A PLATEAU IS NOT A CEILING: EVALUATING HEADROOM IN CONVERSATIONAL IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

A late-turn plateau in conversational image retrieval summarizes the outcome of an interaction, but does not establish which opportunities for improvement have been exhausted. Historical success, query construction, and target access affect that interpretation in different ways. We study retrieval headroom through three controlled comparisons: continuations without additional target observations, alternative representations of a fixed dialogue, and additional target-grounded observations. On FS-COCO-Dialog, reordering a frozen top-100 shortlist after four turns raises cumulative H@10 from 39.7 to 50.5 while current H@10 falls from 30.9 to 6.0. On DAR-preprocessed VisDial states, mean pooling raises current H@10 from 62.2 to 74.9 under ChatIR-FT and from 51.6 to 79.9 under BLIP-ITC, with a positive gain also on prefixes that are never truncated. Across broader fixed-dialogue comparisons, the preferred construction depends on the encoder and input regime. Additional evidence remains useful beyond the tested dialogue representations: under held-out interface selection, a paired human sketch improves all twelve dialogue-state–retrieval-space pairs by 9.8–22.7 points. On the same episodes, rewriting without target access matches or exceeds the sketch on concatenated queries in three of four spaces but adds little to an already reformulated state, an interaction that is significant in every retrieval space. These comparisons separately probe historical exposure, representation headroom, and evidence-access headroom, without assuming that they form independent causes of failure. The results support evaluating saturation relative to the states and observations tested, rather than inferring an information ceiling from a flattened retrieval curve. Code is available at https://anonymous.4open.science/r/iclr2027-DA65/.

1 INTRODUCTION

Conversational image retrieval uses interaction to resolve an incomplete query. A system asks about the target, receives an answer, and updates its ranking, so progress is commonly summarized by a retrieval curve over dialogue turns (Levy et al., 2023; Lee et al., 2024; Lin et al., 2026; Xie et al., 2026; Bai et al., 2025b). When late turns add little, however, the curve does not reveal whether the system should keep asking, represent the existing dialogue differently, or seek another observation.

Three distinctions matter. Historical success does not determine the quality of the current ranking. The same observed dialogue can induce different retrieval states depending on how it is represented. Performance under one evidence regime also does not determine whether another target-grounded observation would help. A plateau along one trajectory therefore does not establish that retrieval opportunities have been exhausted.

Existing work addresses parts of this ambiguity. PlugIR separates current from cumulative retrieval and reformulates dialogue for pretrained encoders (Lee et al., 2024), while MAQIU studies the gap between historical success and round-wise accuracy through memory-based intent modeling (Chen et al., 2026). We ask a broader question: *which retrieval opportunities remain when historical exposure, query representation, and target access are varied separately?* We test these axes under matched target cohorts and galleries.

First, we remove further target observations. After four turns on FS-COCO-Dialog, reordering a frozen top-100 shortlist raises cumulative H@10 from 39.7 to 50.5 while current H@10 falls from

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.

30.9 to 6.0. Historical improvement can therefore arise from repeated exposure without a stronger retrieval state.

Second, we hold the observed dialogue fixed and vary only query construction. On DAR-preprocessed VisDial, mean pooling improves current H@10 by 12.7 points under ChatIR-FT and by 28.3 points under BLIP-ITC. On FS-COCO-Dialog, target-blind reformulation adds 14.5–29.9 points across three retrieval spaces. The direction is not universal: on raw human VisDial under ChatIR-FT, concatenation remains strongest. Fixed observations thus contain substantial retrieval headroom, but the effective representation depends on the encoder–input regime.

Third, we expand target access after reducing representation headroom. Across four retrieval spaces, a dialogue-only rewrite adds 14.0–25.0 H@10 points from concatenated states but only 0.5–1.9 after PlugIR reformulation. A target-grounded human sketch still adds 9.8–12.3 points from those reformulated states. The same separation appears on raw human VisDial, where another rewrite adds only 1.6–3.2 points after PlugIR while target-informed appearance text adds 20.9–23.5. Moreover, all twelve state–space pairs improve under held-out selection, and under CLIP, sketch gains on pooled and reformulated states remain 8.9–10.0 points after extending the same dialogues to twenty turns. These results show that gains from added evidence can overlap with representation headroom, yet substantial evidence-access headroom can remain after stronger tested representations.

Our contributions are threefold:

- We formulate controlled tests of retrieval headroom that separate historical exposure, query representation, and evidence access, making explicit what is held fixed and what each positive result establishes.
- We show substantial headroom without additional target observations. Target-blind continuations can increase historical success while current retrieval deteriorates, and fixed-dialogue representation changes yield 12.7–29.9 H@10 gains in six of seven encoder–input settings.
- We characterize residual headroom under expanded target access. The tested dialogue-only rewrites add little after reformulation, while target-informed text and human sketches retain substantial gains across generated and human dialogue. The interaction is estimated directly with paired intervals in four retrieval spaces. Correspondence, degradation, fusion-harm, source-complementarity, and best-of-eight next-question controls delimit the scope of these gains.

These tests can refute a ceiling within their stated query and observation families, but failure of a finite probe cannot establish that every useful representation or observation has been exhausted.

2 Related Work

2.1 Conversational Image Retrieval

Conversational image retrieval improves ranking through question selection and query refinement. ChatIR learns to retrieve from dialogue histories (Levy et al., 2023), while PlugIR combines context reformulation with candidate-aware questioning (Lee et al., 2024). CAST transforms contextual representations, and MAQIU maintains query intent through explicit memory (Lin et al., 2026; Chen et al., 2026). Ask-to-Retrieve and ICL improve question selection and iterative refinement, respectively (Xie et al., 2026; Qin et al., 2025). These methods address two related decisions: which observation to acquire and how to incorporate it into the retrieval state. Their interaction makes an observed plateau difficult to interpret, since limited progress may reflect either the acquired dialogue or its representation. We hold observations fixed to test how much retrieval value remains available through query construction alone, then examine whether additional target access improves the resulting states.

2.2 Trajectory-Based Evaluation

Cumulative success records whether retrieval has succeeded at any previous turn (Levy et al., 2023). PlugIR distinguishes this objective from current recall and introduces Best log Rank Integral to

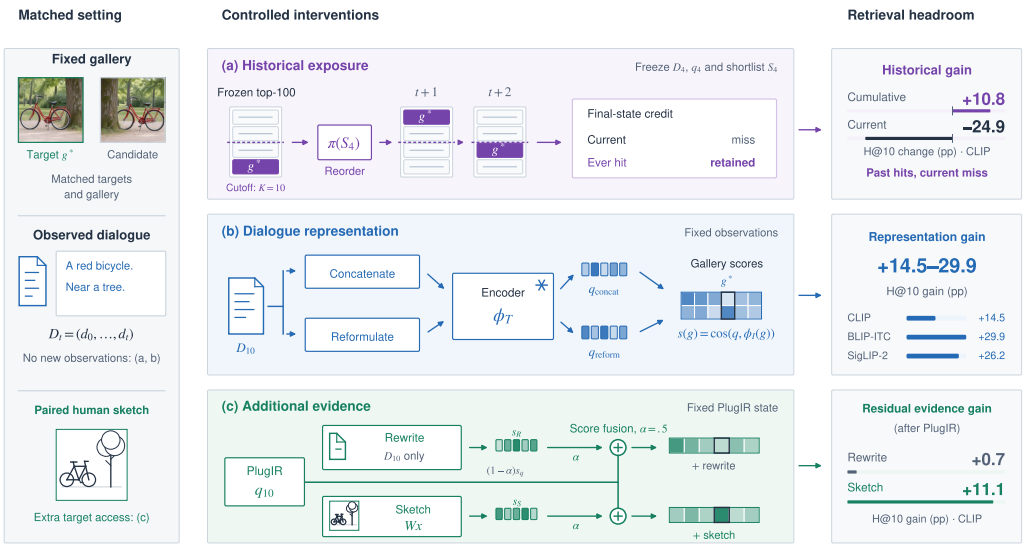

Figure 1: **An observed plateau need not be a retrieval ceiling.** Controlled probes on FS-COCO-Dialog reveal **(a)** historical gains despite worse current retrieval, **(b)** representation gains from fixed observations, and **(c)** residual sketch gains beyond PlugIR. Right-column results are measured. Other visuals are illustrative. Bar scales are shared within each panel.

summarize the running-best trajectory (Lee et al., 2024). MAQIU also reports historical and round-wise performance to expose inconsistent retrieval (Chen et al., 2026). These measures answer different questions: whether a target has been exposed and whether the current state still retrieves it. Historical credit remains useful when users stop after a successful result, but does not establish that later interaction improves ranking. Building on this distinction, we test continuations without additional target observations. These controls quantify historical gains due to repeated candidate exposure and examine whether they coexist with deterioration in current retrieval.

2.3 Multimodal Query Construction

Visual queries can be derived from existing dialogue or acquired as another observation. DAR and ADaFuSE generate images from dialogue (Long et al., 2025; Zhang et al., 2026), while sketch retrieval uses target-grounded drawings, alone or with text (Chowdhury et al., 2022; Sangkloy et al., 2022; Koley et al., 2024; Braghis & Liu, 2024; Demić & Čehovin Zajc, 2025). Pic2Word, SEARLE, LinCIR, CIReVL, and OSrCIR provide alternative ways to incorporate visual inputs (Saito et al., 2023; Baldrati et al., 2023; Gu et al., 2024; Karthik et al., 2024; Tang et al., 2025). These routes differ in both evidence access and the interface used to represent visual content. A gain over dialogue alone therefore does not establish whether the benefit comes from additional observations or a stronger representation of existing ones. We compare dialogue-only rewriting with target-grounded evidence on matched episodes and test whether a human sketch remains useful after dialogue refinement. Evidence access and supervision are specified for each route, so the comparison measures headroom under the tested conditions.

3 Evaluating Retrieval Headroom

A late-turn plateau summarizes one retrieval trajectory. To test which opportunities remain, we vary historical exposure, query construction, and target access under matched conditions while keeping the target cohort and gallery fixed within each comparison.

3.1 Controlled Comparisons

Let g^* denote the target, d_0 the initial caption, and $D_t = (d_0, \ldots, d_t)$ the observed prefix. A query construction maps D_t to a retrieval state q_t, producing target rank r_t. New units are question–answer pairs on FS-COCO-Dialog and standalone descriptions in DAR-preprocessed VisDial. Raw VisDial controls use the original human dialogue. We use ten turns unless stated otherwise and use *plateau* to mean limited late-turn progress rather than statistical convergence.

Measurement controls freeze observations and vary subsequent rankings. *Representation controls* fix the dialogue and vary query construction. *Evidence-access probes* introduce target-informed text or a paired sketch. A representation gain reveals unused retrieval value in the observed dialogue, while an evidence-access gain reveals recoverable errors under expanded access. The latter may reflect new content, better wording, or correction of noisy evidence. These tests establish headroom within the tested query and observation families rather than an independent decomposition of failure causes.

The ordering of these comparisons matters. A gain from an added source over a weak dialogue state can reflect an opportunity that a target-blind rewrite would also recover. We therefore compare evidence routes again after strengthening the dialogue representation. The relevant result is whether the additional source still improves that state, not simply whether it improves concatenation. This conditions the evidence-access claim on a stronger tested baseline without assuming that representation headroom has been exhausted.

3.2 Evaluation Metrics

For cutoff K, current success measures the present ranking while cumulative success retains any earlier hit:

$$I_t = \Pr(r_t \le K), \qquad C_t = \Pr\left(\min_{0 \le u \le t} r_u \le K\right). \tag{1}$$

Their difference $G_t = C_t - I_t$ is the *masked gap*, the share of episodes retrieved previously but missed now. A zero gap is not itself evidence of good retrieval, so historical and current quantities must be interpreted together. We additionally report BRI, the log-rank integral of PlugIR over the running-best rank, and ALRI, the same integral over the current rank. Appendix A gives their definitions and shows that any non-decreasing cumulative curve is compatible with retaining every earlier hit or none of them, so historical success does not identify current retrieval. The empirical controls below test how strongly this ambiguity appears in actual retrieval states.

