-context input, the MLLM processes the prompt through its transformer layers before decoding an answer. At layer k, the prober observes the hidden-state prefix

$$H_{1:k} = [h_1, h_2, \ldots, h_k],$$

where h_k denotes the hidden state at layer k for the current retrieved-context execution. The prober then decides whether to continue processing or to fire a failure signal. Once the prober fires, the decision is final. At each layer k, the prober outputs a conditional firing probability

$$q_k = P(\text{fire at layer } k \mid \text{no earlier fire}, H_{1:k}).$$

This quantity is local to layer k. It represents the probability of firing at the current layer, conditioned on the fact that the prober has not fired at any previous layer. Since the prober is constrained to fire at most once, the probability that it fires for the first time exactly at layer k must also account for not firing at all earlier layers. We denote this first-fire probability by p_k:

$$p_k = q_k \prod_{j<k} (1 - q_j).$$

Thus, q_k is the prober's conditional decision probability at layer k, whereas p_k is the induced probability that the first and only firing event occurs at layer k. For a model with L layers, the probability of never firing is

$$P_{\text{never fire}} = \prod_{k=1}^{L} (1 - q_k),$$

and the probability of firing at least once is

$$P_{\text{fire}} = 1 - P_{\text{never fire}} = \sum_{k=1}^{L} p_k.$$

In our experiments, we instantiate retrieved-context executions under two retrieval conditions: an oracle-only context, which contains only gold evidence, and a mixed context, which contains both gold evidence and hard-negative evidence. Both settings are treated as retrieved-context executions, allowing us to examine whether failure signals emerge under clean evidence as well as in the presence of distractors.

Given a binary label $y \in \{0, 1\}$, where $y = 1$ denotes failure, i.e., the final model answer is incorrect, and $y = 0$ denotes no failure, i.e., the final model answer is correct, we train the prober with the sequence-level binary negative log-likelihood

$$\mathcal{L}_{\text{seq}} = -y \log P_{\text{fire}} - (1 - y) \log P_{\text{never fire}}.$$

This objective directly matches the online single-fire evaluation setting. Failure examples are rewarded if the prober fires at least once, while no-failure examples are rewarded only if the prober never fires.

However, correctness alone is not sufficient. A prober that waits until the final layer may be accurate, but it leaves little room for downstream intervention. Conversely, a prober that fires too aggressively in early layers may interrupt executions that would have eventually produced correct answers. Therefore, the prober should fire only when the failure signal is reliable, while still firing early enough to be useful.

To encourage timely decisions, we introduce an earliness cost on failure examples. Conditioned on the prober firing, the expected relative firing depth is

$$d_{\text{fire}} = \frac{1}{P_{\text{fire}}} \sum_{k=1}^{L} \frac{k}{L} p_k.$$

The final objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seq}} + \lambda_{\text{early}} y d_{\text{fire}}.$$

Here, λ_{early} controls how strongly late firing is penalized. When $\lambda_{\text{early}} = 0$, the prober optimizes only the sequence-level single-fire objective. Increasing λ_{early} encourages the prober to fire earlier on failure examples, but only when the hidden states provide sufficient evidence to preserve the sequence-level objective.

Figure 3 reports the earliness behavior of the proposed prober across the three datasets and MLLMs. We vary λ_{early}, which controls the cost of late firing, and examine how the prober trades off overall accuracy, failure recall, false-fire rate, and firing depth. Diagnosis occurs before the final transformer layer in every evaluated setting, but the point at which a reliable signal emerges is model- and dataset-dependent. In particular, the VISA results indicate later and less stable failure detection than InfoSeek and E-VQA.

Table 1 shows that the accuracy drop at large λ_{early} is largely driven by reduced fire rates. Rather than firing earlier, the prober often becomes overly conservative and chooses not to fire, which lowers false fires but also misses true failures. This collapse is most evident for Gemma 4 on InfoSeek and all three models on VISA at $\lambda_{\text{early}} = 0.5$. Thus, moderate earliness weights encourage useful earlier intervention, whereas overly large weights reduce failure coverage.

The failure signal is also weaker and less stable on VISA than on InfoSeek and E-VQA. VISA evidence consists of dense scientific-document screenshots with substantial textual, visual, and layout information. This makes grounding the question in the evidence more difficult. The filtered three-candidate setting may also make oracle and mixed hidden states less separable. We therefore do not claim uniform performance across evidence formats.

Key Takeaway

Failure signals emerge before answer decoding in all evaluated settings, but the usable prefill window varies considerably across datasets and architectures and is limited in two VISA settings.

3.3 Does the model internally encode the type of failure?

Our second question is whether hidden states encode the failure type as well as failure occurrence. This distinction matters for repair. A DISTRACTION failure motivates evidence filtering or reranking, while a COMPREHENSION failure is unlikely to be fixed by removing distractors.

We extend the prober with a type prediction head in addition to the fire head described in Section 3.2. The fire head decides when to interrupt the current execution, while the type head predicts the reason for failure once the fire head fires. Thus, the type head is conditioned on the fire head.

Prober architecture. At each transformer depth, we extract the final prefill-token hidden state and project it into a shared 256-dimensional space. We

Dataset	Model	λ_{early}										
		0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
InfoSeek	Qwen	$57.5_{\pm 1.7}$	$56.8_{\pm 1.9}$	$56.2_{\pm 1.6}$	$54.4_{\pm 4.4}$	$56.9_{\pm 1.6}$	$54.7_{\pm 5.1}$	$56.1_{\pm 3.5}$	$55.9_{\pm 2.8}$	$55.5_{\pm 1.9}$	$55.9_{\pm 1.2}$	$54.4_{\pm 2.6}$
	InternVL	$75.9_{\pm 1.6}$	$72.4_{\pm 3.2}$	$72.9_{\pm 2.3}$	$74.1_{\pm 2.8}$	$74.2_{\pm 1.3}$	$73.7_{\pm 2.0}$	$74.5_{\pm 1.8}$	$70.2_{\pm 2.9}$	$70.8_{\pm 1.5}$	$41.7_{\pm 38.1}$	$41.3_{\pm 37.8}$
	Gemma	$70.6_{\pm 1.1}$	$69.1_{\pm 0.8}$	$68.4_{\pm 2.5}$	$70.0_{\pm 2.4}$	$69.6_{\pm 1.9}$	$69.5_{\pm 1.6}$	$67.8_{\pm 1.0}$	$68.4_{\pm 2.6}$	$54.1_{\pm 30.2}$	$26.7_{\pm 36.6}$	0.0
E-VQA	Qwen	$57.9_{\pm 1.4}$	$58.1_{\pm 0.8}$	$57.3_{\pm 1.0}$	$58.3_{\pm 2.1}$	$57.4_{\pm 0.9}$	$57.5_{\pm 2.2}$	$56.8_{\pm 1.4}$	$57.6_{\pm 1.6}$	$56.6_{\pm 0.8}$	$56.2_{\pm 0.6}$	$56.2_{\pm 1.8}$
	InternVL	$76.1_{\pm 0.7}$	$76.1_{\pm 1.4}$	$76.3_{\pm 1.0}$	$75.4_{\pm 1.2}$	$74.8_{\pm 0.6}$	$74.2_{\pm 1.1}$	$74.3_{\pm 0.8}$	$72.8_{\pm 1.2}$	$72.1_{\pm 0.7}$	$65.6_{\pm 16.0}$	$64.1_{\pm 15.8}$
	Gemma	$70.9_{\pm 0.4}$	$71.7_{\pm 0.4}$	$71.2_{\pm 1.3}$	$70.9_{\pm 1.1}$	$70.2_{\pm 0.7}$	$70.5_{\pm 0.7}$	$69.6_{\pm 0.4}$	$69.4_{\pm 1.1}$	$67.1_{\pm 12.6}$	$61.9_{\pm 14.2}$	$58.1_{\pm 0.1}$
VISA	Qwen	$20.8_{\pm 7.7}$	$26.0_{\pm 6.5}$	$21.0_{\pm 5.6}$	$19.5_{\pm 5.2}$	$17.6_{\pm 0.9}$	$20.2_{\pm 5.0}$	$17.2_{\pm 6.7}$	$16.5_{\pm 3.6}$	$17.5_{\pm 3.7}$	$5.4_{\pm 7.5}$	0.0
	InternVL	$23.6_{\pm 8.3}$	$28.8_{\pm 7.4}$	$34.1_{\pm 5.3}$	$27.5_{\pm 3.8}$	$29.0_{\pm 3.6}$	$33.9_{\pm 5.8}$	$27.3_{\pm 6.3}$	$30.9_{\pm 2.8}$	$19.8_{\pm 18.1}$	$12.9_{\pm 17.8}$	0.0
	Gemma	$20.5_{\pm 1.0}$	$21.9_{\pm 3.0}$	$18.9_{\pm 3.8}$	$20.2_{\pm 6.3}$	$16.1_{\pm 9.1}$	$21.8_{\pm 3.1}$	$14.6_{\pm 8.5}$	$9.7_{\pm 13.3}$	$8.0_{\pm 11.2}$	0.0	0.0

Table 1: Overall fire rate (%) on InfoSeek, E-VQA and VISA test set across earliness weights λ_{early}. Each cell is $\text{mean}_{\pm \text{std}}$ across five seeds.

Model	Dataset	Fail vs. NF AUROC	Comp. vs. Dist. AUROC
Qwen	InfoSeek	0.945 ± 0.002	0.929 ± 0.003
	E-VQA	0.924 ± 0.011	0.911 ± 0.007
	VISA	0.761 ± 0.021	0.654 ± 0.016
InternVL	InfoSeek	0.928 ± 0.005	0.913 ± 0.004
	E-VQA	0.907 ± 0.013	0.886 ± 0.010
	VISA	0.784 ± 0.010	0.652 ± 0.040
Gemma	InfoSeek	0.907 ± 0.002	0.874 ± 0.005
	E-VQA	0.880 ± 0.012	0.856 ± 0.024
	VISA	0.772 ± 0.035	0.616 ± 0.068

Table 2: Threshold-independent AUROC at $\lambda_{\text{early}} = 0.2$. Values are mean $\pm$ standard deviation over five seeds. NF, Comp., and Dist. denote NO FAILURE, COMPREHENSION, and DISTRACTION.

add a learned layer embedding and process the resulting hidden-state prefix with a two-layer Transformer encoder. A fire head predicts whether the prober should fire at the current depth. Once it fires, a separate type head classifies the execution as COMPREHENSION or DISTRACTION. The prober is trained while the MLLM remains frozen. Full hyperparameters and implementation details are provided in Appendix B. We assign type labels by comparing the model's behavior under paired oracle and mixed retrieval conditions. Let $r_o, r_m \in \{0, 1\}$ denote answer correctness under oracle-only and mixed evidence, respectively. We define three labels:

$$(r_o, r_m) = (1, 1) \Rightarrow \text{NO FAILURE},$$

$$(r_o, r_m) = (0, 0) \Rightarrow \text{COMPREHENSION},$$

$$(r_o, r_m) = (1, 0) \Rightarrow \text{DISTRACTION}.$$

The NO FAILURE label covers executions that remain correct after hard negatives are added. COMPREHENSION covers failures that persist with gold evidence without determining the underlying cause. DISTRACTION captures errors that appear after hard negatives are added. It can reflect semantic interference or context ordering. These categories are derived by changing only the retrieved context for the same image and question, and they support different interventions.

Since the type head is used only after the fire head fires, we train it under the induced first-fire distribution. For an example with type label c, $c \in \{\text{COMPREHENSION}, \text{DISTRACTION}\}$, the type prediction conditioned on firing is

$$P(c \mid \text{fire}) = \frac{1}{P_{\text{fire}}} \sum_{k=1}^{L} p_k \, P(c \mid H_{1:k}).$$

The corresponding type loss is

$$\mathcal{L}_{\text{type}} = -\mathbb{I}[c \neq \text{NO FAILURE}] \log P(c \mid \text{fire}).$$

We apply this loss only to actionable failure types, since NO FAILURE examples should be handled by the fire head not firing rather than by assigning a type after firing.

The final training objective adds type loss to the single-fire objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seq}} + \lambda_{\text{early}} y d_{\text{fire}} + \mathcal{L}_{\text{type}}.$$

This formulation ensures that the prober first learns whether and when to fire, and the type head learns to classify the failure type at the layers where firing is likely.

We compare the proposed Fire-Then-Type (FTT) prober against a Direct 3-Class (D3C) baseline. FTT is the fire-conditioned model described above. The fire head first determines whether and when to interrupt the execution, and the type head predicts the failure type under the induced first-fire distribution. D3C removes this decomposition and directly predicts NO FAILURE, COMPREHENSION, or DISTRACTION from the hidden-state prefix.

Figure 4 compares the two methods across λ_{early} on three metrics. Full per-lambda results are reported in Table 12, 13, 14 in Appendix. Our FTT improves overall accuracy over D3C across the all the MLLMs. The advantage is clearest for Qwen3-VL and InternVL, where FTT also substantially improves DISTRACTION recall while preserving high COMPREHENSION recall. This supports that separating failure detection from failure typing makes

the diagnostic signal easier to extract from hidden states.

Gemma shows a more mixed pattern. FTT improves overall accuracy and COMPREHENSION recall, but does not improve DISTRACTION recall over D3C. Thus, the fire-then-type decomposition is beneficial for typed diagnosis overall, but its benefit for identifying distraction failures is model-dependent.

Threshold-independent reliability. At $\lambda_{\text{early}} = 0.2$, failure-vs.-no-failure AUROC ranges from 0.880 to 0.945 on InfoSeek and E-VQA, and COMPREHENSION-vs.-DISTRACTION AUROC ranges from 0.856 to 0.929. Both tasks are weaker on VISA, especially type classification, where AUROC is between 0.616 and 0.654. This confirms that the learned signal is not equally reliable across evidence formats. Table 2 reports all five-seed results.

Zero-shot cross-dataset transfer. Within each MLLM, we train on one dataset and directly evaluate on the other two. Detector transfer between InfoSeek and E-VQA reduces accuracy by 2.8 to 4.8 points. Transfer from either Wikipedia-derived dataset to VISA loses 6.5 to 8.1 points. Probers trained on VISA lose 1.8 to 2.5 points when transferred to InfoSeek and E-VQA. FTT shows the same asymmetry. We therefore characterize the result as partial cross-dataset generalization rather than domain invariance. Complete detector and FTT results are in Tables 5 and 6. Cross-model transfer is not evaluated because each input projection is tied to its backbone and independently trained hidden coordinates are not aligned.

Key Takeaway

Prefill hidden states encode failure occurrence and distinguish COMPREHENSION from DISTRACTION, with partial rather than domain-invariant cross-dataset transfer.

3.4 Does DISTRACTION failure intervention improve answer generation?

Our final question asks whether failure diagnosis can support downstream repair. We focus on DISTRACTION, where the model answers correctly with oracle evidence but fails with mixed evidence. This behavior directly motivates context removal. COMPREHENSION failures persist even with oracle evidence and therefore call for different strategies. We study context-level repair for DISTRACTION and leave abstention, re-retrieval, or reasoning support for COMPREHENSION to future work.

Our repair policy uses hidden states from the prober's firing layer, i.e., the layer at which the prober commits to a DISTRACTION prediction. This keeps the intervention online-causal. The evidence to remove is selected using only representations available at the time of diagnosis. For each retrieved evidence item, we pool hidden states over its tokens and compare this representation with the final-token representation at the same layer to obtain a model-focus score. We then remove the document with the largest mismatch between model focus and retrieval relevance, corresponding to evidence that receives disproportionately high internal focus relative to its retriever score.

We compare against five controls. TOP-1 RETRIEVAL is prober-free and generates from only the retriever's highest-scoring document. ALWAYS APPLY removes the highest-mismatch document for every example using final-layer hidden states. RANDOM REMOVAL and LOW RELEVANCE use the same prober-triggered routing as our method but remove a random or the lowest-relevance document. ORACLE (UB) provides only gold evidence and is an upper bound for distractor-removal repair. It also exposes failures that persist with gold evidence and cannot be addressed by our context-removal intervention.

Table 3 reports mean accuracy and standard deviation over five seeds with $\lambda_{\text{early}} = 0.2$. OURS consistently improves over both the original retrieved context and the prober-free TOP-1 RETRIEVAL baseline. Random and low-relevance removal do not reliably improve over the original context, indicating that the hidden-state mismatch captures a signal beyond retriever rank alone.

The comparison with ALWAYS APPLY clarifies the role of routing. OURS achieves comparable accuracy in every setting while applying repair only after a DISTRACTION prediction. Full end-to-end measurements show 13% to 36% lower latency than ALWAYS APPLY. The reduction is at least 15% in eight of nine settings. Peak memory remains similar to ALWAYS APPLY (Table 7).

The earliness benefit varies by dataset. On InfoSeek and E-VQA, OURS leaves between 16.3% and 24.3% of the prefill depth. On VISA, Qwen3-VL and InternVL leave only 8.9% and 4.6%, respectively. These cases demonstrate pre-generation routing but limited prefill savings.

Finally, the gap between our method and the oracle upper bound shows both the promise and the limitation of context-level repair. Across models, the strategy recovers a substantial portion of the achievable improvement, but it does not fully

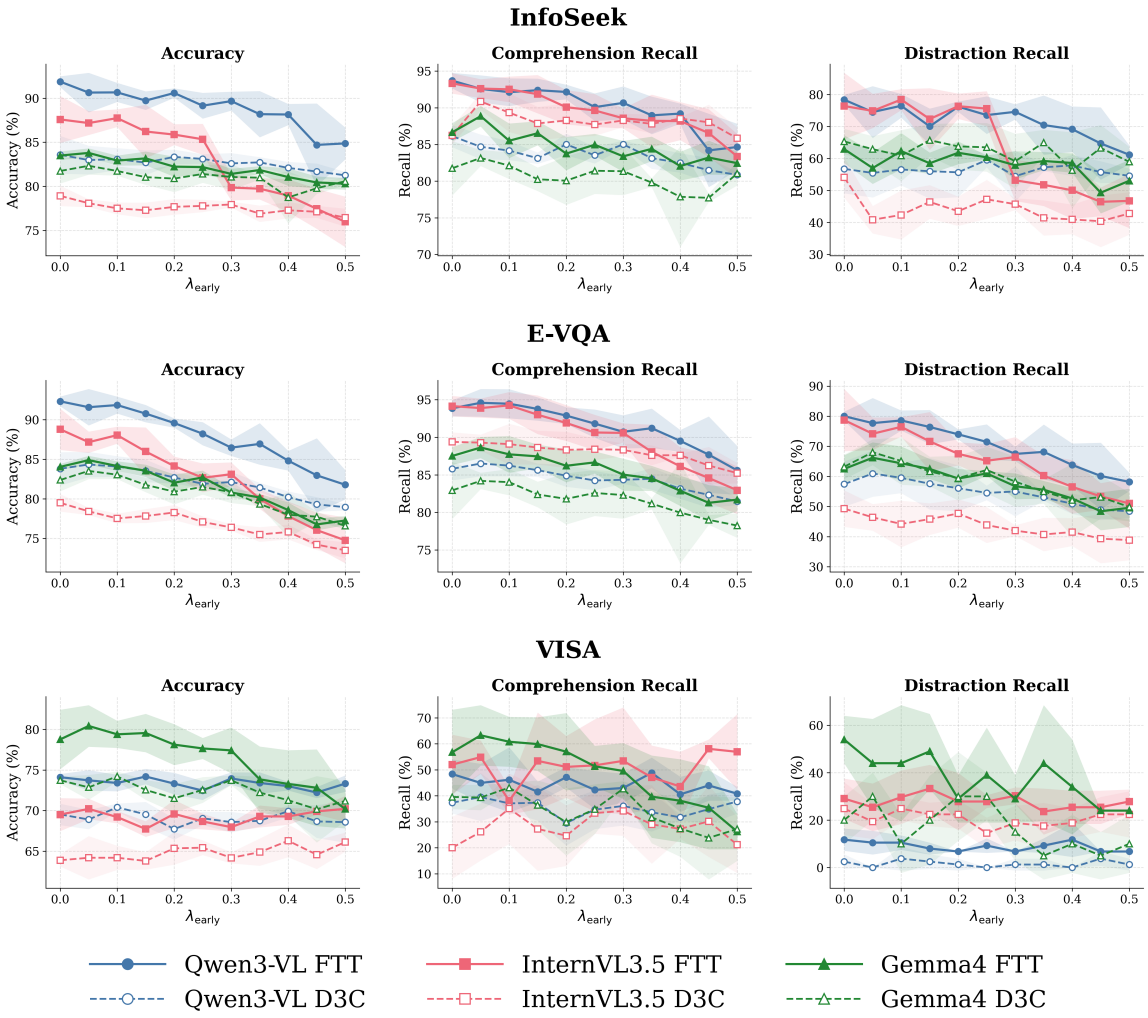


Figure 4: **Failure type prediction from hidden-state prefixes.** We compare the proposed Fire-Then-Type (FTT) prober with a direct 3-class (D3C) baseline across different earliness weights. FTT more reliably separates failure detection from failure typing, yielding stronger overall diagnosis and improved recall for the two failure types. Raw values are present in Table 12, 13, 14.

close the gap to the oracle. This suggests that failure diagnosis can be used not only for post-hoc analysis, but as an actionable signal for improving generation-time behavior.

Key Takeaway

Pre-generation diagnosis enables selective context repair with accuracy comparable to always applying repair and 13% to 36% lower end-to-end latency, although the prefill window is limited in two VISA settings.