3.3 Experimental Setup

Datasets and observations. FS-COCO-Dialog extends FS-COCO (Chowdhury et al., 2022) with ten-turn dialogues, using 7k training targets and 3k disjoint evaluation targets in a 50k gallery of all 10k FS-COCO photographs and 40k MS-COCO distractors. An image-blind Qwen3-VL-32B questioner (Bai et al., 2025a) interacts with a target-grounded MiniCPM-V-4.5 answerer (Yu et al., 2026) under an open protocol without yes/no questions (Appendix H.1). Answer fidelity is uncertain, as two automated judges agree on only 57.8% of sampled answers (Appendix H.2). Sketch comparisons use all 3k evaluation queries. VisDial (Das et al., 2017) uses 2,064 episodes and a 50k gallery under the ChatIR/DAR setting (Levy et al., 2023; Long et al., 2025). We test both the DAR-preprocessed standalone descriptions and the original caption with human question–answer rounds.

Retrieval and comparison protocol. We use CLIP, BLIP-ITC, and SigLIP-2 retrieval spaces (Radford et al., 2021; Li et al., 2022; Tschannen et al., 2025), plus OpenCLIP ViT-H/14 (Cherti et al., 2023). On FS-COCO-Dialog, *ChatIR* denotes the concatenated dialogue in a frozen encoder, and ChatIR-FT a fine-tuned BLIP encoder. FS-COCO-Dialog text queries pass through a linear text map selected on the 3k evaluation targets, the development cohort of the appendix (Appendix H.4), whereas Section 5.1 fits and selects all maps on training targets, which moves CLIP concatenation from 29.2 to 28.8. PlugIR reformulates the shared dialogue, and its candidate-aware questioning loop is excluded when observations must remain fixed. DAR and CAST adaptations are detailed in Appendix H. Unless specified, results report state-ten H@10. Confidence intervals use 2,000 paired

query-bootstrap resamples, while seed variability is assessed separately. All differences are computed from unrounded per-query results and may deviate by 0.1 points from subtracting rounded entries.

4 Headroom Within the Observed Dialogue

4.1 Historical Gains Without New Observations

We share turns 0–4 across all continuations, then either retain the turn-four query, perturb it independently, or reorder its frozen top-100 shortlist. Query perturbation uses

$$\widetilde{q}_t = \frac{q_4 + \epsilon_t}{\|q_4 + \epsilon_t\|_2}, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2 I), \quad t = 5, \ldots, 10. \tag{2}$$

Every draw starts from q_4. Neither target-blind control receives further target observations, and shortlist reordering makes no additional retrieval-model call.

Table 1 shows that historical gains can accompany substantial ranking deterioration. From state four to ten, 47 targets outside the state-four top ten enter it while 98 state-four hits leave, producing only a -1.7 point net change. A nearly flat curve therefore need not reflect a stable retrieved set.

	H@10 at state ten		Change from frozen (pp)	
Continuation	Current	Cumulative	ΔCurrent	ΔCumulative
Frozen at turn 4	30.9	39.7	–	–
Dialogue continuation	29.2	40.5	-1.7	$+0.8$
Query perturbation ($\sigma = .05$)	14.2	45.9	-16.7	$+\mathbf{6.2}$
Shortlist reordering	6.0	**50.5**	-24.9	$+\mathbf{10.8}$

Table 1: **Historical credit can improve without a stronger retrieval state.** State-ten H@10 on FS-COCO-Dialog with CLIP. All continuations share turns 0–4, and only dialogue continuation receives further target observations.

The target-blind controls sharpen the distinction. Query perturbation gains 6.2 cumulative points while losing 16.7 current points, with paired 95% intervals of $[+5.4, +7.1]$ and $[-18.1, -15.2]$. Shortlist reordering raises cumulative H@10 from 39.7 to 50.5 while current H@10 falls from 30.9 to 6.0, and 88.1% of credited historical successes are absent from the current ranking. BRI likewise improves while ALRI deteriorates (Figure 3). Historical improvement therefore does not imply a stronger retrieval state. Repeated exposure alone raises it, since each uniform shuffle places a shortlisted target in the top ten with probability 0.1 (Appendix A). The trend persists across five seeds, multiple query constructions and retrieval spaces, and both datasets. Reordering lowers current retrieval in all eleven tested settings and exceeds the real cumulative endpoint in eight (Table 5).

4.2 Query Construction Under Fixed Observations

We fix all observed units and vary only their representation. Concatenation uses $q_t^{\text{concat}} = \phi_T(d_0 \oplus \cdots \oplus d_t)$. For independently encoded, normalized units e_k, mean pooling uses $q_t^{\text{mean}} \propto \sum_{k=0}^{t} e_k$, while redundancy-weighted pooling downweights units similar to earlier observations:

$$q_t^{\text{rw}} \propto \sum_{k=0}^{t} w_k e_k, \qquad w_0 = 1, \quad w_k = 1 - \max_{j<k} \cos(e_k, e_j). \tag{3}$$

We also include a PlugIR-style target-blind reformulation by Qwen3-VL-32B. All non-oracle constructions therefore receive exactly the same observations.

Changing query construction produces large retrieval differences despite fixed observations (Table 2). On FS-COCO-Dialog, reformulation improves over concatenation by 14.5–29.9 H@10 points across three retrieval spaces. On DAR-preprocessed VisDial, mean pooling raises H@10 from 62.2 to 74.9 under ChatIR-FT and from 51.6 to 79.9 under zero-shot BLIP-ITC. Across the six positive settings, paired gains range from $+12.7$ $[+11.0, +14.3]$ to $+29.9$ $[+28.1, +31.7]$, with all intervals excluding zero. The same direction holds at R@1, R@5, MRR, and median rank (Appendix C.1, Table 8).

	FS-COCO-Dialog			VisDial, raw		VisDial, preprocessed	
Query construction	CLIP	BLIP-ITC	SigLIP-2	ChatIR-FT	ZS BLIP-ITC	ChatIR-FT	ZS BLIP-ITC
Concatenation	29.2	27.6	41.1	**75.8**	44.2	62.2	51.6
Mean pooling	28.6	36.0	40.1	54.7	30.0	**74.9**	**79.9**
Weighted pooling	31.8	38.1	43.5	64.6	53.2	70.9	79.0
Reformulation	**43.7**	**57.5**	**67.3**	73.4	**72.6**	–	–
Δ vs. concat.	+14.5	+29.9	+26.2	−2.5	+28.4	+12.7	+28.3
Best-unit oracle†	48.3	58.1	60.4	69.9	74.9	85.9	89.4

Table 2: **The same observations support substantially different retrieval states.** Current H@10 at state ten. Non-oracle rows are target-blind and use identical dialogue. Bold marks the best of them. Δ: best alternative minus concatenation. † selects the best-ranked observed unit (diagnostic only).

Effect of truncation. The VisDial gains are not explained solely by context truncation. Among the 230 episodes whose state-ten concatenation never exceeds the 256-token limit of the BLIP tokenizer shared by both encoders, mean pooling still improves H@10 from 59.6 to 71.3, and zero-shot BLIP-ITC from 52.2 to 78.7. Truncation may contribute, but cannot account for the observed representation headroom by itself. On FS-COCO-Dialog, concatenation exceeds the CLIP and SigLIP-2 text contexts, so ten further turns leave it unchanged (Table 18), yet reformulation still exceeds weighted pooling, which encodes every unit, by 12–24 points.

The effect is not universal. On raw human VisDial under ChatIR-FT, concatenation reaches 75.8 H@10 versus 73.4 for the strongest tested target-blind alternative. Under the same encoder, pooling helps preprocessed standalone descriptions but hurts the raw caption–QA sequence, possibly because short human answers lose their context when encoded independently. The preferred construction thus depends on the encoder–input pairing, and a representation control must be evaluated on the actual observation format.

4.3 Retrieval Value Beyond the Aggregate State

A diagnostic best-unit oracle selects the independently encoded observed unit with the best target rank. It reaches 48.3, 58.1, and 60.4 H@10 on FS-COCO-Dialog under CLIP, BLIP-ITC, and SigLIP-2, and 85.9 and 89.4 on preprocessed VisDial (Table 2). These values identify successful units already present among the observations, but are neither deployable selectors nor upper bounds on arbitrary query representations.

The tested aggregate constructions reduce much of this aggregate score gap without reproducing the oracle's retrieved set. On FS-COCO-Dialog, reformulation reduces the concatenation-to-oracle H@10 gap by 76% and 98% under CLIP and BLIP-ITC, while under SigLIP-2 it exceeds the single-unit oracle by 6.9 points. Under CLIP it still misses 408 targets that some observed unit retrieves while retrieving 270 that no single unit does (Appendix C). On preprocessed VisDial, mean pooling reduces the corresponding gap by 54% and 75%. For a trajectory of separately encoded unit queries, cumulative H@10 equals oracle success because both count whether any observed unit retrieves the target. Meanwhile, the final three mean-pooled turns add only 2.5 points on VisDial, and weighted pooling adds 0.6 on FS-COCO-Dialog. Small late-turn gains can therefore coexist with substantial unused retrieval value in observations already acquired.

5 Headroom Under Additional Target Access

A privileged recovery probe first verifies that dialogue failures remain retrievable in the same gallery and embedding space. Among 349 targets never retrieved by DAR in its top ten, mean pooling reaches 8.0 H@10 and the observed-turn oracle 33.2, while target-informed spatial and appearance text reach 56.2 and 77.7. Because this subset is conditioned on DAR failure and the probe changes both access and wording, we use it only to establish recoverability (full results in Appendix D).

5.1 Rewriting Versus Additional Evidence

We ask whether additional-evidence gains remain after representation headroom has been reduced. On the same 3,000 FS-COCO-Dialog targets, a Qwen3-VL-32B rewrite uses only facts already stated in the dialogue, while a human sketch provides a target-grounded observation. A target-informed appearance description gives a privileged reference. We add each source to concatenated, pooled, and PlugIR-reformulated states.

Interfaces are fit on 5,000 training targets and selected on a disjoint 1,000. No checkpoint or fusion weight is selected on the 3,000 evaluation targets. An additional 1,000 training targets on which no component of this protocol was fit or selected provide a held-out replication (Appendix D.3). All added sources use $\alpha = .5$.

Figure 2 shows the key interaction. From ChatIR, rewriting and sketch fusion add 14.0 and 13.4 points. After PlugIR reformulation, another rewrite adds only $+0.7$, whereas the sketch adds 11.1. The same pattern holds in all four retrieval spaces (Table 10). The interaction is estimated directly: the sketch-minus-rewrite difference after reformulation exceeds the same difference from concatenation by $+11.0$ $[+9.4, +12.6]$ points under CLIP and by 8.8–16.2 across the four spaces, with every paired interval excluding zero (Appendix D.2). Thus, a sketch gain over a weak representation can overlap with representation headroom, while its residual gain after reformulation provides stronger evidence of headroom from expanded target access.

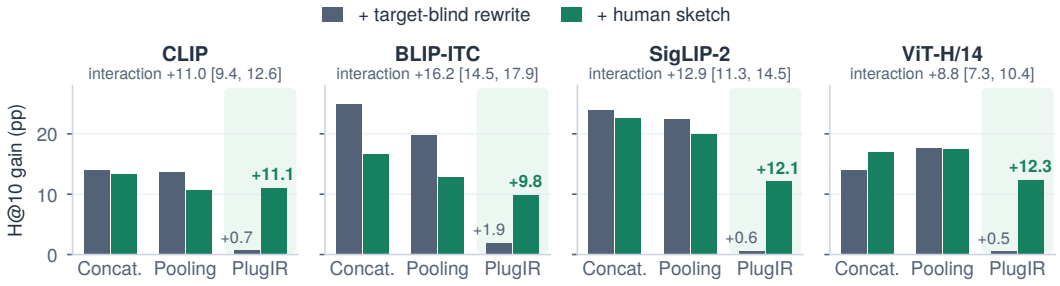

Figure 2: **Representation refinement and additional evidence expose different headroom.** Paired H@10 gain from adding a target-blind rewrite or a human sketch to each dialogue state (3,000 FS-COCO-Dialog targets, held-out interface selection, $\alpha = .5$). Pooling is weighted pooling. Rewriting adds only 0.5–1.9 after PlugIR, while the sketch still adds 9.8–12.3. Subtitles give the interaction contrast with paired 95% intervals. Absolute values are in Table 10.

The interaction also persists across dialogue prefixes. Rewrite gains over ChatIR grow from $+5.6$ at state two to $+14.0$ at state ten but remain $+0.5$–$+0.8$ after PlugIR, while sketch gains remain $+10.5$–$+14.1$. On an untouched cohort, rewriting changes ChatIR from 27.9 to 42.0 and PlugIR from 44.3 to 43.9, while sketch fusion reaches 40.6 and 54.7. Wrong-target controls reverse the gain (Appendix D.5).

The sequential control makes this distinction more direct. Starting from ChatIR, a rewrite reaches 42.7 H@10. Adding a sketch to that rewritten state reaches 52.4, while adding a second sampled rewrite reaches 43.7, a paired difference of $+8.7$ $[+7.2, +10.2]$. On the untouched cohort, the corresponding endpoints are 42.0, 52.0, and 45.9 ($+6.1$ $[+3.3, +8.8]$), so a second rewrite is not always negligible from a concatenated state (Appendix D.3). The sketch nevertheless retains value when both routes start from the same strengthened state. This supports residual evidence-access headroom beyond the tested rewriting procedure, rather than attributing the entire sketch gain to information unavailable in the dialogue.

Raw human VisDial. We repeat the matched comparison on the same 2,064 human VisDial episodes and fixed 50k gallery under ChatIR-FT BLIP and zero-shot BLIP-ITC.

Under ChatIR-FT, concatenation is already the strongest tested base at 75.8 H@10. Another rewrite raises it only to 77.9, whereas target-informed appearance text reaches 95.1. The same separation remains after PlugIR: rewriting changes 73.4 to 75.0 while appearance text reaches 94.3. Under zero-shot BLIP-ITC, rewriting substantially improves weak states, but after PlugIR it adds only 3.2

points, from 72.6 to 75.8, whereas appearance text reaches 96.1. A second sampled rewrite reproduces the first within 0.4 points (Table 13). Residual evidence-access headroom therefore persists on human dialogue after further tested target-blind refinement. Because the privileged description changes both target access and wording, this establishes recoverability under expanded access rather than a causal effect of specific previously unobserved facts.

5.2 Sketch Evidence Across Retrieval States

We align each paired human sketch to the frozen retrieval space using a bias-free map $W \in \mathbb{R}^{d\times d}$ trained on 7k sketch–photo pairs. The retrieval towers remain frozen. For base query q_T and sketch feature x,

$$s(g) = (1-\alpha)\cos(q_T, \phi_I(g)) + \alpha\cos(Wx, \phi_I(g)), \qquad \alpha = .5. \tag{4}$$

We fix $\alpha = .5$ across pipelines and backbones (details in Appendix H). The matrix below uses evaluation-selected text and sketch checkpoints and is a breadth analysis. The residual-headroom claim rests on the held-out-selection comparison of Section 5.1, replicated across four retrieval spaces and an untouched cohort.

Direct sketch features improve all twelve sketch-unaware pairs by 8.3–24.2 points (Table 3), and PlugIR with SigLIP-2 rises from 67.3 to 80.7. Paired 95% intervals exclude zero for all fifteen comparisons shown, including the CIReVL-style controls, where direct sketch features add 8.8–11.6 points. OpenCLIP and task-trained results are reported in Appendix G.3.

	CLIP		BLIP-ITC		SigLIP-2	
Dialogue state	Base	+ Sketch	Base	+ Sketch	Base	+ Sketch
ChatIR (Levy et al., 2023)	29.2	43.3	27.6	48.1	41.1	65.3
Weighted pooling	31.8	41.8	38.1	52.1	43.5	64.1
DAR adaptation (Long et al., 2025)	27.3	35.6	31.1	40.3	37.8	52.8
PlugIR reform. (Lee et al., 2024)	43.7	54.8	57.5	69.5	67.3	**80.7**
Sketch-aware control: base already uses a caption of the drawing						
CIReVL-style† (Karthik et al., 2024)	44.5	53.3	58.8	67.7	66.7	78.3

Table 3: **Sketch gains persist across retrieval states and backbones.** Direct sketch features improve all twelve sketch-unaware pairs by 8.3–24.2 H@10 points and remain beneficial in the sketch-aware control. Evaluation-selected checkpoints (held-out results in Table 10).

The result is also robust to training and alignment: local ChatIR-FT improves 58.8→67.2 and CAST 52.5→64.2, four OpenCLIP pairs gain 10.4–18.8 points, three alignment seeds change gains by at most 0.4, and artist-held-out and Sketchy-trained maps reach 42.5 and 37.6 ± 0.5 versus 43.3 (Appendix G.4). Comparisons with DAR, Pic2Word, SEARLE, LinCIR, CIReVL-style, and Qwen3-VL reranking (Appendix G.1) differ in evidence access, supervision, and candidate scope, so we treat them as interface comparisons rather than matched architectural rankings.

5.3 Headroom After More Interaction

We extend every FS-COCO-Dialog dialogue to twenty turns with the same simulator families while keeping turns 1–10 unchanged.

The best-unit oracle gains 5.9 points, while weighted pooling and PlugIR gain only 1.7 and 2.2 (Table 4). The same sketch still adds 8.9 and 10.0 points at turn twenty. Under BLIP-ITC and SigLIP-2, the oracle gains 6.2 and 6.7 points while their reformulated states lose 3.1 and 1.3, yet sketch fusion still adds 11.6 and 14.6 points (Appendix E). Of the 177 targets whose first successful unit lies in turns 11–20 under CLIP, the state-twenty reformulated state retrieves 81 and concatenation 11, while the sketch fused at state ten retrieves 77 of them without any further turn. A best-of-eight oracle over sampled reformulations of the unchanged ten-turn dialogue already gains 3.7–4.6 points at temperature .7 and 4.8–5.9 at 1.0 (Appendix E), whereas the deployed states gain at most 2.2 points from ten real turns and the sketch fused at state ten adds 11.0. More interaction thus does not exhaust the observed sketch headroom.

State	$t = 10$	$t = 20$	$\Delta_{10\to20}$	+ Sketch at $t = 20$
Weighted pooling	31.8	33.4	$+1.7$	$+\mathbf{8.9}$
PlugIR reform.	43.7	45.9	$+2.2$	$+\mathbf{10.0}$
Best-unit oracle†	48.3	54.2	$+\mathbf{5.9}$	–

Table 4: **Single-unit oracle success grows more than aggregate-state success over ten more turns.** CLIP current H@10 after extending the same episodes from ten to twenty turns. † uses target ranks.

A finite best-of-eight next-question probe provides a complementary test. Target-rank selection raises H@10 from 29.9 to 34.2 and from 29.7 to 35.4 at state four, and from 32.5 to 36.4 for weighted pooling at state ten, whereas the tested target-blind, post-answer novelty selector recovers none of this headroom in the diversified candidate pool. Under concatenation at state ten, the added answer falls outside CLIP's context window and all alternatives remain at 29.2 H@10. Question choice therefore leaves measurable headroom, but exploiting it also requires a representation capable of incorporating the resulting answer (Appendix F).

5.4 Complementarity and Boundary Conditions

Recovery must also be distinguished from preservation. In the matched CLIP comparison, adding the correct sketch to the reformulated state recovers 445 targets but loses 111 earlier hits (Appendix D.7). The net gain thus hides per-query losses. Replacing it with a category-matched wrong-target sketch lowers H@10 from the 42.9 base to 30.2, versus 54.0 with the correct sketch. A randomly permuted sketch lowers it further to 23.6 (Appendix D.5). The benefit depends on correspondence to the target beyond broad category overlap, while these controls do not isolate which individual visual details were missing from the dialogue.

The sketch gain is not explained solely by selecting whichever source succeeds. For ChatIR–CLIP, source selection retrieves 1,236 of 3,000 targets, while joint ranking retrieves 1,299 by rescuing 235 targets missed by both sources and losing 172, a net $+2.1$ points $[+0.8, +3.4]$. These rescues are 13.3% of queries missed by both sources. Across sixteen configurations, fourteen margins are positive and twelve paired intervals exclude zero (Appendix D, Table 17).

The benefit is not universal. Weighted pooling with CLIP reaches 41.8 under joint ranking, below its 43.1 source-selection oracle. Added evidence can also hurt: severe sketch degradation reduces H@10 to 19.5, high sketch weight destroys 56–59% of existing top-ten successes, and wrong-target controls reverse the gain (Appendices G.5, D.6, and D.5). Expanded target access thus exposes recoverable headroom, but its value depends on evidence correspondence, quality, and integration.

These results suggest evaluating a continuation against three references: the current state, a strengthened representation of the same observations, and that strengthened state with the proposed evidence, reporting recovered and lost targets alongside current and cumulative success. A saturation claim is then specific to an observation regime and a tested family of retrieval states.

6 Conclusion

We study retrieval headroom beyond a flattened dialogue curve. Historical success can increase without another observation, the same dialogue can support a stronger query, and additional target access can recover residual failures. Longer interaction exposes additional successful units faster than deployed states capture them, while substantial sketch headroom remains. A finite next-question probe further shows that question choice can leave recoverable headroom, but exploiting it depends on both evidence acquisition and representation. Limited late-turn progress therefore does not establish an information ceiling. Evaluating saturation requires separating historical from current retrieval, testing query construction under fixed observations, and measuring additional evidence against those strengthened states. Our evidence is limited to generated FS-COCO-Dialog answers of uncertain fidelity, human sketches from FS-COCO alone, and a fixed linear fusion rule.

AI USE STATEMENT

Generative AI tools were used in two roles in this work. The vision–language models specified in Appendix H were used as experimental components to generate the FS-COCO-Dialog dialogues, reformulations, dialogue-only rewrites, privileged descriptions, and automated fidelity labels. These generated outputs are themselves objects of study. An AI coding assistant was additionally used to provide feedback on experimental controls, assist with implementation and analysis scripts, verify the authors' elementary derivations, and assist with manuscript drafting and editing. The authors made the experimental and methodological decisions, reviewed all AI-assisted code and derivations, regenerated reported results from recorded per-query ranks, and independently verified the reported claims. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Metric definitions and the derivations behind Section 3.2 are given in Appendix A. Benchmark construction and the questioner, answerer, reformulation, rewrite, description, and judge prompts, together with model versions and decoding settings, are given in Appendix H, as are the training recipe for the sketch and text maps, the held-out interface-selection protocol, the fixed fusion weight, and the statistical protocol. All paired comparisons share episode identifiers. Confidence intervals use 2,000 paired bootstrap resamples, and seed variability is reported separately. Per-query ranks are recorded for every reported state, so every table and paired estimate can be regenerated from them. Code is available at `https://anonymous.4open.science/r/iclr2027-DA65/`.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3952–3962. IEEE, 2025b.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15292–15301, 2023.

Daniel D. Braghis and Haiming Liu. Conversational image search: A sketch-based approach. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*, pp. 1265–1269, 2024.