4 Conclusion

We investigated whether MMRAG failures can be diagnosed before answer generation. By probing MLLM hidden states under controlled retrieval conditions, we show that pre-generation representations encode both failure occurrence and failure type. Our early failure routing framework separates DISTRACTION from COMPREHENSION failures, with routing decisions emerging during the pre-generation forward pass at intermediate-to-late prefill depths. For DISTRACTION failures, prober-guided repair improves answer accuracy by suppressing hard negatives. Our results show that failure diagnosis is possible before answer generation across all evaluated settings, while the point at which a reliable signal emerges varies considerably across datasets and architectures.

Limitations

This work studies pre-generation failure diagnosis under paired oracle and mixed retrieval conditions. The oracle condition is used only offline to construct labels for failures that appear after hard negatives are added and failures that persist with gold evidence. At inference, the prober predicts

Dataset	Model	No fix	Top-1 Ret.	Always	Random	Low Rel.	Oracle (UB)	OURS	Avg. layer when failure detected	Prefill saved
InfoSeek	Qwen3-VL-8B	38.1±.08	41.3±.14	42.5±.21	39.2±.09	39.7±.12	**43.8±.07**	42.7±.18	28.7 / 37	22.4%
	InternVL3.5-14B	21.6±.11	24.1±.19	25.4±.24	22.5±.13	22.1±.08	**27.3±.06**	25.3±.22	33.9 / 41	17.3%
	Gemma4-31B	28.4±.04	28.8±.10	29.7±.17	29.2±.15	29.7±.09	**30.9±.05**	30.1±.20	46.2 / 61	24.3%
E-VQA	Qwen3-VL-8B	54.9±.16	60.8±.28	65.7±.33	54.1±.18	58.6±.21	**69.8±.12**	66.1±.29	28.7 / 37	22.4%
	InternVL3.5-14B	28.7±.20	31.1±.30	36.2±.41	28.7±.25	31.0±.17	**51.4±.14**	36.0±.35	34.3 / 41	16.3%
	Gemma4-31B	45.2±.07	47.7±.13	51.4±.23	46.2±.19	47.9±.11	**54.2±.06**	51.5±.27	47.5 / 61	22.1%
VISA	Qwen3-VL-8B	63.8±.03	64.0±.08	64.1±.12	62.0±.10	64.1±.09	**65.9±.04**	64.7±.15	33.7 / 37	8.9%
	InternVL3.5-14B	45.0±.18	48.7±.26	50.6±.38	46.4±.22	49.8±.16	**55.2±.11**	51.1±.31	39.1 / 41	4.6%
	Gemma4-31B	61.7±.05	62.2±.09	62.8±.14	60.9±.13	62.1±.07	**66.1±.03**	63.2±.17	48.3 / 61	20.8%

Table 3: **Prober-triggered context repair results on the three datasets.** Values are mean accuracy ± standard deviation over five seeds (percentage). TOP-1 RETRIEVAL is a prober-free baseline that retains only the retriever's highest-scoring document. ALWAYS applies the removal policy to all samples, while RANDOM and LOW REL. use prober-triggered routing but remove a random or lowest-ranked document. ORACLE (UB) uses only gold evidence. Avg. layer is the mean firing layer over five seeds, followed by the model's total layer count. Prefill saved is the remaining fraction before the final layer. Best results are bold and second-best results are underlined. Firing layers use $\lambda_{\text{early}} = 0.2$.

from hidden states of the mixed context alone. This setup does not cover deployment cases where gold evidence is absent, partially relevant, or ambiguous. The COMPREHENSION and DISTRACTION labels do not form a complete taxonomy of MMRAG failure causes. Failures that persist with gold evidence can include visual-understanding, OCR or layout, reasoning, answer-scoring, parametric-knowledge conflict, and prompt-sensitivity errors. Failures that appear with mixed context can arise from semantic interference or context-order effects. Missing or insufficient evidence is excluded because gold evidence is always injected. Separating these causes would require targeted annotations and controlled perturbations. Our repair module is intended as a proof-of-concept rather than a complete intervention strategy. We focus on DISTRACTION failures and use a simple hidden-state-guided context removal policy to show that early diagnosis can support targeted generation-time repair. Future work can explore stronger interventions, such as adaptive re-retrieval, evidence rewriting, or reasoning support for COMPREHENSION failures. Finally, the diagnostic signal varies across datasets and models, with weaker results on VISA than on InfoSeek and E-VQA. This suggests that diagnosability may depend on dataset structure, evidence granularity, and retrieval-candidate quality. False negatives may allow incorrect answers to reach users without intervention, while false positives may trigger unnecessary context removal and damage otherwise correct answers. The operating threshold should therefore reflect the relative cost of these errors. This is especially important in medical, legal, and financial applications. Our benchmark results do not establish safety for these domains, where deployment would require domain-specific calibration, abstention or human review for uncertain cases, and external validation. A separate prober is trained for each MLLM. Its input projection is tied to the backbone's hidden dimensionality and layer structure, and hidden coordinates are not aligned across independently trained models. Cross-model transfer would therefore require explicit representation alignment or a model-agnostic projection method.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations Workshop*.

Wenbin An, Jiahao Nie, Feng Tian, Haonan Lin, mingxiang cai, Yaqiang Wu, QianYing Wang, Xiaoqin Zhang, and Shijian Lu. 2025. Boosting knowledge utilization in multimodal large language models via adaptive logits fusion and attention reallocation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*.

Wenhu Chen, Hexiang Hu, Xi Chen, Patrick Verga, and William W. Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570.

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

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Alex Graves. 2016. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968.

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

Chan-Wei Hu, Yueqi Wang, Shuo Xing, Chia-Ju Chen, Suofei Feng, Ryan Rossi, and Zhengzhong Tu. 2025. mrag: Elucidating the design space of multi-modal retrieval-augmented generation. *arXiv preprint arXiv:2505.24073*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. 2023. REVEAL: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23369–23379.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan Arik. 2025. Long-context llms meet rag: Overcoming challenges for long inputs in rag. In *International Conference on Learning Representations*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, HUI WANG, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and Philip Yu. 2025. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. In *International Conference on Learning Representations*, volume 2025, pages 95582–95604.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xinwei Long, Zhiyuan Ma, Ermo Hua, Kaiyan Zhang, Biqing Qi, and Bowen Zhou. 2025. Retrieval-augmented visual question answering via built-in autoregressive search engines. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. 2025. Visa: Retrieval augmented generation with visual source attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30154–30169.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder,

Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594.

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

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

Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M. Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. 2025. DeepMMSearch-R1: Empowering multimodal LLMs in multimodal web search. *arXiv preprint arXiv:2510.12801*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in neural information processing systems*, volume 36, pages 8634–8652.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, and 1 others. 2026. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*.

Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. 2016. BranchyNet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition*, pages 2464–2469.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. FVQA: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2413–2427.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Keshu Wu, Chenchen Kuai, Zihao Li, Jiwan Jiang, Shiyu Shen, Shian Wang, Chan-Wei Hu, Zhengzhong Tu, and Yang Zhou. 2026. Knowledge is not static: Order-aware hypergraph rag for language models. *arXiv preprint arXiv:2604.12185*.

Shuo Xing, Peiran Li, Yuping Wang, Ruizheng Bai, Yueqi Wang, Chan-Wei Hu, Chengxuan Qian, Huaxiu Yao, and Zhengzhong Tu. 2025. Re-align: Aligning vision language models via retrieval-augmented direct preference optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2379–2397.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Jiayu Yao, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Yuyao Ge, Zhecheng Li, and Xueqi Cheng. 2025. Who is in the spotlight: The hidden bias undermining multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15183–15193, Suzhou, China. Association for Computational Linguistics.

Samuel Yeh and Sharon Li. 2026. How retrieved context shapes internal representations in RAG. *arXiv preprint arXiv:2602.20091*.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025a. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025b. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025c. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Details of Dataset

We evaluate early failure diagnosis on three knowledge-intensive multimodal QA datasets: InfoSeek, E-VQA, and VISA. Table 4 summarizes the dataset statistics and the retrieval evidence source used in our experiments.

Dataset	Evidence source	KB / candidates	Train	Test / Eval
InfoSeek	Wikipedia	100K docs	934K	73K val
E-VQA	Wikipedia	2M pages	221K	5.7K test
VISA-paper	Paper screenshots	-	71K usable / 100K total	2.16K

Table 4: Dataset statistics used in our experiments. For InfoSeek, we evaluate on the validation split because the public test answers are not available. For VISA, we use the Paper-VISA subset rather than Wiki-VISA, since InfoSeek and E-VQA already provide Wikipedia/encyclopedic evidence settings.

InfoSeek. InfoSeek is a visual information-seeking QA benchmark designed for questions that cannot be answered from the image or common-sense knowledge alone (Chen et al., 2023). The dataset contains about 1.3M image-question-answer triplets over roughly 11K visual entities, with approximately 934K training examples, 73K validation examples, and 348K test examples. Since the public test labels are not available, we evaluate on the validation split. In our MMRAG setup, InfoSeek is paired with a 100K-document Wikipedia knowledge base. For each query, the oracle condition contains the target evidence document, while the mixed condition combines the target evidence with hard negatives retrieved from this knowledge base.

E-VQA. E-VQA, or Encyclopedic-VQA, is a large-scale visual question answering benchmark focused on detailed properties of fine-grained categories and instances (Mensink et al., 2023). It contains 221K unique question-answer pairs, each paired with up to five images, resulting in about 1M image-question-answer triplets. The dataset is supported by a controlled Wikipedia-derived knowledge base with about 2M pages. We use E-VQA as a Wikipedia-grounded MMRAG setting and follow the dataset's prescribed answer evaluation protocol.

VISA-paper. VISA is a retrieval-augmented generation benchmark with visual source attribution, where models must answer a question and identify supporting evidence in retrieved document screenshots (Ma et al., 2025). VISA contains multiple subsets, including Wiki-VISA, based on Wikipedia webpage screenshots, and Paper-VISA, derived from PubLayNet and focused on paper-style scientific documents. Since InfoSeek and E-VQA already cover Wikipedia or encyclopedic evidence, we use the Paper-VISA subset as a complementary non-Wikipedia setting. Paper-VISA contains 100K training examples and 3,000 test examples in the original release. In our experiments, after filtering to examples with three retrieved candidates, we use 71,010 training examples and 2,160 test examples. The mixed condition is constructed from the target evidence screenshot and the provided candidate distractors.

B Experimental Protocol

B.1 Prober Architecture

All probers for binary failure detection, Fire-Then-Type (FTT), and Direct 3-Class (D3C) share a single backbone architecture and differ only in their classification heads and training objective. The input is the hidden state prefix until layer-k, $H_{1:k} = [h_1, h_2, \ldots, h_k]$. $k \leq N$, where N is the depth of the underlying MLLM, namely embedding plus transformer layers ($N = 37$ for Qwen3-VL-8B, $N = 41$ for InternVL3.5-14B, and $N = 61$ for Gemma4-31B). We linearly project to $d_{\text{model}} = 256$, add a learnable per-layer positional embedding, prepend a learnable [CLS] token, and run a 2-layer transformer encoder with $n_{\text{heads}} = 8$, FFN dimension $4 \times d_{\text{model}}$, GELU, and dropout $= 0.3$, followed by LayerNorm.