Xianke Chen, Daizong Liu, Yushuo Lou, Xin Tan, Xun Yang, Shuhui Wang, Xun Wang, and Jianfeng Dong. Memory-augmented query intent understanding for efficient chat-based image retrieval. *arXiv preprint arXiv:2605.17365*, 2026.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Pinaki Nath Chowdhury, Aneeshan Sain, Yulia Gryaditskaya, Ayan Kumar Bhunia, Tao Xiang, and Yi-Zhe Song. FS-COCO: Towards understanding of freehand sketches of common objects in context. In *European Conference on Computer Vision (ECCV)*, 2022.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1080–1089, 2017.

Emil Demić and Luka Čehovin Zajc. Back to the drawing board: Rethinking scene-level sketch-based image retrieval. In *British Machine Vision Conference (BMVC2025)*, 2025.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13225–13234, 2024.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *International Conference on Learning Representations (ICLR)*, 2024.

Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. You'll never walk alone: A sketch and text duet for fine-grained image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16509–16519, 2024.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 791–809, 2024.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 61437–61449, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 12888–12900, 2022.

Xuanzuo Lin, Min Zhang, Daizong Liu, Zhiwen Zuo, Xun Yang, Changting Lin, Xun Wang, and Jianfeng Dong. CAST: Context-aware dynamic latent space transformation for interactive text-to-image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38794–38803, 2026.

Zijun Long, Kangheng Liang, Gerardo Aragon Camarasa, Richard McCreadie, and Paul Henderson. Diffusion augmented retrieval: A training-free approach to interactive text-to-image retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 823–832, 2025.

Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. Human-centered interactive learning via mllms for text-to-image person re-identification. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14390–14399. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2Word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19305–19314, 2023.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The Sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics*, 35(4):1–12, 2016.

Patsorn Sangkloy, Wittawat Jitkrittum, Diyi Yang, and James Hays. A sketch is worth a thousand words: Image retrieval with text and sketch. In *European Conference on Computer Vision (ECCV)*, pp. 251–267, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision (ECCV)*, pp. 87–103, 2024.

Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14400–14410, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Shuaiwei Xie, Yating Yang, Bo Ma, Xi Zhou, Zhen Wang, and Ahtamjan Ahmat. Ask-to-Retrieve: VQA-guided information-gain question selection for interactive text-to-image retrieval. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*, pp. 1740–1748, 2026.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Zhuocheng Zhang, Xingwu Zhang, Kangheng Liang, Guanxuan Li, Richard McCreadie, and Zijun Long. ADaFuSE: Adaptive diffusion-generated image and text fusion for interactive text-to-image retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026.

A Historical Credit in Multi-Turn Metrics

A.1 Current Success, Cumulative Success, and the Masked Gap

For episode i, let r_{it} be the target rank at state t and let $E_{it} = \{r_{it} \le K\}$. The current and cumulative rates are

$$I_t = \Pr(E_t), \qquad C_t = \Pr\left(\bigcup_{u=0}^{t} E_u\right).$$

Because $E_t \subseteq \cup_{u \le t} E_u$, their difference is

$$G_t = C_t - I_t = \Pr\left(\left(\bigcup_{u=0}^{t} E_u\right) \setminus E_t\right).$$

Thus G_t counts episodes that succeeded at least once but are not successful at the current state. A zero gap is not, by itself, evidence of good retrieval: a system that never succeeds also has $G_t = 0$. We therefore report the gap alongside current and cumulative success rather than using it as a standalone quality metric.

For MHR@10, defined as the average of current and cumulative H@10,

$$\mathrm{MHR}_t = \frac{I_t + C_t}{2} = I_t + \frac{G_t}{2}.$$

This identity shows that MHR retains half of the historical/current disagreement. It does not imply that MHR is inappropriate for every application.

A.2 A Cumulative Curve Does Not Determine Current Retrieval

For any non-decreasing sequence $C_0, \ldots, C_T$ in $[0, 1]$, assign a fraction C_0 of episodes a first hit at state zero and a fraction $C_t - C_{t-1}$ a first hit at state t. In one construction, every successful episode remains in the top K thereafter. In a second, each episode is in the top K only at its first-hit state. Both constructions have the same cumulative curve, while the first has $I_t = C_t$ and the second has $I_t = C_t - C_{t-1}$ for $t \ge 1$. A flat cumulative segment is therefore compatible with stable current success or its complete loss. This construction establishes non-identifiability of current retrieval from cumulative success alone. It does not identify the semantic causes of an empirical plateau or prove that a finite set of probes exhausts representation and acquisition headroom.

A.3 Matched Trajectory Summaries

Let $b_{it} = \min_{u \le t} r_{iu}$ and $\ell(r) = \ln(1 + r)$. The reported running-best integral follows the principle of PlugIR, with the explicit transformation and normalization used here

$$\mathrm{BRI} = \mathbb{E}_i\left[\frac{\ell(b_{i0}) + \ell(b_{iT})}{2T} + \frac{1}{T}\sum_{t=1}^{T-1} \ell(b_{it})\right].$$

Our ALRI uses the same transformation and trapezoidal weights but replaces the running best rank with the current rank:

$$\mathrm{ALRI} = \mathbb{E}_i\left[\frac{\ell(r_{i0}) + \ell(r_{iT})}{2T} + \frac{1}{T}\sum_{t=1}^{T-1} \ell(r_{it})\right].$$

The comparison therefore isolates one operation: BRI replaces each current rank with the best rank seen so far, whereas ALRI leaves the trajectory unchanged. Lower is better for both.

A.4 Repeated Opportunities and First-Hit Success

Suppose an episode has not succeeded before state t_0, and later target-blind attempts are conditionally independent with fixed per-attempt success probability p_i. After m attempts,

$$\Pr(\text{at least one later hit}) = 1 - (1 - p_i)^m,$$

even though the accuracy of any individual attempt remains p_i. This independence assumption is used only to illustrate the mechanism. Cumulative success is non-decreasing without it, by set inclusion.

For the top-100 control in the main paper, the target's position within the frozen shortlist is randomized uniformly. If the target is in that shortlist, one shuffle places it in the top ten with probability 0.1. If it is outside the shortlist, the shuffle cannot create a hit. No target label is used to decide which candidate to promote.

Let S_i indicate that target i belongs to the frozen shortlist and H_i indicate a hit by state four. For a top-M shortlist and independent uniform reorderings, the conditional expected cumulative endpoint is

$$\mathbb{E}[C_{4+m} \mid \{H_i, S_i\}] = C_4 + \frac{1}{N} \sum_i (1 - H_i) S_i \left[1 - (1 - K/M)^m\right].$$

For $M = 100$, $K = 10$, and $m = 6$, the conditional later-hit probability is $1 - 0.9^6$. This expression requires no independence between shortlist membership and earlier success. Evaluating an empirical prediction additionally requires their joint per-query indicators.

B Stress Tests Beyond the Main Setting

B.1 Perturbation and Shortlist Controls

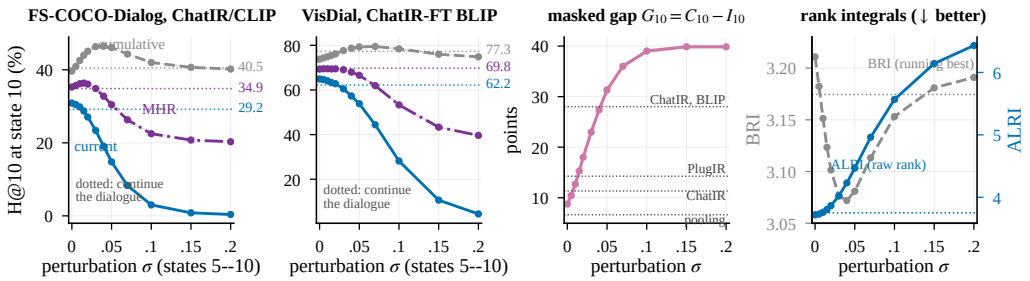

Figure 3: Target-blind perturbation on FS-COCO-Dialog and VisDial. Historical success can improve as current retrieval deteriorates. Dotted lines denote dialogue continuation. Bands show ± one standard deviation over five seeds.

Robustness across retrieval states. The trend persists across five seeds, with standard deviation below 0.8 H@10 points. On FS-COCO-Dialog, perturbations increase cumulative success over the frozen state across ChatIR, pooled, and reformulated queries while reducing current retrieval. On VisDial, perturbed ChatIR reaches 79.5 cumulative H@10 at $\sigma = .07$, above the real continuation's 77.3, while current H@10 falls to 44.4. Shortlist reordering reaches 82.2 cumulative H@10. Across eleven settings, reordering has lower current retrieval in all eleven and exceeds the real cumulative endpoint in eight (Table 5). The separation is therefore not specific to one dataset or query construction.

All post-prefix perturbations start from the same state-four query. The prefix, target cohort, and gallery are unchanged. Gaussian perturbations are centered on that query independently at each later state, and the full strength sweep is reported rather than selecting a single favorable value. The top-100 control is stronger as a falsification test because the candidate set itself is frozen and only its order changes.

On FS-COCO-Dialog, the same qualitative effect appears when the perturbed query is formed by concatenation, mean pooling, or redundancy-weighted pooling. For PlugIR and CIReVL-style, the real continuation improves current retrieval more strongly, so the perturbations do not exceed the real cumulative endpoint. They still raise cumulative success over the frozen query while lowering current retrieval. This is the comparison relevant to the paper's claim: the stress test need not beat a strong method to show that historical credit and current quality are different quantities.

The same diagnosis appears in the ChatIR VisDial trajectory. Perturbations can increase cumulative H@10 while current H@10 decreases, and the model-free shortlist control exceeds the real cumulative

endpoint in the concatenation and pooling settings tested on both datasets. These experiments test the interpretation of historical scores. Reordering may improve exposure for an inspect-and-stop user, but does not acquire another target-grounded observation.

		Frozen		Dialogue		Reordered	
Dataset / space	Query	I	C	I	C	I	C
FS-COCO / CLIP	ChatIR	30.9	39.7	29.2	40.5	6.0	50.5
	Weighted pool	28.4	31.6	31.8	38.4	5.8	45.3
	Mean pool	26.8	36.9	28.6	41.6	5.6	47.9
	PlugIR reform.	36.9	46.9	43.7	58.0	7.0	57.3
	CIReVL-style	36.5	44.5	44.5	56.4	7.2	56.3
FS-COCO / BLIP	ChatIR	35.6	51.3	27.6	55.6	7.2	60.7
FS-COCO / SigLIP-2	ChatIR	41.8	53.9	41.1	54.9	7.5	63.2
FS-COCO / CLIP	ChatIR + sketch	44.5	51.4	43.3	52.6	8.0	64.1
VisDial / ChatIR-FT	Concatenation	64.9	73.8	62.2	77.3	9.1	82.2
	Mean pool	69.0	73.4	74.9	79.2	9.5	83.0
	Latest turn	65.5	77.7	71.3	85.9	9.1	84.5

Table 5: **Shortlist reordering separates exposure from current retrieval across settings.** State-ten current (I) and cumulative (C) H@10. Frozen and reordered continuations receive no further observations after state four. Reordering independently permutes the fixed top 100 and retains all lower positions. Values are five-seed means, with reported standard deviations at most 0.7 points. FS-COCO denotes FS-COCO-Dialog and VisDial uses ChatIR-FT BLIP.

B.2 A Simulated Inspection Objective

The inspection model $U(p, K)$ reviews each screen with probability p, with $K = 10$. A reviewed screen is scanned in rank order until success or until all K images are inspected. If the target is still unfound after the dialogue, the final ranking is scanned to a give-up depth of 1,000. The implementation counts rank as the number of images above the target, so successful inspection costs that count plus one. We use one-based target positions in the main metric notation. Inspection is charged per viewed screen rather than deduplicated across the entire session. The outcomes are simulated retrieval utility under these assumptions, not measurements of human behavior.