To realise online prediction, at each candidate fire layer $k \in \{1, \ldots, N\}$ we re-run the encoder with a key-padding mask that zeroes out positions larger than k, read the [CLS] state, and apply the relevant classification head.

The fire head (used in RQ1 and FTT) is an MLP mapping from $d \to 128 \to 1$, producing $q_k = P(\text{fire at } k \mid \neg\text{fired before})$. The type head in FTT is an MLP mapping from $d \to 128 \to 2$ over {COMP, DIST}, read out at the fired layer. The D3C is an MLP mapping from $d \to 128 \to 3$) over {NF, COMP, DIST}, with no separate firing step. For FTT we use the *soft-type* loss, in which the type cross-entropy at layer k is weighted by q_k, making the type head answerable wherever the hazard fires. All three variants add an earliness penalty $\lambda_{\text{early}}\, \mathbb{E}[k_{\text{fire}}]/N$ on failure examples, and the sweep over λ_{early} traces the recall-depth Pareto frontier.

B.2 Answer Evaluation in Section 3.4

We apply each dataset's prescribed scorer uniformly across the two execution conditions (Oracle and Mixed). For **InfoSeek** we use the upstream STRING / NUMERICAL / TIME exact-match scorer with the dataset's normalization (Chen et al., 2023). For **E-VQA** we use the upstream encyclopedic-VQA scorer (Mensink et al., 2023), which applies pipe-separated alternative EM

for single-answer questions and set-IoU ≥ 0.5 for `multi_answer` questions. A BEM (Bulian et al., 2022) fallback for paraphrase equivalence is wired in but disabled on our hosts, so reported numbers are EM-only. For **VISA** we apply the E-VQA scorer to the `short_answer` field.

B.3 Hard-Negative Construction

For each query, the mixed condition is built by shuffling the oracle evidence with $K-1$ distractors drawn from the source dataset's top-K retrieval candidates. "Hard" here is in the retrieval-similarity sense, not adversarially mined. For **InfoSeek** and **E-VQA**, we use the top-5 candidates from an EVA-CLIP retriever over the Wikipedia knowledge base. For **VISA**, we use the queries that contains three candidates ($K = 3$) shipped with the source release leaving 71,010 out of 100,000 usable samples for training, and 2,160 out of 3,000 samples for testing.

B.4 Hyperparameter Selection

We train each prober for up to 1000 epochs with AdamW with lr $= 1 \times 10^{-4}$, weight decay $= 1 \times 10^{-2}$ and batch size 256. We also apply early stopping on validation accuracy with patience $= 10$. We do not tune λ_{early}. We sweep the entire grid $\lambda_{\text{early}} \in \{0.00, 0.05, \ldots, 0.50\}$. We report mean and standard deviation over five random seeds numbered 42 to 46. Training for each model, dataset, and prober combination runs on a single NVIDIA A6000 in under 10 hours for Qwen3-VL-8B and InternVL3.5-14B. Gemma4-31B takes $\sim$20 hours owing to its 61-layer hidden-state input.

All paired oracle and mixed executions use deterministic greedy decoding with temperature $= 0$. We will release the implementation and dataset-construction details with the public paper.

C Licenses

We use publicly released datasets, models, and software resources in accordance with their original licenses and intended research use. We do not redistribute third-party datasets, model checkpoints, or external knowledge bases.

Datasets. InfoSeek (Chen et al., 2023) is released under the Apache-2.0 license. E-VQA (Mensink et al., 2023) is released through the Google Research repository, where datasets are released under CC BY 4.0 and source code under Apache-2.0. For VISA (Ma et al., 2025), we use the released Paper-VISA subset following the terms of the original benchmark release. Since the public paper/release page does not clearly specify a separate dataset license, we do not redistribute VISA data and only use it for research evaluation

Models and software. We evaluate publicly released MLLMs, including Qwen3-VL-8B (Bai et al., 2025), InternVL3.5-14B (Wang et al., 2025), and Gemma4-31B (Team et al., 2026), under Apache-2.0. Our implementation builds on standard open-source libraries, including PyTorch and Hugging Face Transformers, under their corresponding licenses.

D Additional Analyses

Threshold-independent AUROC supports the same pattern as the thresholded results. On InfoSeek and E-VQA, failure-vs.-no-failure AUROC ranges from 0.880 to 0.945, while type AUROC ranges from 0.856 to 0.929. On VISA, failure detection AUROC is between 0.761 and 0.784, while type AUROC is between 0.616 and 0.654. The lower VISA values show that the weaker result is not caused only by the selected decision threshold.

Transfer between InfoSeek and E-VQA is relatively strong, while transfer from either dataset to VISA is more difficult. This is consistent with VISA's distinct scientific-document screenshots and weaker in-domain separability. VISA-trained probers transfer comparatively well to InfoSeek and E-VQA. Cross-model transfer is not directly meaningful here because input dimensions, layer structures, and hidden coordinates differ. It would require explicit representation alignment or a model-agnostic projection.

FTT transfer follows the same direction as detector transfer. Between InfoSeek and E-VQA, transferred FTT accuracy decreases by 3.3 to 4.8 points. Transfer from either dataset to VISA decreases accuracy by 7.1 to 9.1 points. In the reverse direction, VISA-trained FTT probers lose 2.1 to 2.8 points on InfoSeek and E-VQA. Both failure-type recalls also decline after transfer. These results support partial cross-dataset reuse, but not dataset-invariant diagnosis.

The end-to-end measurements also confirm the benefit of selective repair. Across all nine dataset and model settings, OURS stays within 0.6 accuracy points of ALWAYS APPLY while reducing latency by 13% to 36%. Latency falls by 22% to 29% on InfoSeek, 23% to 36% on E-VQA, and 13% to 25% on VISA. The reduction reaches at least 15% in eight settings. Peak memory differs from ALWAYS APPLY by at most 0.3 GiB. NO FIX remains the least expensive configuration because it does not perform repair.

The TOP-1 RETRIEVAL baseline improves over NO FIX in all nine settings, showing that a smaller context can itself be helpful. It remains below

OURS in all nine settings. The full gain therefore cannot be explained only by retaining the highest-ranked document. Diagnosis and hidden-state-guided removal provide additional value beyond retriever rank.

Source	Target	Model	Accuracy	Failure recall	Relative fire depth
InfoSeek	E-VQA	Qwen3-VL-8B	90.1 → 85.3 (−4.8)	90.0 → 84.4 (−5.6)	.777 → .783 (+.006)
		InternVL3.5-14B	86.1 → 83.0 (−3.1)	90.5 → 82.3 (−8.2)	.837 → .845 (+.008)
		Gemma4-31B	88.5 → 84.1 (−4.4)	91.5 → 87.3 (−4.2)	.779 → .805 (+.026)
	VISA	Qwen3-VL-8B	73.2 → 66.1 (−7.1)	37.8 → 29.4 (−8.4)	.912 → .915 (+.003)
		InternVL3.5-14B	70.0 → 63.5 (−6.5)	48.6 → 41.3 (−7.3)	.953 → .960 (+.007)
		Gemma4-31B	73.9 → 66.2 (−7.7)	35.9 → 28.8 (−7.1)	.791 → .816 (+.025)
E-VQA	InfoSeek	Qwen3-VL-8B	90.2 → 86.1 (−4.1)	89.5 → 83.0 (−6.5)	.777 → .785 (+.008)
		InternVL3.5-14B	86.6 → 83.8 (−2.8)	90.4 → 83.5 (−6.9)	.827 → .838 (+.011)
		Gemma4-31B	88.5 → 84.7 (−3.8)	91.1 → 86.5 (−4.6)	.757 → .781 (+.024)
	VISA	Qwen3-VL-8B	73.2 → 65.7 (−7.5)	37.8 → 28.5 (−9.3)	.912 → .918 (+.006)
		InternVL3.5-14B	70.0 → 63.0 (−7.0)	48.6 → 40.1 (−8.5)	.953 → .962 (+.009)
		Gemma4-31B	73.9 → 65.8 (−8.1)	35.9 → 27.5 (−8.4)	.791 → .821 (+.030)
VISA	InfoSeek	Qwen3-VL-8B	90.2 → 87.8 (−2.4)	89.5 → 86.2 (−3.3)	.777 → .771 (−.006)
		InternVL3.5-14B	86.6 → 84.5 (−2.1)	90.4 → 87.0 (−3.4)	.827 → .823 (−.004)
		Gemma4-31B	88.5 → 86.0 (−2.5)	91.1 → 87.8 (−3.3)	.757 → .753 (−.004)
	E-VQA	Qwen3-VL-8B	90.1 → 87.9 (−2.2)	90.0 → 86.9 (−3.1)	.777 → .781 (+.004)
		InternVL3.5-14B	86.1 → 84.3 (−1.8)	90.5 → 87.6 (−2.9)	.837 → .831 (−.006)
		Gemma4-31B	88.5 → 86.2 (−2.3)	91.5 → 88.1 (−3.4)	.779 → .782 (+.003)

Table 5: Zero-shot cross-dataset transfer for failure detection. Each entry is in-domain → transferred performance, with the change in parentheses. A separate prober is retained for each MLLM.

Source	Target	Model	Accuracy	Comp. recall	Dist. recall
InfoSeek	E-VQA	Qwen3-VL-8B	89.6 → 84.9 (−4.7)	92.9 → 86.8 (−6.1)	74.0 → 71.4 (−2.6)
		InternVL3.5-14B	84.1 → 80.6 (−3.5)	91.9 → 83.4 (−8.5)	67.5 → 64.6 (−2.9)
		Gemma4-31B	82.0 → 77.2 (−4.8)	86.2 → 81.1 (−5.1)	59.3 → 56.8 (−2.5)
	VISA	Qwen3-VL-8B	73.3 → 65.6 (−7.7)	47.1 → 38.2 (−8.9)	6.8 → 3.9 (−2.9)
		InternVL3.5-14B	69.6 → 62.5 (−7.1)	51.1 → 42.8 (−8.3)	27.8 → 19.4 (−8.4)
		Gemma4-31B	78.1 → 69.7 (−8.4)	57.0 → 48.5 (−8.5)	29.0 → 20.9 (−8.1)
E-VQA	InfoSeek	Qwen3-VL-8B	90.6 → 86.1 (−4.5)	92.2 → 85.3 (−6.9)	76.2 → 73.7 (−2.5)
		InternVL3.5-14B	85.9 → 82.6 (−3.3)	90.1 → 82.7 (−7.4)	76.3 → 73.5 (−2.8)
		Gemma4-31B	82.2 → 78.0 (−4.2)	83.8 → 78.6 (−5.2)	61.8 → 59.2 (−2.6)
	VISA	Qwen3-VL-8B	73.3 → 65.1 (−8.2)	47.1 → 37.5 (−9.6)	6.8 → 3.4 (−3.4)
		InternVL3.5-14B	69.6 → 61.9 (−7.7)	51.1 → 41.5 (−9.6)	27.8 → 18.1 (−9.7)
		Gemma4-31B	78.1 → 69.0 (−9.1)	57.0 → 47.2 (−9.8)	29.0 → 19.2 (−9.8)
VISA	InfoSeek	Qwen3-VL-8B	90.6 → 87.9 (−2.7)	92.2 → 88.3 (−3.9)	76.2 → 70.8 (−5.4)
		InternVL3.5-14B	85.9 → 83.5 (−2.4)	90.1 → 86.1 (−4.0)	76.3 → 70.5 (−5.8)
		Gemma4-31B	82.2 → 79.4 (−2.8)	83.8 → 79.9 (−3.9)	61.8 → 56.6 (−5.2)
	E-VQA	Qwen3-VL-8B	89.6 → 87.1 (−2.5)	92.9 → 89.2 (−3.7)	74.0 → 68.8 (−5.2)
		InternVL3.5-14B	84.1 → 82.0 (−2.1)	91.9 → 88.4 (−3.5)	67.5 → 62.7 (−4.8)
		Gemma4-31B	82.0 → 79.4 (−2.6)	86.2 → 82.1 (−4.1)	59.3 → 54.0 (−5.3)

Table 6: Zero-shot cross-dataset transfer for Fire-Then-Type (FTT). Each entry is in-domain → transferred performance, with the change in parentheses. Comp. and Dist. denote COMPREHENSION and DISTRACTION.