At $p = 1$, the real dialogue retrieves 95.2% of targets at an average cost of 245 inspected images, whereas the strong perturbation retrieves 82.8% at a cost of 366. For $p = .5$, the corresponding outcomes are 95.2% at 213 images and 81.9% at 354 images. For $p = 0$, they are 95.2% at 180 images and 80.2% at 369 images. The comparison is stable across these three inspection probabilities. We use this result only to show that the perturbation's larger first-hit count does not imply a better outcome under this alternative simulated objective.

Metric	**11 pipelines**	**+ noise controls**
Cumulative H@10	+0.60	+0.43
BRI	+0.82	+0.56
Current H@10	+0.85	+0.91
ALRI	**+0.96**	**+0.98**

Table 6: **Agreement with a simulated inspection objective.** Kendall τ compares system orderings after metric directions are aligned. These values are conditional on the inspection model described in the text and are not measurements of human behavior.

The table above compares metric orderings with the same simulated cost over the recorded systems. Its values are conditional on that cost model. We do not infer that ALRI is universally preferable, nor do we use this simulation as evidence about real inspection behavior.

Construction	0	1	2	3	4	5	6	7	8	9	10
Last turn only	20.1	15.9	15.2	15.0	14.3	14.3	13.0	13.4	13.5	13.3	12.1
Full dialogue	20.1	24.8	27.1	30.1	30.9	29.9	29.4	29.2	29.2	29.2	29.2
Recent three turns	20.1	24.8	27.1	25.5	24.6	23.6	22.2	22.1	22.2	22.5	23.0
Mean pooling	20.1	23.7	24.6	26.9	26.8	27.4	27.7	27.6	28.1	28.5	28.6
Fast decay	20.1	22.6	23.4	24.7	23.5	23.7	23.1	23.6	24.2	24.6	23.3
Slow decay	20.1	23.3	24.3	26.1	26.3	26.3	26.9	26.6	27.7	27.6	27.5
Max turn score	20.1	21.9	23.1	24.6	24.9	25.1	25.3	25.6	25.8	26.3	26.3
Redundancy filtering	20.1	23.1	23.9	26.6	27.8	27.7	28.6	29.1	29.4	29.8	30.0
Weighted pooling	20.1	22.5	24.9	26.7	28.4	29.7	30.4	31.1	31.7	31.7	31.8
Learned: position	20.1	24.0	26.2	28.0	28.7	30.0	30.1	30.0	30.2	30.8	30.0
Learned: content	20.1	24.0	27.0	30.1	31.0	32.3	32.8	33.6	33.8	34.3	35.0
Learned: full	20.1	24.1	27.0	30.1	31.0	32.3	32.8	33.5	33.7	34.5	34.9
Oracle: single turn	20.1	27.6	32.3	36.5	39.0	41.2	42.9	44.8	46.2	47.3	48.3

Table 7: **Turn-wise preservation on FS-COCO-Dialog.** Current H@10 with CLIP. All non-oracle rows use the same acquired dialogue. The final row selects the best separately encoded turn using the target rank and is therefore diagnostic rather than deployable.

C Representation Headroom Under Fixed Observations

The text-only rows above use the same acquired dialogue and differ only in how its turns are encoded or aggregated. The single-turn oracle selects among separately encoded caption/QA units using the target rank. It is therefore an upper bound only for that candidate family, not for arbitrary dialogue representations.

For a fixed family of separately encoded turn queries $\{q_0, \ldots, q_T\}$, cumulative H@10 and target-informed single-turn selection have the same success event:

$$\mathbf{1}\left[\min_{t \le T} r(q_t) \le 10\right].$$

This is the sense in which running-best evaluation becomes oracle selection in the main paper. The identity does not extend to a different query family or to an online selector that lacks the target rank.

At state ten, the aggregate states and the oracle retrieve overlapping but distinct sets. Under CLIP, concatenation, weighted pooling, and reformulation miss 658/546/408 targets that some single observed unit retrieves, while retrieving 84/49/270 targets that no single unit does. The corresponding counts are 985/656/362 and 71/55/345 under BLIP-ITC and 684/558/216 and 103/50/422 under SigLIP-2. On preprocessed VisDial, mean pooling misses 230 (ChatIR-FT) and 198 (zero-shot BLIP-ITC) oracle-retrieved targets and retrieves 4 and 1 that no unit does. A reduction of the aggregate score gap therefore does not mean that the aggregate state retains the oracle's successes.

The VisDial preservation result is not explained solely by context truncation. Under ChatIR-FT, mean pooling improves current H@10 from 62.2 to 74.9 on the full cohort. Restricting the comparison to the 230 episodes whose state-ten prefix never exceeds the encoder context still gives 59.6 for concatenation and 71.3 for mean pooling. On the 1,834 truncated episodes, the corresponding gain is 12.8 points $[+11.0, +14.7]$. Both groups exhibit representation headroom, although their difference does not estimate a causal effect of truncation. We therefore avoid describing the effect simply as "forgetting." Serialization, weighting, and compatibility with the retrieval encoder can all contribute.

The same intervention is not uniformly strongest across datasets. On FS-COCO-Dialog with CLIP, mean pooling reaches 28.6 H@10, concatenation 29.2, redundancy-weighted pooling 31.8, and PlugIR reformulation 43.7. These results support preservation headroom without implying that pooling is a generally superior retrieval architecture.

C.1 Representation Gains Across Retrieval Metrics

Table 8 repeats the fixed-observation comparison using several rank-sensitive metrics. The comparison remains paired within each encoder–input setting: concatenation and the reported target-blind alternative receive exactly the same observations, targets, and gallery.

Setting	Encoder	Query construction	R@1	R@5	H@10	MRR	medR
FS-COCO-Dialog	CLIP	Concatenation	10.1	22.9	29.2	16.7	50
		Target-blind reformulation	19.6	35.6	43.7	27.7	17
FS-COCO-Dialog	BLIP-ITC	Concatenation	8.2	20.4	27.6	14.7	52
		Target-blind reformulation	27.7	48.4	57.5	37.8	6
FS-COCO-Dialog	SigLIP-2	Concatenation	17.6	33.4	41.1	25.6	21
		Target-blind reformulation	38.8	59.4	67.3	48.6	3
VisDial, raw	ChatIR-FT	Concatenation	44.4	67.4	75.8	54.9	2
		Target-blind reformulation	42.0	65.4	73.4	52.7	2
VisDial, raw	ZS BLIP-ITC	Concatenation	17.5	35.8	44.2	26.6	15
		Target-blind reformulation	40.9	63.2	72.6	51.6	2
VisDial, preprocessed	ChatIR-FT	Concatenation	31.4	52.4	62.2	41.8	4
		Mean pooling	41.8	65.7	74.9	52.8	2
VisDial, preprocessed	ZS BLIP-ITC	Concatenation	21.3	41.9	51.6	31.2	9
		Mean pooling	48.5	72.5	79.9	59.3	2

Table 8: **Fixed-observation representation gains are not specific to H@10.** Current retrieval at state ten under the same seven encoder–input settings as Table 2. In all six settings with positive H@10 headroom, the best tested target-blind alternative also improves R@1, R@5, MRR, and median rank. MRR is reported in percent and medR is the one-based median target rank.

In all six positive settings, the target-blind alternative improves every reported retrieval metric. The exception is also consistent across metrics: under ChatIR-FT on raw human VisDial, reformulation is worse than the concatenated state. The fixed-observation result therefore does not depend on the H@10 cutoff.

D Evidence Access and Source Complementarity

D.1 Privileged Descriptions as Accessibility Probes

	Residual cohort			Reference cohort
Query	H@10 ↑	Δ (pp)	medR ↓	H@10 ↑
Concatenation	2.6	−5.4	105	72.3
Latest turn	14.6	+6.6	47	81.7
Mean pooling	8.0	–	45	88.0
Best turn†	33.2	+25.2	19	95.0
DAR	0.0	−8.0	79	80.7
Spatial description‡	56.2	+48.2	7	80.3
Appearance description‡	**77.7**	**+69.7**	**2**	92.3

Table 9: **Expanded target access recovers failures beyond dialogue representations.** VisDial with ChatIR-FT and a fixed 50k gallery. The residual cohort contains 349 targets missed by DAR at every turn, so its zero is by construction. The reference cohort samples 300 of the remaining 1,715 episodes. Δ is relative to mean pooling within the residual cohort. † uses target ranks for turn selection. ‡ uses descriptions generated with direct target-image access, changing both content and wording. These cohorts support a conditional recoverability analysis, not a population-level method ranking.

The 349-target VisDial cohort in the main paper is selected by failure of DAR over its trajectory. DAR's zero on this subset is therefore true by construction, and the other percentages should not be interpreted as population-level method comparisons. The privileged descriptions also change

both target information and wording. Their role is narrower: they demonstrate that many method-conditioned failures remain retrievable in the same gallery when the query has direct target access.

The rightmost column of Table 9 is a separate random sample of 300 of the remaining episodes. It is not the complement of the 349-target subset in a partition-based estimate. The privileged-text and human-sketch experiments also use different cohorts, so their effect sizes are not compared as if they formed a matched modality study.

D.2 Matched Evidence Routes on a Common Cohort

Figure 2 uses the same 3,000 FS-COCO-Dialog queries, CLIP space with the text adapter, and 50k gallery as the main sketch comparison. Qwen3-VL-32B-Instruct (FP8) generates all three text probes. The dialogue-only prompt is: "Rewrite the dialogue about an image into ONE description of the image for retrieval, in at most 25 words. Include only visual facts stated in the dialogue; add nothing." Its inputs are the initial caption and ten QA turns, without the target image.

Target-informed prompts request one sentence of at most 25 words. The spatial instruction is: "Describe ONLY the spatial layout of this photo: where each main object sits in the frame, how many of each there are, which way they face, and how they overlap; no colours, materials, textures, lighting or mood." The appearance instruction is: "Describe ONLY the appearance of this photo: colours, materials, textures, lighting, weather, style and mood; no information about where anything is, no counts, no spatial relations." Actual mean lengths are 28.9 words for spatial text, 28.4 for appearance text, and 27.6 for the rewrite. The requested word limit is therefore not a hard generation cap.

With normalized source features, adding one source s to a state q uses $\mathrm{normalize}(.5q + .5s)$. The sequential routes ("rewrite, then sketch" and "rewrite, then second rewrite") apply this rule twice, $\mathrm{normalize}(.5\,\mathrm{normalize}(.5q + .5s_1) + .5s_2)$, so the second source receives a larger effective weight than the first. The first rewrite is decoded deterministically, while the second and third are sampled at temperature .7 (seeds 1 and 2) from the same prompt on the original dialogue, not on the first rewrite's output. Every rewrite is added to the base state rather than substituted for it.

Interaction estimate. The representation–evidence interaction is the paired per-query contrast $(H_{\text{sketch}} - H_{\text{rewrite}})_{\text{PlugIR}} - (H_{\text{sketch}} - H_{\text{rewrite}})_{\text{base}}$, resampled over the same episode identifiers for all four conditions (2,000 resamples). Against the ChatIR base it is $+11.0$ $[+9.4, +12.6]$ (CLIP), $+16.2$ $[+14.5, +17.9]$ (BLIP-ITC), $+12.9$ $[+11.3, +14.5]$ (SigLIP-2), and $+8.8$ $[+7.3, +10.4]$ (ViT-H/14). Against weighted pooling, it is $+13.5$ $[+11.9, +15.2]$, $+15.0$ $[+13.3, +16.6]$, $+13.9$ $[+12.2, +15.6]$, and $+12.0$ $[+10.4, +13.6]$. On the untouched cohort under CLIP the two contrasts are $+12.2$ $[+9.4, +15.3]$ and $+14.5$ $[+11.5, +17.3]$. This establishes that the sketch-minus-rewrite difference depends on the base state. The contrast is not a causal decomposition of semantic information.