Dataset	Model	Method	Accuracy / Δ	E2E ms/query	vs. No Fix	vs. Always	Peak GiB
InfoSeek	Qwen3-VL-8B	No Fix	38.1	17,066	1.00×	0.35×	35.4
		Always Apply	42.5 (+4.4)	49,159	2.88×	1.00×	44.4
		Ours	42.7 (+4.6)	36,253	2.12×	0.74×	44.4
	InternVL3.5-14B	No Fix	21.6	8,862	1.00×	0.35×	36.9
		Always Apply	25.4 (+3.8)	25,206	2.84×	1.00×	40.1
		Ours	25.3 (+3.7)	19,551	2.21×	0.78×	40.1
	Gemma4-31B	No Fix	28.4	4,134	1.00×	0.36×	65.2
		Always Apply	29.7 (+1.3)	11,495	2.78×	1.00×	65.5
		Ours	30.1 (+1.7)	8,160	1.97×	0.71×	65.5
E-VQA	Qwen3-VL-8B	No Fix	54.9	15,841	1.00×	0.34×	35.6
		Always Apply	65.7 (+10.8)	46,274	2.92×	1.00×	44.7
		Ours	66.1 (+11.2)	35,894	2.26×	0.77×	45.0
	InternVL3.5-14B	No Fix	28.7	8,125	1.00×	0.34×	37.1
		Always Apply	36.2 (+7.5)	23,841	2.93×	1.00×	40.4
		Ours	36.0 (+7.3)	17,212	2.12×	0.72×	40.6
	Gemma4-31B	No Fix	45.2	3,971	1.00×	0.35×	65.4
		Always Apply	51.4 (+6.2)	11,287	2.84×	1.00×	65.7
		Ours	51.5 (+6.3)	7,190	1.81×	0.64×	65.9
VISA	Qwen3-VL-8B	No Fix	63.8	14,180	1.00×	0.35×	35.5
		Always Apply	64.1 (+0.3)	41,059	2.90×	1.00×	44.6
		Ours	64.7 (+0.9)	34,769	2.45×	0.85×	44.9
	InternVL3.5-14B	No Fix	45.0	7,348	1.00×	0.34×	37.0
		Always Apply	50.6 (+5.6)	21,694	2.95×	1.00×	40.3
		Ours	51.1 (+6.1)	18,863	2.57×	0.87×	40.5
	Gemma4-31B	No Fix	61.7	3,681	1.00×	0.35×	65.3
		Always Apply	62.8 (+1.1)	10,492	2.85×	1.00×	65.6
		Ours	63.2 (+1.5)	7,912	2.15×	0.75×	65.8

Table 7: End-to-end latency and peak-memory measurements. OURS reduces latency relative to ALWAYS APPLY by 13% to 36%, with similar peak memory to ALWAYS APPLY.

Method	**Multi-modal**	**RAG-specific**	**Online sequential decision**	**Typed failure diagnosis**	**Drives inter-vention**	**Pre-generation**
Static hidden-state probing						
Kadavath et al. (2022)	✗	✗	✗	✗	✗	N/A
Burns et al. (2023)	✗	✗	✗	✗	✗	N/A
Azaria and Mitchell (2023)	✗	✗	✗	✗	✗	N/A
Zhang et al. (2025a)	✗	✗	✗	✗	✗	✗
Yeh and Li (2026)	✗	✓	✗	✗	✗	N/A
Self-correction / corrective retrieval						
Self-Refine (Madaan et al., 2023)	✗	✗	✗	✗	✓	✗
Reflexion (Shinn et al., 2023)	✗	✗	✗	✗	✓	✗
Self-RAG (Asai et al., 2024)	✗	✓	✗	✗	✓	✗
CRAG (Yan et al., 2024)	✗	✓	✗	✗	✓	✗
OmniSearch (Li et al., 2025)	✓	✓	✗	✗	✓	✗
DeepMMSearch-R1 (Narayan et al., 2025)	✓	✓	✗	✗	✓	✗
Attention-level intervention						
ALFAR (An et al., 2025)	✓	✓	✗	✗	✓	✓
Ours	✓	✓	✓	✓	✓	✓

Table 8: Positioning of our framework relative to closely related lines of work. **Multimodal**: handles image-question inputs, not text-only. **RAG-specific**: targets failures induced by retrieved context, rather than general model behavior. **Online sequential decision**: makes a per-layer fire/no-fire decision during prefill with an earliness cost, rather than reading a single fixed representation or waiting for full generation. **Typed failure diagnosis**: distinguishes the two failure types (e.g., DISTRACTION vs. COMPREHENSION) rather than producing a single correct/incorrect signal. **Drives intervention**: the diagnostic signal is used to modify inputs or generation, not only for analysis. **Pre-generation**: diagnosis is available before any output token is decoded. N/A indicates that the dimension does not apply to the work's setting.

Model	λ_{early}	Accuracy (%)	Failure Recall (%)	False Fire Rate (%)	Fire Depth
Qwen3-VL-8B	0.00	90.85 ± 0.99	90.13 ± 1.40	8.08 ± 0.50	0.881 ± 0.011
	0.05	89.99 ± 1.18	89.06 ± 2.29	8.64 ± 0.86	0.845 ± 0.028
	0.10	89.99 ± 0.91	88.72 ± 1.97	8.14 ± 0.80	0.829 ± 0.035
	0.15	89.38 ± 2.42	86.65 ± 5.27	6.60 ± 1.92	0.808 ± 0.028
	0.20	90.20 ± 0.94	89.50 ± 1.95	8.76 ± 0.89	0.777 ± 0.020
	0.25	88.46 ± 2.63	86.05 ± 5.95	7.96 ± 2.36	0.773 ± 0.025
	0.30	89.82 ± 2.00	88.68 ± 4.28	8.48 ± 2.12	0.752 ± 0.024
	0.35	89.21 ± 0.74	87.37 ± 2.05	8.05 ± 1.28	0.740 ± 0.007
	0.40	89.18 ± 0.71	87.47 ± 2.01	8.30 ± 1.51	0.730 ± 0.006
	0.45	88.66 ± 0.42	87.34 ± 1.11	9.38 ± 1.05	0.724 ± 0.005
	0.50	87.83 ± 1.16	85.25 ± 2.70	8.36 ± 1.18	0.725 ± 0.012
InternVL3.5-14B	0.00	86.47 ± 0.88	91.41 ± 1.34	28.57 ± 2.32	0.897 ± 0.013
	0.05	85.35 ± 1.66	88.38 ± 2.99	23.87 ± 2.74	0.881 ± 0.016
	0.10	85.68 ± 1.40	88.92 ± 2.26	24.20 ± 1.45	0.878 ± 0.022
	0.15	85.81 ± 0.92	89.75 ± 2.19	26.22 ± 3.73	0.847 ± 0.013
	0.20	86.60 ± 0.41	90.36 ± 0.70	24.87 ± 2.69	0.827 ± 0.017
	0.25	86.26 ± 0.98	89.81 ± 1.74	24.54 ± 2.46	0.819 ± 0.024
	0.30	86.22 ± 1.42	90.30 ± 1.72	26.22 ± 3.16	0.799 ± 0.016
	0.35	84.27 ± 1.41	86.17 ± 2.67	21.51 ± 2.74	0.810 ± 0.019
	0.40	85.19 ± 0.92	87.16 ± 1.20	20.84 ± 2.88	0.788 ± 0.005
	0.45	60.62 ± 29.34	51.52 ± 42.07	11.60 ± 9.65	0.780 ± 0.002
	0.50	59.54 ± 28.47	50.58 ± 41.31	13.11 ± 11.07	0.786 ± 0.025
Gemma4-31B	0.00	89.17 ± 0.80	92.24 ± 1.11	18.25 ± 1.52	0.868 ± 0.015
	0.05	88.71 ± 1.02	90.83 ± 1.20	16.42 ± 0.86	0.807 ± 0.017
	0.10	88.46 ± 0.83	90.35 ± 1.87	16.08 ± 1.96	0.780 ± 0.012
	0.15	88.54 ± 0.68	91.38 ± 1.89	18.33 ± 3.11	0.758 ± 0.025
	0.20	88.54 ± 0.21	91.10 ± 1.18	17.67 ± 2.97	0.757 ± 0.019
	0.25	88.36 ± 0.44	90.90 ± 1.02	17.75 ± 2.76	0.740 ± 0.020
	0.30	87.02 ± 1.20	88.76 ± 1.35	17.17 ± 1.38	0.742 ± 0.004
	0.35	86.29 ± 1.16	88.69 ± 2.38	19.50 ± 2.10	0.740 ± 0.023
	0.40	75.41 ± 23.08	70.86 ± 35.44	13.58 ± 6.85	0.745 ± 0.016
	0.45	52.10 ± 27.96	35.03 ± 42.91	6.67 ± 8.20	0.729 ± 0.001
	0.50	29.27 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	N/A

Table 9: Raw multi-seed results behind Figure 3 on **InfoSeek** (Chen et al., 2023) test set. Values are mean $\pm$ std across 5 seeds.