Held-out interface selection. To remove evaluation-split checkpoint selection from the matched comparison, we split the 7,000 FS-COCO-Dialog training targets into 5,000 fitting targets, 1,000 selection targets, and 1,000 held-out targets. The text and sketch maps are fit on the first split and selected on the second, with the fusion weight fixed at $\alpha = .5$. No checkpoint or fusion weight is selected on the 3,000-query development cohort. We evaluate the same comparison in CLIP, BLIP-ITC, SigLIP-2, and OpenCLIP ViT-H/14 retrieval spaces.

The same interaction appears in every retrieval space. From the ChatIR concatenated state, rewriting the observed dialogue adds 14.0–25.0 H@10 points, comparable in scale to the sketch. After PlugIR reformulation, however, another target-blind rewrite adds only 0.5–1.9 points, whereas the sketch still adds 9.8–12.3. Thus a sketch gain over concatenation can overlap with representation headroom, while the gain remaining after reformulation is not reproduced by another tested target-blind rewrite.

Interaction across dialogue states. The endpoint comparison above can also be tracked over dialogue prefixes. Table 11 reports the paired gain from the same dialogue-only rewrite and human sketch at states 2, 4, 7, and 10.

Across states two through ten, the rewrite gain over ChatIR increases from $+5.6$ to $+14.0$ points and its gain over weighted pooling from $+8.6$ to $+13.7$, whereas its gain after PlugIR reformulation remains within $+0.5$–$+0.8$. The sketch gain instead remains between $+10.5$ and $+14.1$ points across

	CLIP			BLIP-ITC			SigLIP-2			ViT-H/14		
Added source	ChatIR	Pool.	PlugIR	ChatIR	Pool.	PlugIR	ChatIR	Pool.	PlugIR	ChatIR	Pool.	PlugIR
Dialogue state alone	28.8	31.3	42.9	30.3	40.0	59.5	40.8	43.6	67.5	38.9	35.8	55.5
Dialogue-only rewrite	42.7	45.0	43.6	55.3	59.9	61.4	64.8	66.0	68.1	52.8	53.4	55.9
Human sketch	42.2	41.9	54.0	47.0	52.8	69.3	63.5	63.6	79.6	55.8	53.3	67.8
Rewrite, then sketch	52.4	52.9	54.2	65.5	67.5	70.1	78.6	78.7	79.9	65.5	65.9	67.9
Privileged appearance	62.7	67.7	66.7	86.5	89.9	89.5	90.8	92.5	92.8	80.1	82.7	83.3
Paired gain over the same dialogue state												
Rewrite	+14.0	+13.7	+0.7	+25.0	+19.8	+1.9	+24.0	+22.4	+0.6	+14.0	+17.6	+0.5
Sketch	+13.4	+10.7	+11.1	+16.6	+12.8	+9.8	+22.7	+20.0	+12.1	+17.0	+17.5	+12.3
Sketch − rewrite	−0.5	−3.1	+10.5	−8.3	−7.1	+7.9	−1.3	−2.4	+11.5	+3.0	−0.1	+11.8

Table 10: **The representation–evidence interaction persists across retrieval spaces.** Current H@10 at state ten on the same 3,000-query development cohort under held-out interface selection. A target-blind rewrite yields large gains from concatenated states but only 0.5–1.9 points after PlugIR reformulation, whereas the target-grounded sketch still yields 9.8–12.3 points.

	Dialogue-only rewrite gain			Human sketch gain		
State	ChatIR	Pool.	PlugIR	ChatIR	Pool.	PlugIR
2	+5.6	+8.6	+0.5	+13.9	+13.5	+14.1
4	+6.7	+9.6	+0.8	+13.1	+11.7	+13.5
7	+11.6	+11.9	+0.7	+13.5	+10.5	+12.0
10	+14.0	+13.7	+0.7	+13.4	+10.7	+11.1

Table 11: **Representation and evidence headroom evolve differently across dialogue states.** Paired current H@10 gains over the same dialogue state under CLIP on the 3,000-query development cohort with held-out interface selection. The dialogue-only rewrite has no target-image access, whereas the human sketch is a target-grounded observation.

the same states and query constructions. This pattern is consistent with increasing representation headroom in later unreformulated dialogue states, while the residual sketch benefit persists after the dialogue has already been reformulated.

D.3 Untouched-Cohort and Repeated-Rewrite Controls

	Development (3,000)			Untouched (1,000)		
Added to the state	ChatIR	Pool.	PlugIR	ChatIR	Pool.	PlugIR
Dialogue state alone	28.8	31.3	42.9	27.9	31.9	44.3
Dialogue-only rewrite	42.7	45.0	43.6	42.0	44.8	43.9
Second sampled rewrite	42.6	44.7	43.6	42.4	45.1	44.7
Human sketch	42.2	41.9	54.0	40.6	41.1	54.7
Rewrite, then sketch	52.4	52.9	54.2	52.0	53.2	55.0
Rewrite, then second rewrite	43.7	45.2	43.6	45.9	45.9	44.9

Table 12: **The matched pattern survives an untouched cohort and a second rewrite.** The untouched cohort consists of 1,000 training targets on which no component of this protocol was fit or selected and which no earlier evaluation in this paper used. They belong to the 7k training set of the evaluation-selected maps of Table 3, which are not used here. A second target-blind rewrite adds 0.0–0.9 points after the first on the development cohort and 1.0–3.9 on the untouched cohort, whereas the sketch added to the same rewritten state adds 7.9–11.1.

On the untouched cohort, rewriting adds 14.1 points from concatenation but -0.4 after PlugIR reformulation, whereas the sketch adds 12.7 and 10.4 points, respectively. Repeating the target-blind intervention qualifies rather than removes this pattern. A second independently sampled rewrite reproduces the first within 0.3 points on the development cohort and within 0.8 on the untouched

cohort. Applied sequentially after the first rewrite, it adds 0.9, 0.2, and 0.0 points from the ChatIR, pooled, and reformulated states on the development cohort but 3.9, 1.1, and 1.0 on the untouched cohort, so a second rewrite is not always negligible from a concatenated state. The sketch added to the same rewritten state adds 9.7, 7.9, and 10.7 points on the development cohort and 10.0, 8.4, and 11.1 on the untouched cohort. The direct paired contrast between the two sequential routes, sketch minus second rewrite after the same first rewrite, is $+8.7$ $[+7.2, +10.2]$, $+7.7$ $[+6.0, +9.2]$, and $+10.6$ $[+9.2, +12.1]$ on the development cohort and $+6.1$ $[+3.3, +8.8]$, $+7.3$ $[+4.5, +10.2]$, and $+10.1$ $[+7.5, +12.6]$ on the untouched cohort.

D.4 Matched Evidence Routes on Raw Human VisDial Dialogue

We repeat the matched text comparison on the original human VisDial dialogue for the same 2,064 episodes and fixed 50k gallery. ChatIR-FT uses the released fine-tuned BLIP checkpoint, while the second setting uses zero-shot BLIP-ITC. The dialogue-only rewrite sees only the observed caption and human QA rounds. Spatial and appearance descriptions are privileged probes with direct access to the target image. Added sources use the fixed fusion weight $\alpha = .5$.

	ChatIR-FT BLIP				Zero-shot BLIP-ITC			
Added to the state	Concat.	Pool.	PlugIR	Alone	Concat.	Pool.	PlugIR	Alone
Dialogue state alone	75.8	64.6	73.4	–	44.2	53.2	72.6	–
Dialogue-only rewrite	77.9	76.5	75.0	74.1	70.8	75.8	75.8	75.0
Second sampled rewrite	78.3	76.1	75.2	74.9	70.9	76.2	76.2	76.1
Privileged spatial text	90.3	87.5	89.2	79.0	86.0	88.4	92.1	84.6
Privileged appearance text	95.1	93.8	94.3	90.3	92.4	95.1	96.1	93.1
Rewrite, then appearance	94.8	94.9	94.1	–	96.0	96.5	96.4	–

Table 13: **Matched representation and evidence probes on raw human VisDial dialogue.** Current H@10 at state ten on the same 2,064 episodes and fixed 50k gallery, with $\alpha = .5$ for added-source fusion. Under the fine-tuned encoder, concatenation is already the strongest of the three base dialogue states. Under zero-shot BLIP-ITC, reformulation absorbs much of the gain available from another target-blind rewrite, while privileged target-grounded text retains substantial residual gains.

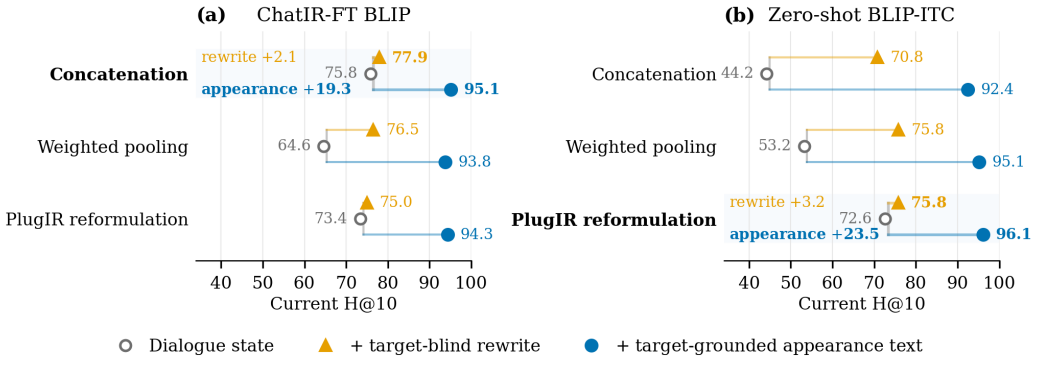

Figure 4: **Target-grounded evidence remains valuable after target-blind refinement on raw human VisDial.** Current H@10 on the same 2,064 episodes and fixed 50k gallery. Gray circles denote the dialogue state, orange triangles add a target-blind rewrite, and blue circles add privileged appearance text. From the three strongest dialogue states (concatenation and reformulation under ChatIR-FT, reformulation under zero-shot BLIP-ITC), rewriting adds only 1.6–3.2 points while appearance text adds 19.3–23.5.

Under zero-shot BLIP-ITC, the target-blind rewrite adds 26.6 points to concatenation but only 3.2 points to the PlugIR-reformulated state. Privileged appearance text adds 48.2 and 23.5 points from

those same states, respectively. Under ChatIR-FT, raw concatenation is already strong: another rewrite adds 2.1 points, while appearance text adds 19.3. After PlugIR reformulation, the corresponding gains are 1.6 and 20.9 points.

The second sampled rewrite closely reproduces the first on the three dialogue states used in Figure 4, differing by at most 0.4 H@10 points within each encoder. The residual appearance gain is therefore not explained by one particular sampled rewrite. These privileged descriptions nevertheless change both target access and wording, so the comparison diagnoses recoverability under expanded access rather than identifying which individual facts were absent from the human dialogue.

D.5 Target Specificity of Added Evidence

Added source	ChatIR	Weighted pool.	PlugIR reform.
Dialogue state alone	28.8	31.3	42.9
Human sketch	42.2	41.9	54.0
category-matched wrong target	18.8	13.4	30.2
randomly permuted target	12.5	7.4	23.6
Dialogue-only rewrite	42.7	45.0	43.6
category-matched wrong target	15.1	11.6	24.3
randomly permuted target	8.2	4.1	15.6
Privileged appearance text	62.7	67.7	66.7
category-matched wrong target	16.9	12.4	25.5
randomly permuted target	9.0	4.5	17.1

Table 14: **The added source must correspond to the target, regardless of modality.** Current H@10 under the clean CLIP protocol, development cohort, $\alpha = .5$. Each added source is replaced by the same source type from a different target. Category-matched targets share nearly the same object categories. Randomly permuted rows average five seeds.

The sign of the intervention reverses when the added source belongs to another target, for both visual and textual sources. Category matching recovers only part of the loss. These controls rule out an explanation in which the gain arises merely from introducing another channel or another query, while still not identifying which semantic details of the correct source are novel relative to the dialogue.