Model	λ_{early}	Accuracy (%)	Failure Recall (%)	False Fire Rate (%)	Fire Depth
Qwen3-VL-8B	0.00	92.05 ± 0.46	91.64 ± 0.69	7.33 ± 0.20	0.881 ± 0.006
	0.05	91.45 ± 0.47	91.50 ± 0.94	8.61 ± 0.35	0.860 ± 0.014
	0.10	91.58 ± 0.36	91.44 ± 0.83	8.21 ± 0.37	0.829 ± 0.018
	0.15	90.76 ± 0.98	90.64 ± 2.15	9.06 ± 0.77	0.798 ± 0.012
	0.20	90.07 ± 0.47	89.99 ± 0.90	9.82 ± 0.42	0.777 ± 0.011
	0.25	89.22 ± 1.12	89.05 ± 2.47	10.53 ± 0.95	0.769 ± 0.012
	0.30	88.13 ± 0.85	87.84 ± 1.83	11.44 ± 0.96	0.748 ± 0.010
	0.35	88.33 ± 0.31	88.03 ± 0.84	11.25 ± 0.53	0.738 ± 0.003
	0.40	87.12 ± 0.36	86.69 ± 0.91	12.24 ± 0.62	0.729 ± 0.003
	0.45	86.30 ± 0.17	85.69 ± 0.50	12.80 ± 0.42	0.722 ± 0.001
	0.50	85.98 ± 0.44	84.94 ± 1.07	12.47 ± 0.57	0.723 ± 0.006
InternVL3.5-14B	0.00	88.94 ± 0.43	93.18 ± 0.66	24.00 ± 0.94	0.899 ± 0.007
	0.05	88.05 ± 0.77	92.62 ± 1.36	25.86 ± 1.16	0.886 ± 0.009
	0.10	88.52 ± 0.61	93.04 ± 0.99	25.24 ± 0.59	0.866 ± 0.010
	0.15	87.12 ± 0.42	91.50 ± 0.96	26.25 ± 1.49	0.849 ± 0.008
	0.20	86.08 ± 0.18	90.46 ± 0.32	27.26 ± 1.18	0.837 ± 0.008
	0.25	85.08 ± 0.57	89.36 ± 1.01	27.97 ± 1.09	0.809 ± 0.012
	0.30	85.50 ± 0.60	89.74 ± 0.75	27.43 ± 1.36	0.790 ± 0.007
	0.35	83.90 ± 0.61	87.66 ± 1.14	27.55 ± 1.12	0.785 ± 0.009
	0.40	83.03 ± 0.38	86.61 ± 0.59	27.89 ± 1.21	0.783 ± 0.004
	0.45	77.61 ± 12.36	78.65 ± 17.67	25.61 ± 3.97	0.774 ± 0.002
	0.50	76.54 ± 11.95	76.97 ± 17.31	24.77 ± 4.56	0.776 ± 0.010
Gemma4-31B	0.00	90.65 ± 0.29	93.55 ± 0.45	16.34 ± 0.66	0.869 ± 0.006
	0.05	90.83 ± 0.49	94.23 ± 0.61	17.37 ± 0.36	0.845 ± 0.008
	0.10	90.54 ± 0.37	93.82 ± 0.80	17.31 ± 0.80	0.802 ± 0.007
	0.15	89.76 ± 0.30	92.89 ± 0.91	17.79 ± 1.21	0.801 ± 0.011
	0.20	88.52 ± 0.15	91.54 ± 0.38	18.77 ± 1.26	0.779 ± 0.009
	0.25	89.06 ± 0.16	92.12 ± 0.48	18.34 ± 1.17	0.780 ± 0.010
	0.30	87.67 ± 0.47	90.49 ± 0.50	19.13 ± 0.67	0.750 ± 0.001
	0.35	87.16 ± 0.49	90.02 ± 0.98	19.71 ± 0.87	0.751 ± 0.011
	0.40	84.95 ± 9.62	86.79 ± 14.80	19.46 ± 2.90	0.751 ± 0.008
	0.45	80.27 ± 10.62	79.82 ± 16.48	18.65 ± 3.54	0.753 ± 0.002
	0.50	77.63 ± 0.09	75.30 ± 0.11	16.75 ± 0.08	0.737 ± 0.001

Table 10: Raw multi-seed results behind Figure 3 on **E-VQA** (Mensink et al., 2023) test set. Values are mean $\pm$ std across 5 seeds.

Model	λ_{early}	Accuracy (%)	Failure Recall (%)	False Fire Rate (%)	Fire Depth
Qwen3-VL-8B	0.00	72.72 ± 1.34	41.43 ± 8.75	8.47 ± 5.95	0.936 ± 0.013
	0.05	73.26 ± 1.56	48.98 ± 6.98	12.15 ± 5.41	0.918 ± 0.015
	0.10	73.03 ± 0.71	42.04 ± 7.23	8.34 ± 3.69	0.926 ± 0.030
	0.15	72.64 ± 1.50	39.59 ± 6.07	7.48 ± 4.20	0.922 ± 0.038
	0.20	73.18 ± 0.69	37.75 ± 1.83	5.52 ± 0.55	0.912 ± 0.022
	0.25	72.34 ± 0.29	40.00 ± 5.75	8.22 ± 3.73	0.891 ± 0.023
	0.30	71.80 ± 0.79	35.31 ± 8.63	6.26 ± 4.45	0.904 ± 0.028
	0.35	73.10 ± 0.61	36.12 ± 5.03	4.66 ± 2.15	0.879 ± 0.030
	0.40	73.03 ± 1.40	37.35 ± 4.11	5.52 ± 3.27	0.871 ± 0.015
	0.45	66.44 ± 4.89	12.45 ± 15.41	1.10 ± 1.66	0.880 ± 0.009
	0.50	62.45 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	N/A
InternVL3.5-14B	0.00	71.19 ± 0.56	43.33 ± 10.16	10.94 ± 5.66	0.978 ± 0.010
	0.05	70.57 ± 1.50	49.22 ± 6.66	15.72 ± 6.60	0.957 ± 0.014
	0.10	69.88 ± 0.79	55.10 ± 5.35	20.63 ± 4.39	0.935 ± 0.019
	0.15	70.65 ± 0.79	47.65 ± 4.90	14.59 ± 2.62	0.947 ± 0.027
	0.20	69.96 ± 1.53	48.63 ± 2.37	16.35 ± 3.88	0.953 ± 0.023
	0.25	70.34 ± 1.62	55.49 ± 6.19	20.13 ± 4.87	0.937 ± 0.013
	0.30	70.88 ± 0.42	47.65 ± 7.03	14.21 ± 4.71	0.942 ± 0.022
	0.35	69.73 ± 1.45	50.78 ± 3.24	18.11 ± 2.68	0.933 ± 0.014
	0.40	66.59 ± 4.67	32.55 ± 26.58	11.57 ± 9.48	0.931 ± 0.001
	0.45	64.60 ± 4.51	21.18 ± 26.05	7.55 ± 9.45	0.912 ± 0.002
	0.50	60.92 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	N/A
Gemma4-31B	0.00	77.70 ± 1.58	47.53 ± 2.49	5.60 ± 1.44	0.873 ± 0.027
	0.05	76.63 ± 1.33	47.96 ± 5.20	7.50 ± 1.67	0.841 ± 0.046
	0.10	75.86 ± 1.19	42.58 ± 4.74	5.71 ± 2.93	0.831 ± 0.036
	0.15	75.79 ± 0.29	44.30 ± 8.23	6.79 ± 4.27	0.805 ± 0.030
	0.20	73.87 ± 4.97	35.91 ± 18.34	5.12 ± 2.68	0.791 ± 0.012
	0.25	75.71 ± 1.42	46.45 ± 4.73	8.10 ± 2.25	0.787 ± 0.029
	0.30	73.26 ± 4.53	32.90 ± 16.75	4.40 ± 3.05	0.779 ± 0.032
	0.35	68.89 ± 5.54	20.00 ± 24.50	4.05 ± 4.96	0.717 ± 0.009
	0.40	69.04 ± 5.74	17.85 ± 22.07	2.62 ± 3.40	0.725 ± 0.007
	0.45	64.37 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	N/A
	0.50	64.37 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	N/A

Table 11: Raw multi-seed numbers behind the RQ1 hazard panels on **VISA** (Ma et al., 2025) test set with three candidates. Values are mean $\pm$ std across 5 seeds.