D.6 Fusion Weight and Harm Rate

For a fixed dialogue state, define the *harm rate* as the fraction of targets retrieved in the base top ten that leave the top ten after fusion. Table 15 contrasts the fixed main weight $\alpha = .5$ with a high-weight condition.

Base	Added source	$\alpha = .5$		$\alpha = .9$	
		H@10	Harm (%)	H@10	Harm (%)
ChatIR (base 28.8)	Rewrite	42.7	3	44.0	12
	Sketch	42.2	11	26.7	56
	Appearance	62.7	2	66.6	11
PlugIR reform. (base 42.9)	Rewrite	43.6	2	43.2	5
	Sketch	54.0	9	28.9	59
	Appearance	66.7	3	67.0	13

Table 15: **The fusion weight can determine whether added evidence helps or hurts.** Current H@10 and harm rate on the clean development cohort. Harm is the share of the base state's top-ten successes that are destroyed after adding the source. The main matched comparison fixes $\alpha = .5$, and $\alpha = .9$ is reported as a boundary-condition control.

The sketch remains beneficial at the fixed main weight but becomes harmful when heavily weighted: at $\alpha = .9$ it falls below both dialogue bases and destroys more than half of their existing top-ten successes. The textual rewrite and privileged appearance description are substantially less destructive at the same weight. Rather than identifying an optimal fusion rule, this control shows that access to an informative source is distinct from using that source safely in the retrieval state.

D.7 Per-Query Turnover Behind Aggregate Changes

Aggregate H@10 changes can conceal substantial changes in which targets are retrieved. For two matched states, let an *entry* denote a target that is outside the top ten before the intervention and inside it afterward, and let an *exit* denote the reverse. Their difference determines the aggregate change in H@10, but the two counts reveal how much the retrieved set itself changes.

Change	Before	After	Enter top 10	Leave top 10	Δ H@10
Six more dialogue turns	30.9	29.2	47	98	−1.7
Re-encode same dialogue	29.2	31.8	281	204	+2.6
Dialogue-only rewrite	28.8	42.7	449	30	+14.0
Human sketch, concat. state	28.8	42.2	498	95	+13.4
Human sketch, reform. state	42.9	54.0	445	111	+11.1
Permuted-target sketch	28.8	12.5	45	532	−16.2

Table 16: **Aggregate retrieval changes conceal target turnover.** Counts compare top-ten success before and after each matched intervention on the same 3,000 FS-COCO-Dialog targets. The net H@10 change equals entries minus exits, divided by the cohort size.

For example, the nearly flat dialogue continuation from state four to state ten changes current H@10 by only -1.7 points, yet 47 targets outside the state-four top ten enter it while 98 leave it. Likewise, the sketch added to a reformulated state recovers 445 targets while losing 111. Thus, a small aggregate change need not correspond to a stable retrieved set, while a large positive change can still contain nontrivial losses.

D.8 Joint Ranking Beyond Source Selection

Let A, B, and J denote top-K success for dialogue, sketch, and joint ranking, and let $U = A \cup B$. Then

$$\Pr(J) - \Pr(U) = \Pr(J \setminus U) - \Pr(U \setminus J).$$

No independence assumption is required. Joint ranking can therefore outperform a source-selection oracle when it rescues enough queries that neither source retrieves individually. For ChatIR with CLIP, the recorded counts are 235 rescues and 172 losses, giving a net gain of 63 of 3,000 queries. The rescue rate among queries missed by both individual sources is $235/1764 = 13.3\%$. The source-selection oracle is strong for choosing between the two sources, but it is not an upper bound on all score-level combinations.

A simple margin view explains how this is possible. For distractor g, define

$$m_D(g) = s_D(g^*) - s_D(g), \qquad m_S(g) = s_S(g^*) - s_S(g).$$

Under fixed linear score fusion,

$$m_J(g) = (1 - \alpha)m_D(g) + \alpha m_S(g).$$

The two sources can make errors against different distractors, so a positive margin from one can compensate for a negative margin from the other. This is a mechanistic interpretation, not a measured decomposition of every fusion gain.

Backbone	Query	$\lvert A\rvert$	$\lvert B\rvert$	$\lvert A \cup B\rvert$	$\lvert J\rvert$	Rescue	Loss	Net pp [95% CI]
CLIP	ChatIR	876	669	1236	1299	235	172	$+2.1$ $[+0.8, +3.4]$
	Weighted pool	953	669	1294	1253	222	263	-1.4 $[-2.8, +0.1]$
	PlugIR reform.	1312	669	1589	1643	237	183	$+1.8$ $[+0.4, +3.2]$
	CIReVL-style	1334	669	1573	1599	219	193	$+0.9$ $[-0.5, +2.2]$
BLIP-ITC	ChatIR	829	654	1248	1442	379	185	$+6.5$ $[+5.0, +7.9]$
	Weighted pool	1142	654	1470	1564	309	215	$+3.1$ $[+1.7, +4.7]$
	PlugIR reform.	1726	654	1933	2086	288	135	$+5.1$ $[+3.8, +6.4]$
	CIReVL-style	1765	654	1948	2031	227	144	$+2.8$ $[+1.5, +4.0]$
SigLIP-2	ChatIR	1232	1342	1903	1958	230	175	$+1.8$ $[+0.6, +3.1]$
	Weighted pool	1305	1342	1954	1923	206	237	-1.0 $[-2.4, +0.3]$
	PlugIR reform.	2019	1342	2342	2422	205	125	$+2.7$ $[+1.5, +3.9]$
	CIReVL-style	2000	1342	2306	2350	179	135	$+1.5$ $[+0.3, +2.6]$
ViT-H/14	ChatIR	1176	926	1642	1731	271	182	$+3.0$ $[+1.6, +4.4]$
	Weighted pool	1071	926	1568	1634	281	215	$+2.2$ $[+0.7, +3.7]$
	PlugIR reform.	1667	926	1981	2037	234	178	$+1.9$ $[+0.6, +3.2]$
	CIReVL-style	1701	926	1978	2012	225	191	$+1.1$ $[-0.1, +2.5]$

Table 17: **Joint ranking exceeds source selection in 14 of 16 settings.** All counts use the same 3,000 queries, state ten, and $\alpha = .5$. A, B, and J denote dialogue, sketch, and joint success. Rescue counts $J \setminus (A \cup B)$ and loss counts $(A \cup B) \setminus J$. Net gains are positive in fourteen settings and their paired intervals exclude zero in twelve. Negative margins are retained.

E Extended Dialogue Across Retrieval Spaces

We extend the same 3,000 FS-COCO-Dialog evaluation dialogues from ten to twenty turns, keeping turns 1–10 unchanged. Table 18 repeats the comparison under CLIP, BLIP-ITC, and SigLIP-2. The best-unit oracle uses target ranks and is diagnostic rather than deployable.

Retrieval space	State	$t = 10$	$t = 20$	$\Delta_{10\to20}$	+ sketch at $t = 20$
CLIP	ChatIR concatenation	29.2	29.2	$+0.0$	$+14.1$
	Weighted pooling	31.8	33.4	$+1.7$	$+8.9$
	PlugIR reformulation	43.7	45.9	$+2.2$	$+10.0$
	Best-unit oracle†	48.3	54.2	$+5.9$	–
BLIP-ITC	ChatIR concatenation	27.6	20.9	-6.7	$+22.9$
	Weighted pooling	38.1	38.4	$+0.4$	$+14.4$
	PlugIR reformulation	57.5	54.5	-3.1	$+11.6$
	Best-unit oracle†	58.1	64.3	$+6.2$	–
SigLIP-2	ChatIR concatenation	41.1	41.1	$+0.0$	$+24.2$
	Weighted pooling	43.5	44.9	$+1.4$	$+19.5$
	PlugIR reformulation	67.3	66.0	-1.3	$+14.6$
	Best-unit oracle†	60.4	67.1	$+6.7$	–

Table 18: **The single-unit oracle gains more from ten additional turns than the deployed states.** Current H@10 on the same 3,000 extended FS-COCO-Dialog dialogues. The best-unit oracle improves in all three retrieval spaces, while deployed states improve much less or deteriorate. Sketch gains remain substantial at state twenty. † uses target ranks.

The best-unit oracle gains 5.9, 6.2, and 6.7 H@10 points under CLIP, BLIP-ITC, and SigLIP-2, respectively, showing that later turns contain retrieval-relevant observations. The deployed states capture substantially less of this opportunity: weighted pooling gains at most 1.7 points, while PlugIR reformulation decreases by 3.1 points under BLIP-ITC and 1.3 points under SigLIP-2. BLIP-ITC concatenation, whose encoder can read the longer prefix, also decreases by 6.7 points. The remaining sketch gains therefore cannot be attributed solely to stopping the dialogue at ten turns.

The oracle is monotone in the number of units, so part of its gain reflects a larger candidate family rather than new facts. The targets whose first successful unit lies in turns 11–20 number 177, 187, and 200 under CLIP, BLIP-ITC, and SigLIP-2. Of these, the state-twenty reformulated state retrieves 81, 122, and 140, weighted pooling 26, 42, and 51, and concatenation 11, 21, and 33. The sketch fused at state ten retrieves 77, 109, and 135 of the same targets without any further turn.

Fixed-observation reformulation control. To calibrate how much of the oracle gain a larger candidate family alone can produce, we sample eight reformulations of the unchanged ten-turn dialogue from the deployed reformulator (same prompt and model, temperature .7 and 1.0, one request with eight completions per episode) and score them in the retrieval spaces of Table 18. Table 19 reports the greedy reformulation, the range over the eight samples, and the expected best of k samples. The eight samples stay within 0.7 H@10 points of one another at both temperatures, while the best-of-eight oracle adds 3.9, 4.6, and 3.7 points under CLIP, BLIP-ITC, and SigLIP-2 at temperature .7 and 4.8, 5.9, and 5.1 at temperature 1.0. Ten additional units add 5.9, 6.2, and 6.7 points to the single-unit oracle, so most of that gain is matched by paraphrase choice over fixed observations. The deployed reformulated state gains at most 2.2 points from those units, and the sketch fused at state ten adds 11.0, 12.0, and 13.4 points, more than any paraphrase choice. Of the late-only targets, the best of eight fixed-observation reformulations retrieves 60, 99, and 111 at temperature .7 (61, 104, and 116 at 1.0), the state-twenty reformulation 81, 122, and 140, and the sketch fused at state ten 77, 109, and 135.

				Best of k samples†			+ sketch at $t = 10$	
Retrieval space	Temp.	Greedy	Eight samples	$k = 2$	$k = 4$	$k = 8$	Best of 8†	Greedy
CLIP	.7	43.7	43.4–43.9	45.4	46.8	47.6 (+3.9)	58.0	54.8
	1.0	43.7	43.5–44.2	45.8	47.5	48.5 (+4.8)	58.9	54.8
BLIP-ITC	.7	57.5	57.2–57.9	59.5	61.0	62.2 (+4.6)	72.5	69.5
	1.0	57.5	57.3–58.0	59.9	61.9	63.5 (+5.9)	73.5	69.5
SigLIP-2	.7	67.3	66.9–67.5	68.9	70.1	71.0 (+3.7)	82.5	80.7
	1.0	67.3	67.0–67.7	69.4	71.0	72.4 (+5.1)	83.2	80.7

Table 19: **Eight reformulations of the unchanged ten-turn dialogue bound the candidate effect of the oracle.** Current H@10 on the 3,000 extended FS-COCO-Dialog episodes in the retrieval spaces of Table 18, for the greedy PlugIR reformulation of the first ten turns, the range over eight reformulations sampled from the same reformulator at the given temperature, and the expected best of k of them (per query, from target ranks, with the gain over the greedy reformulation in parentheses). The last two columns fuse the sketch at state ten ($\alpha = .5$). † uses target ranks.