Model	**Method**	**Metric**	λ_{early}										
			0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Qwen3-VL	FTT	Acc	$91.9_{\pm 0.6}$	$90.6_{\pm 2.2}$	$90.7_{\pm 1.1}$	$89.7_{\pm 1.1}$	$90.6_{\pm 0.5}$	$89.2_{\pm 1.5}$	$89.7_{\pm 1.1}$	$88.2_{\pm 2.6}$	$88.2_{\pm 1.2}$	$84.7_{\pm 4.7}$	$84.9_{\pm 1.8}$
		Comp Rec	$93.7_{\pm 1.1}$	$92.6_{\pm 1.9}$	$92.1_{\pm 1.9}$	$92.4_{\pm 1.6}$	$92.2_{\pm 1.0}$	$90.1_{\pm 1.9}$	$90.7_{\pm 2.2}$	$89.0_{\pm 2.6}$	$89.2_{\pm 1.4}$	$84.2_{\pm 5.0}$	$84.7_{\pm 3.2}$
		Dist Rec	$78.4_{\pm 1.5}$	$74.5_{\pm 8.1}$	$76.5_{\pm 3.7}$	$70.0_{\pm 5.6}$	$76.2_{\pm 1.6}$	$73.6_{\pm 5.9}$	$74.6_{\pm 2.4}$	$70.4_{\pm 9.2}$	$69.2_{\pm 7.1}$	$64.7_{\pm 11.3}$	$61.1_{\pm 2.9}$
	D3C	Acc	$83.6_{\pm 2.1}$	$83.0_{\pm 1.6}$	$83.1_{\pm 1.2}$	$82.7_{\pm 1.4}$	$83.3_{\pm 0.8}$	$83.1_{\pm 0.5}$	$82.6_{\pm 1.2}$	$82.7_{\pm 0.6}$	$82.1_{\pm 0.7}$	$81.7_{\pm 0.9}$	$81.2_{\pm 0.4}$
		Comp Rec	$86.1_{\pm 1.5}$	$84.7_{\pm 1.7}$	$84.2_{\pm 1.3}$	$83.1_{\pm 0.8}$	$85.0_{\pm 1.5}$	$83.5_{\pm 1.1}$	$85.0_{\pm 0.5}$	$83.1_{\pm 1.4}$	$82.5_{\pm 1.3}$	$81.5_{\pm 1.7}$	$80.8_{\pm 0.4}$
		Dist Rec	$56.7_{\pm 7.6}$	$55.5_{\pm 7.7}$	$56.5_{\pm 5.0}$	$56.0_{\pm 8.5}$	$55.7_{\pm 7.8}$	$59.5_{\pm 2.3}$	$54.4_{\pm 7.9}$	$57.2_{\pm 1.2}$	$57.8_{\pm 3.3}$	$55.7_{\pm 3.4}$	$54.6_{\pm 2.3}$
InternVL3.5	FTT	Acc	$87.6_{\pm 2.7}$	$87.2_{\pm 1.4}$	$87.7_{\pm 1.0}$	$86.2_{\pm 3.0}$	$85.9_{\pm 1.2}$	$85.3_{\pm 1.8}$	$79.9_{\pm 1.5}$	$79.7_{\pm 1.2}$	$79.0_{\pm 1.9}$	$77.5_{\pm 2.3}$	$76.0_{\pm 2.9}$
		Comp Rec	$93.3_{\pm 1.4}$	$92.6_{\pm 1.3}$	$92.5_{\pm 1.7}$	$91.9_{\pm 2.6}$	$90.1_{\pm 2.3}$	$89.6_{\pm 2.2}$	$88.6_{\pm 1.4}$	$88.2_{\pm 3.6}$	$88.2_{\pm 2.6}$	$86.6_{\pm 3.2}$	$83.4_{\pm 3.1}$
		Dist Rec	$76.4_{\pm 10.4}$	$74.9_{\pm 5.3}$	$78.4_{\pm 3.3}$	$72.3_{\pm 9.6}$	$76.3_{\pm 4.0}$	$75.6_{\pm 5.5}$	$53.2_{\pm 6.4}$	$51.7_{\pm 6.6}$	$50.0_{\pm 8.9}$	$46.4_{\pm 4.5}$	$46.7_{\pm 8.5}$
	D3C	Acc	$78.9_{\pm 1.1}$	$78.1_{\pm 0.4}$	$77.5_{\pm 0.7}$	$77.3_{\pm 0.9}$	$77.7_{\pm 0.9}$	$77.8_{\pm 0.7}$	$77.9_{\pm 0.3}$	$76.9_{\pm 0.9}$	$77.3_{\pm 0.9}$	$77.1_{\pm 0.7}$	$76.5_{\pm 1.0}$
		Comp Rec	$86.2_{\pm 1.2}$	$90.9_{\pm 1.1}$	$89.3_{\pm 2.6}$	$87.9_{\pm 2.2}$	$88.3_{\pm 2.3}$	$87.7_{\pm 1.5}$	$88.3_{\pm 0.9}$	$87.8_{\pm 1.3}$	$88.5_{\pm 0.2}$	$88.0_{\pm 2.4}$	$85.8_{\pm 1.4}$
		Dist Rec	$54.1_{\pm 6.3}$	$40.8_{\pm 4.3}$	$42.3_{\pm 7.7}$	$46.4_{\pm 5.2}$	$43.5_{\pm 4.6}$	$47.2_{\pm 6.1}$	$45.7_{\pm 1.8}$	$41.4_{\pm 5.5}$	$40.9_{\pm 4.6}$	$40.4_{\pm 8.1}$	$42.8_{\pm 6.6}$
Gemma4	FTT	Acc	$83.5_{\pm 0.5}$	$83.8_{\pm 0.5}$	$82.9_{\pm 0.6}$	$83.2_{\pm 0.8}$	$82.2_{\pm 1.1}$	$82.2_{\pm 0.9}$	$81.4_{\pm 0.6}$	$81.8_{\pm 0.4}$	$81.0_{\pm 1.0}$	$80.4_{\pm 0.7}$	$80.3_{\pm 0.7}$
		Comp Rec	$86.7_{\pm 1.2}$	$88.9_{\pm 1.4}$	$85.5_{\pm 2.4}$	$86.5_{\pm 1.6}$	$83.8_{\pm 2.3}$	$85.0_{\pm 1.5}$	$83.4_{\pm 2.1}$	$84.4_{\pm 1.7}$	$82.0_{\pm 2.0}$	$83.2_{\pm 2.3}$	$82.5_{\pm 1.6}$
		Dist Rec	$62.9_{\pm 4.3}$	$57.0_{\pm 4.9}$	$62.3_{\pm 5.7}$	$58.5_{\pm 3.3}$	$61.8_{\pm 3.6}$	$60.4_{\pm 0.6}$	$57.9_{\pm 4.4}$	$59.2_{\pm 2.8}$	$58.6_{\pm 3.0}$	$49.3_{\pm 6.5}$	$53.0_{\pm 6.5}$
	D3C	Acc	$81.7_{\pm 1.4}$	$82.3_{\pm 1.0}$	$81.7_{\pm 0.5}$	$81.0_{\pm 0.9}$	$80.9_{\pm 1.5}$	$81.4_{\pm 1.0}$	$81.1_{\pm 1.4}$	$81.0_{\pm 0.2}$	$78.7_{\pm 3.0}$	$79.8_{\pm 0.6}$	$80.6_{\pm 0.4}$
		Comp Rec	$81.8_{\pm 3.7}$	$83.2_{\pm 1.1}$	$82.1_{\pm 1.8}$	$80.2_{\pm 1.6}$	$80.1_{\pm 3.3}$	$81.4_{\pm 3.1}$	$81.3_{\pm 3.2}$	$79.8_{\pm 0.6}$	$77.9_{\pm 6.8}$	$77.7_{\pm 0.6}$	$81.0_{\pm 1.6}$
		Dist Rec	$65.3_{\pm 3.9}$	$62.9_{\pm 4.1}$	$60.9_{\pm 8.0}$	$65.7_{\pm 5.1}$	$63.8_{\pm 7.2}$	$63.5_{\pm 7.6}$	$59.1_{\pm 8.6}$	$64.9_{\pm 4.0}$	$56.3_{\pm 11.9}$	$63.3_{\pm 7.2}$	$58.9_{\pm 4.7}$

Table 12: Values plotted in Figure 4 for **InfoSeek**. Each cell is $\text{mean}_{\pm \text{std}}$ across five seeds.

Model	Method	Metric	λ_{early}										
			0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Qwen3-VL	FTT	Acc	$92.3_{\pm 0.6}$	$91.6_{\pm 2.3}$	$91.8_{\pm 1.1}$	$90.8_{\pm 1.1}$	$89.6_{\pm 0.6}$	$88.2_{\pm 1.5}$	$86.5_{\pm 1.0}$	$86.9_{\pm 2.6}$	$84.8_{\pm 1.2}$	$83.0_{\pm 4.7}$	$81.8_{\pm 1.9}$
		Comp Rec	$93.9_{\pm 1.1}$	$94.6_{\pm 1.8}$	$94.5_{\pm 1.9}$	$93.8_{\pm 1.6}$	$92.9_{\pm 1.1}$	$91.8_{\pm 1.9}$	$90.7_{\pm 2.2}$	$91.2_{\pm 2.6}$	$89.5_{\pm 1.4}$	$87.7_{\pm 5.1}$	$85.6_{\pm 3.2}$
		Dist Rec	$80.0_{\pm 1.5}$	$77.7_{\pm 8.3}$	$78.6_{\pm 3.3}$	$76.4_{\pm 5.7}$	$74.0_{\pm 1.7}$	$71.5_{\pm 5.8}$	$67.5_{\pm 2.4}$	$68.1_{\pm 9.1}$	$63.8_{\pm 7.1}$	$60.1_{\pm 11.0}$	$58.2_{\pm 3.0}$
	D3C	Acc	$83.8_{\pm 2.1}$	$84.3_{\pm 1.6}$	$84.1_{\pm 1.3}$	$83.5_{\pm 1.4}$	$82.7_{\pm 0.8}$	$81.9_{\pm 0.6}$	$82.1_{\pm 1.3}$	$81.4_{\pm 0.6}$	$80.2_{\pm 0.7}$	$79.3_{\pm 0.9}$	$79.0_{\pm 0.4}$
		Comp Rec	$85.8_{\pm 1.5}$	$86.5_{\pm 1.7}$	$86.2_{\pm 1.4}$	$85.6_{\pm 0.9}$	$84.9_{\pm 1.5}$	$84.2_{\pm 1.3}$	$84.3_{\pm 0.5}$	$84.5_{\pm 1.4}$	$83.2_{\pm 1.3}$	$82.3_{\pm 1.7}$	$81.5_{\pm 0.5}$
		Dist Rec	$57.4_{\pm 7.5}$	$61.0_{\pm 7.7}$	$59.5_{\pm 5.1}$	$57.6_{\pm 8.5}$	$56.1_{\pm 7.6}$	$54.5_{\pm 2.4}$	$55.0_{\pm 8.2}$	$53.0_{\pm 1.1}$	$50.9_{\pm 3.4}$	$49.0_{\pm 3.4}$	$48.4_{\pm 2.2}$
InternVL3.5	FTT	Acc	$88.8_{\pm 2.6}$	$87.2_{\pm 1.3}$	$88.0_{\pm 1.1}$	$86.0_{\pm 3.0}$	$84.1_{\pm 1.2}$	$82.7_{\pm 1.8}$	$83.1_{\pm 1.5}$	$80.0_{\pm 1.2}$	$77.8_{\pm 1.8}$	$76.0_{\pm 2.2}$	$74.8_{\pm 2.9}$
		Comp Rec	$94.2_{\pm 1.4}$	$93.9_{\pm 1.3}$	$94.2_{\pm 1.8}$	$93.0_{\pm 2.6}$	$91.9_{\pm 2.4}$	$90.6_{\pm 2.4}$	$90.6_{\pm 1.3}$	$88.1_{\pm 3.6}$	$86.1_{\pm 2.5}$	$84.6_{\pm 3.3}$	$83.0_{\pm 3.1}$
		Dist Rec	$78.7_{\pm 10.3}$	$74.2_{\pm 5.4}$	$76.5_{\pm 3.2}$	$71.6_{\pm 9.8}$	$67.5_{\pm 3.8}$	$65.2_{\pm 5.5}$	$66.4_{\pm 6.6}$	$60.2_{\pm 6.7}$	$56.5_{\pm 8.9}$	$53.3_{\pm 4.4}$	$51.0_{\pm 8.4}$
	D3C	Acc	$79.5_{\pm 1.1}$	$78.4_{\pm 0.5}$	$77.5_{\pm 0.6}$	$77.8_{\pm 0.9}$	$78.3_{\pm 0.9}$	$77.1_{\pm 0.7}$	$76.4_{\pm 0.4}$	$75.5_{\pm 0.9}$	$75.8_{\pm 0.9}$	$74.2_{\pm 0.7}$	$73.5_{\pm 1.0}$
		Comp Rec	$89.4_{\pm 1.3}$	$89.3_{\pm 1.1}$	$89.1_{\pm 2.7}$	$88.6_{\pm 2.2}$	$88.3_{\pm 2.4}$	$88.4_{\pm 1.4}$	$88.3_{\pm 0.8}$	$87.6_{\pm 1.3}$	$87.6_{\pm 0.2}$	$86.2_{\pm 2.5}$	$85.2_{\pm 1.4}$
		Dist Rec	$49.3_{\pm 6.0}$	$46.4_{\pm 4.4}$	$44.2_{\pm 7.7}$	$45.8_{\pm 5.2}$	$47.7_{\pm 4.7}$	$43.8_{\pm 5.9}$	$42.0_{\pm 1.8}$	$40.7_{\pm 5.5}$	$41.5_{\pm 4.6}$	$39.3_{\pm 8.1}$	$38.8_{\pm 6.7}$
Gemma4	FTT	Acc	$84.0_{\pm 0.4}$	$84.9_{\pm 0.6}$	$84.2_{\pm 0.6}$	$83.5_{\pm 0.8}$	$82.0_{\pm 1.0}$	$82.7_{\pm 0.9}$	$80.8_{\pm 0.6}$	$80.2_{\pm 0.5}$	$78.6_{\pm 1.0}$	$76.8_{\pm 0.7}$	$77.3_{\pm 0.6}$
		Comp Rec	$87.5_{\pm 1.2}$	$88.6_{\pm 1.4}$	$87.7_{\pm 2.5}$	$87.5_{\pm 1.6}$	$86.2_{\pm 2.2}$	$86.7_{\pm 1.5}$	$85.0_{\pm 2.1}$	$84.5_{\pm 1.7}$	$82.9_{\pm 2.0}$	$81.3_{\pm 2.4}$	$81.7_{\pm 1.5}$
		Dist Rec	$62.5_{\pm 4.5}$	$66.2_{\pm 5.1}$	$64.3_{\pm 5.6}$	$62.5_{\pm 3.3}$	$59.3_{\pm 3.2}$	$61.0_{\pm 0.8}$	$56.7_{\pm 4.3}$	$55.6_{\pm 2.9}$	$52.7_{\pm 3.4}$	$48.4_{\pm 6.3}$	$49.6_{\pm 6.3}$
	D3C	Acc	$82.4_{\pm 1.4}$	$83.5_{\pm 1.0}$	$83.0_{\pm 0.5}$	$81.8_{\pm 0.9}$	$80.9_{\pm 1.6}$	$81.5_{\pm 1.0}$	$80.8_{\pm 1.4}$	$79.3_{\pm 0.3}$	$78.0_{\pm 2.9}$	$77.8_{\pm 0.6}$	$76.6_{\pm 0.4}$
		Comp Rec	$82.9_{\pm 3.8}$	$84.2_{\pm 1.1}$	$84.0_{\pm 1.8}$	$82.4_{\pm 1.7}$	$81.8_{\pm 3.4}$	$82.6_{\pm 3.1}$	$82.3_{\pm 3.1}$	$81.2_{\pm 0.7}$	$80.0_{\pm 6.7}$	$79.0_{\pm 0.6}$	$78.2_{\pm 1.6}$
		Dist Rec	$63.2_{\pm 4.0}$	$68.0_{\pm 4.3}$	$65.1_{\pm 7.9}$	$61.6_{\pm 4.9}$	$59.4_{\pm 7.4}$	$62.1_{\pm 7.7}$	$58.3_{\pm 8.3}$	$55.0_{\pm 4.0}$	$52.1_{\pm 12.2}$	$53.1_{\pm 7.1}$	$49.9_{\pm 4.7}$

Table 13: Values plotted in Figure 4 for **E-VQA**. Each cell is $\text{mean}_{\pm \text{std}}$ across five seeds.

Model	Method	Metric	λ_{early}										
			0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Qwen3-VL	FTT	Acc	$74.1_{\pm 0.4}$	$73.7_{\pm 1.3}$	$73.4_{\pm 1.0}$	$74.2_{\pm 0.9}$	$73.3_{\pm 1.3}$	$72.5_{\pm 1.1}$	$74.0_{\pm 0.8}$	$73.4_{\pm 1.3}$	$73.0_{\pm 0.5}$	$72.2_{\pm 0.7}$	$73.3_{\pm 1.0}$
		Comp Rec	$48.4_{\pm 5.6}$	$45.0_{\pm 7.4}$	$46.2_{\pm 12.5}$	$41.5_{\pm 4.0}$	$47.1_{\pm 4.0}$	$42.3_{\pm 7.3}$	$43.0_{\pm 5.5}$	$48.9_{\pm 5.8}$	$40.6_{\pm 6.8}$	$44.0_{\pm 5.8}$	$40.8_{\pm 2.6}$
		Dist Rec	$11.8_{\pm 4.7}$	$10.5_{\pm 5.0}$	$10.5_{\pm 3.1}$	$8.0_{\pm 2.5}$	$6.8_{\pm 0.0}$	$9.2_{\pm 3.1}$	$6.8_{\pm 0.0}$	$9.2_{\pm 3.1}$	$11.8_{\pm 7.3}$	$6.8_{\pm 0.0}$	$6.8_{\pm 0.0}$
	D3C	Acc	$69.5_{\pm 1.3}$	$68.9_{\pm 1.2}$	$70.4_{\pm 0.5}$	$69.5_{\pm 1.5}$	$67.7_{\pm 1.3}$	$69.0_{\pm 0.8}$	$68.6_{\pm 1.3}$	$68.7_{\pm 0.7}$	$69.9_{\pm 0.7}$	$68.7_{\pm 2.0}$	$68.6_{\pm 2.5}$
		Comp Rec	$37.3_{\pm 4.1}$	$39.5_{\pm 9.2}$	$37.1_{\pm 3.3}$	$37.3_{\pm 6.8}$	$29.5_{\pm 8.6}$	$34.6_{\pm 7.4}$	$36.1_{\pm 7.5}$	$33.7_{\pm 7.4}$	$31.7_{\pm 1.1}$	$34.9_{\pm 5.6}$	$37.8_{\pm 4.6}$
		Dist Rec	$2.5_{\pm 3.1}$	$0.0_{\pm 0.0}$	$3.8_{\pm 5.0}$	$2.5_{\pm 3.1}$	$1.2_{\pm 2.5}$	$0.0_{\pm 0.0}$	$1.2_{\pm 2.5}$	$1.2_{\pm 2.5}$	$0.0_{\pm 0.0}$	$3.8_{\pm 3.1}$	$1.2_{\pm 2.5}$
InternVL3.5	FTT	Acc	$69.5_{\pm 2.0}$	$70.2_{\pm 1.3}$	$69.2_{\pm 0.8}$	$67.7_{\pm 1.0}$	$69.6_{\pm 1.7}$	$68.7_{\pm 1.0}$	$68.0_{\pm 0.5}$	$69.3_{\pm 1.4}$	$69.3_{\pm 1.0}$	$69.9_{\pm 0.4}$	$70.2_{\pm 1.6}$
		Comp Rec	$52.0_{\pm 11.5}$	$54.9_{\pm 6.4}$	$38.1_{\pm 5.3}$	$53.4_{\pm 18.6}$	$51.1_{\pm 11.9}$	$51.7_{\pm 17.7}$	$53.4_{\pm 20.6}$	$47.1_{\pm 12.3}$	$43.6_{\pm 13.4}$	$58.1_{\pm 3.6}$	$56.9_{\pm 14.5}$
		Dist Rec	$29.1_{\pm 8.5}$	$25.4_{\pm 10.0}$	$29.7_{\pm 11.3}$	$33.3_{\pm 10.0}$	$27.8_{\pm 12.1}$	$27.8_{\pm 2.7}$	$30.3_{\pm 3.5}$	$23.6_{\pm 9.7}$	$25.4_{\pm 7.0}$	$25.4_{\pm 7.0}$	$27.8_{\pm 5.1}$
	D3C	Acc	$63.9_{\pm 0.8}$	$64.2_{\pm 2.7}$	$64.2_{\pm 1.5}$	$63.8_{\pm 1.1}$	$65.3_{\pm 2.0}$	$65.4_{\pm 0.9}$	$64.2_{\pm 0.7}$	$64.9_{\pm 1.7}$	$66.3_{\pm 2.0}$	$64.6_{\pm 1.8}$	$66.1_{\pm 1.2}$
		Comp Rec	$20.0_{\pm 11.9}$	$26.1_{\pm 11.0}$	$35.1_{\pm 13.6}$	$27.2_{\pm 16.1}$	$24.6_{\pm 11.6}$	$33.3_{\pm 7.9}$	$34.2_{\pm 9.6}$	$29.0_{\pm 15.1}$	$27.2_{\pm 3.7}$	$30.1_{\pm 13.6}$	$21.2_{\pm 10.9}$
		Dist Rec	$24.9_{\pm 11.1}$	$19.4_{\pm 7.3}$	$24.9_{\pm 4.8}$	$22.4_{\pm 4.9}$	$22.4_{\pm 5.9}$	$14.5_{\pm 6.5}$	$18.8_{\pm 7.3}$	$17.6_{\pm 9.7}$	$18.8_{\pm 6.2}$	$22.4_{\pm 7.3}$	$22.4_{\pm 9.7}$
Gemma4	FTT	Acc	$78.8_{\pm 3.7}$	$80.4_{\pm 2.6}$	$79.4_{\pm 1.7}$	$79.5_{\pm 2.4}$	$78.1_{\pm 2.5}$	$77.6_{\pm 1.3}$	$77.4_{\pm 2.8}$	$73.8_{\pm 4.0}$	$73.3_{\pm 4.2}$	$72.8_{\pm 4.7}$	$70.2_{\pm 2.4}$
		Comp Rec	$56.8_{\pm 16.4}$	$63.3_{\pm 11.6}$	$60.8_{\pm 9.6}$	$59.9_{\pm 10.3}$	$57.0_{\pm 14.9}$	$51.4_{\pm 7.9}$	$49.6_{\pm 10.8}$	$39.7_{\pm 15.8}$	$38.1_{\pm 16.0}$	$35.4_{\pm 16.0}$	$26.2_{\pm 6.7}$
		Dist Rec	$54.0_{\pm 10.0}$	$44.0_{\pm 18.7}$	$44.0_{\pm 24.5}$	$49.0_{\pm 15.8}$	$29.0_{\pm 10.0}$	$39.0_{\pm 20.0}$	$29.0_{\pm 10.0}$	$44.0_{\pm 24.5}$	$34.0_{\pm 20.0}$	$24.0_{\pm 0.0}$	$24.0_{\pm 0.0}$
	D3C	Acc	$73.7_{\pm 0.5}$	$72.9_{\pm 1.5}$	$74.2_{\pm 0.6}$	$72.5_{\pm 1.2}$	$71.5_{\pm 1.8}$	$72.5_{\pm 1.6}$	$73.7_{\pm 0.6}$	$72.2_{\pm 2.0}$	$71.3_{\pm 1.9}$	$70.2_{\pm 4.1}$	$71.2_{\pm 2.5}$
		Comp Rec	$39.5_{\pm 4.6}$	$39.3_{\pm 4.9}$	$43.1_{\pm 1.8}$	$36.4_{\pm 9.5}$	$29.9_{\pm 12.9}$	$34.8_{\pm 9.6}$	$42.7_{\pm 2.9}$	$31.7_{\pm 9.9}$	$27.4_{\pm 10.9}$	$23.8_{\pm 16.0}$	$27.2_{\pm 12.2}$
		Dist Rec	$20.0_{\pm 10.0}$	$30.0_{\pm 10.0}$	$10.0_{\pm 12.2}$	$20.0_{\pm 18.7}$	$30.0_{\pm 18.7}$	$30.0_{\pm 10.0}$	$15.0_{\pm 12.2}$	$5.0_{\pm 10.0}$	$10.0_{\pm 12.2}$	$5.0_{\pm 10.0}$	$10.0_{\pm 12.2}$

Table 14: Values plotted in Figure 4 for **VISA**. Each cell is $\text{mean}_{\pm \text{std}}$ across five seeds.