The three recorded rewrites of the held-out protocol (one greedy and two sampled at temperature .7 from the dialogue-only rewrite prompt, Appendix D.3) give the same picture in four spaces and on the untouched cohort (Table 20): individual rewrites differ by at most 0.7 H@10 points alone on the development cohort and 2.3 on the untouched cohort, while the oracle over three of them adds 1.8–2.6 points alone and 0.8–1.3 points fused into the reformulated state on the development cohort, and 2.2–2.7 and 1.3–1.9 points on the untouched cohort.

		Rewrite alone			PlugIR reformulation + rewrite		
Retrieval space	Cohort	Greedy	Range	Best of 3	Greedy	Range	Best of 3
CLIP	development	42.8	42.7–42.8	45.1 (+2.3)	43.6	43.3–43.6	44.8 (+1.2)
	untouched	44.0	43.5–44.6	46.5 (+2.5)	43.9	43.8–44.7	45.4 (+1.5)
BLIP-ITC	development	61.2	60.8–61.2	63.8 (+2.6)	61.4	61.0–61.4	62.5 (+1.2)
	untouched	63.3	62.3–64.6	66.0 (+2.7)	63.0	63.0–64.1	64.9 (+1.9)
SigLIP-2	development	67.3	67.3–68.0	69.9 (+2.6)	68.1	68.1–68.4	69.4 (+1.3)
	untouched	68.4	68.1–68.4	70.6 (+2.2)	68.6	68.6–68.7	69.9 (+1.3)
OpenCLIP ViT-H/14	development	55.5	55.4–55.5	57.3 (+1.8)	55.9	55.7–55.9	56.8 (+0.8)
	untouched	57.5	57.5–57.7	60.1 (+2.6)	57.5	57.4–57.8	58.8 (+1.3)

Table 20: **Choice among paraphrases of fixed observations bounds the candidate effect.** Current H@10 at state ten under the held-out protocol ($\alpha = .5$) for the greedy dialogue-only rewrite, the range over the greedy rewrite and two rewrites sampled at temperature .7 from the same prompt, and the best-of-three oracle over them (gain over the greedy rewrite in parentheses, from unrounded per-query results). The oracle uses target ranks.

F Alternative Next Questions

At states four and ten of every FS-COCO-Dialog evaluation dialogue, the deployed image-blind Qwen3-VL-32B questioner proposes eight candidate questions about different visual details in one response. Separately, we draw eight independent samples from the deployed questioner at temperature .7. MiniCPM-V-4.5 answers every distinct candidate with access to the target image. Each question–answer pair defines an alternative next state under concatenation and weighted pooling.

For each candidate pool, *Random* averages performance over the eight candidate questions. *Best of k* is the expected best result among k candidates drawn uniformly from the pool, computed per query using target ranks. *Novelty* is target-blind and selects the candidate whose encoded unit has the lowest cosine similarity to the observed prefix. Because it scores the encoded question–answer unit, it is a post-answer selection control that requires all eight candidates to be answered rather than a deployable pre-answer question policy. *Worst* uses target ranks and is included to expose the range induced by question choice.

State	Query	Committed	Random	Novelty	Best of 2†	Best of 4†	Best of 8†	Worst†
Alternatives about different visual details								
4	Concatenation	29.9	29.3	29.7	31.0	32.6	34.2	24.9
4	Weighted pooling	29.7	28.7	29.5	31.0	33.2	35.4	23.4
10	Concatenation	29.2	29.2	29.2	29.2	29.2	29.2	29.2
10	Weighted pooling	32.5	32.0	32.6	33.5	34.9	36.4	28.6
Independent resamples of the deployed policy ($T = .7$)								
4	Concatenation	29.9	30.0	30.7	30.8	31.3	31.7	28.3
4	Weighted pooling	29.7	29.6	30.1	30.4	31.0	31.4	27.8
10	Concatenation	29.2	29.2	29.2	29.2	29.2	29.2	29.2
10	Weighted pooling	32.5	32.3	32.2	32.9	33.4	33.7	30.8

Table 21: **Question-selection headroom under two candidate pools.** Current H@10 after one additional turn on the same 3,000 FS-COCO-Dialog episodes with CLIP. Committed is the benchmark question. The first candidate pool explicitly requests questions about different visual details, whereas the second independently resamples the deployed questioner. Novelty is a target-blind selector based on dissimilarity to the observed prefix. † uses target ranks and is therefore privileged.

The explicitly diversified candidate pool exposes substantially more question-selection headroom than independent resampling. At state four, best-of-eight selection raises concatenation from 29.9 to 34.2 H@10 and weighted pooling from 29.7 to 35.4. At state ten, weighted pooling rises from 32.5

to 36.4. The worst candidate instead lowers retrieval to 24.9, 23.4, and 28.6 in these three settings, showing that the choice of question can materially change the next retrieval state.

This headroom is not recovered by the tested target-blind selector. Novelty selection stays within 0.3 H@10 points of the committed question in the diversified pool. Best-of-k therefore measures a privileged finite-candidate opportunity rather than the performance of a deployable question-selection rule.

Independent resampling also produces much less diverse candidate sets. The eight resampled candidates contain on average 3.2 and 3.6 distinct questions at states four and ten, respectively, with 2.5 and 3.0 distinct answers. In contrast, explicitly requesting different visual details produces 8.0 distinct questions and approximately 7.8 distinct answers. Mean pairwise question cosine similarity falls from .91 and .88 under resampling to .70 and .68 under explicit diversification. Correspondingly, the best-of-eight gain under resampling is at most 1.8 H@10 points.

The state-ten concatenation result provides a separate representation control. Every alternative question yields the same 29.2 H@10 because the additional question–answer turn falls outside CLIP's 77-token context window. Thus, even when alternative observations are available, their value can be bounded by the query interface used to construct the retrieval state.

These experiments diagnose question-selection headroom under the tested finite candidate pools. They do not provide an upper bound on possible question policies, and they do not equate the interaction cost or information budget of one additional question with that of another target-grounded source.

G Representation Controls for Visual Evidence

G.1 Representation Routes and Candidate Access

The following comparisons ask how available sketch evidence can enter retrieval. They do not form a fully matched architecture benchmark. Pic2Word, SEARLE, and LinCIR use released mappings that were not trained with the same sketch supervision as the direct alignment used in our evidence probe, while reranking methods operate on a first-stage shortlist.

Table 22 compares the published representation routes under different sketch features. Table 23 further tests candidate access.

Sketch representation	R@1	R@5	H@10	medR	ΔH@10
Released mappings with their original visual features					
Pic2Word (Saito et al., 2023)	7.2	17.4	23.4	82	–
SEARLE (Baldrati et al., 2023)	8.3	19.1	25.5	70	–
LinCIR (Gu et al., 2024)	10.0	21.8	28.7	57	–
Same mapping, our adapter's sketch features					
Pic2Word	7.3	17.6	23.5	82	+0.1
SEARLE	8.1	19.2	25.9	67	+0.4
LinCIR	6.2	15.0	21.5	104	−7.2
Direct sketch	**17.8**	**35.1**	**43.3**	**16**	–

Table 22: **Released pseudo-token mappings do not match direct sketch scoring in this adaptation.** CLIP with ChatIR at turn 10. The same human sketch is used throughout, but the released mappings were not trained with the same sketch supervision as the direct alignment used in our probe.

For any reranker restricted to a first-stage candidate set C_M and any $K \leq M$,

$$\{r_{\text{rerank}} \leq K\} \subseteq \{g^* \in C_M\}, \qquad \text{H@}K_{\text{rerank}} \leq \text{H@}M_{\text{first}}.$$

The candidate-access experiment raises first-stage R@20 from 40.0 to 54.0 when the first stage receives the sketch, and reranked H@10 rises from 35.5 to 47.0. Because the candidate distribution also changes, this experiment does not assign a unique fraction of the gain to candidate access.

Candidate source	First stage H@10	First stage R@20	After reranking R@1	After reranking R@5	After reranking H@10
Weighted pooling, text only	31.8	40.0	23.8	32.3	35.5
+ drawn sketch (α=.3)	44.2	**54.0**	**28.6**	**41.5**	**47.0**

Table 23: **Candidate access constrains reranking.** Qwen3-VL-32B and the top-20 reranking budget are fixed. This diagnostic uses sketch weight $\alpha = .3$, separate from the main $\alpha = .5$ comparison.

G.2 Fixed Fusion and Exploratory Alternatives

The main paired comparison fixes $\alpha = .5$ for every pipeline and backbone. We do not select the main operating point from the evaluation-target sweep below.

α	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
Evaluation, Weighted pooling	31.8	37.0	42.3	44.2	**44.5**	41.8	38.3	33.8	30.1	25.7	22.3
Train, Weighted pooling	36.4	48.1	58.7	66.6	71.1	**72.8**	71.7	69.0	63.9	57.1	48.4
Train, ChatIR	32.3	39.2	47.8	56.1	63.0	68.0	**70.3**	69.6	66.5	58.9	48.4

Table 24: **Sensitivity to the sketch weight.** Current H@10. The evaluation-target sweep is exploratory and is not used to retune the fixed $\alpha = .5$ main comparison.

Combination rule	R@1	R@5	H@10	medR
Individual channels				
Text only	11.0	24.7	31.8	40
Sketch only	7.2	16.4	22.3	79
Fixed combination				
Embedding sum, text-emphasized	18.1	35.5	44.2	16
Score sum, text-emphasized	18.1	35.5	44.2	16
Embedding sum, intermediate	17.8	35.1	**44.5**	**15**
Z-scored sum	18.0	34.7	44.3	16
Min–max sum	17.8	34.7	**44.5**	**15**
Adaptive and rank-based combination				
Text-margin weighting	17.6	33.5	42.7	18
Sketch-margin weighting	16.9	33.1	42.2	17
Joint-margin weighting	17.4	33.5	43.0	17
Max z-score	12.5	26.5	34.2	30
Reciprocal rank fusion	15.6	30.9	40.7	20
Borda	14.5	27.8	36.9	26
CombMNZ	17.6	33.8	43.2	17
Orthogonal residual	16.8	33.9	42.4	18
Learned combiner	14.6	30.2	39.6	21
Privileged oracle controls				
Route per query	16.4	34.2	43.1	16
α *per query*	26.7	45.2	53.4	8

Table 25: **Alternative combination rules for weighted pooling and sketch evidence.** CLIP. The route-per-query and per-query-α rows use target ranks and are privileged controls. The main paper does not substitute the best value from this exploratory table for the fixed $\alpha = .5$ comparison.

The exploratory weighted-pooling sweep reaches 44.5 H@10 at $\alpha = .4$, while the fixed main setting at $\alpha = .5$ reaches 41.8. The source-selection oracle for this query family reaches 43.1. We therefore do not use the tuned $\alpha = .4$ row to claim that every fixed fusion exceeds source selection. The main complementarity claim is based on the separately recorded fixed-weight comparisons and reports both positive and negative cases.

G.3 Additional Backbones and Task-Specific Training

Retrieval space	Query construction	Base H@10	+ sketch	Gain
BLIP, task-trained	Local ChatIR-FT	58.8	67.2	+8.4
BLIP, task-trained	CAST adaptation	52.5	64.2	+11.7
OpenCLIP ViT-H/14	ChatIR	39.2	57.7	+18.5
OpenCLIP ViT-H/14	Weighted pooling	35.7	54.5	+18.8
OpenCLIP ViT-H/14	PlugIR reform.	55.6	67.9	+12.3
OpenCLIP ViT-H/14	CIReVL-style	56.7	67.1	+10.4

Table 26: **The sketch gain persists with task-specific text training and a fourth retrieval backbone.** Values are current H@10. CIReVL-style already uses a caption of the sketch, so its gain is additional representation rather than a new observation.

Across three recorded alignment seeds for CLIP, the gain of any tested sketch pair changes by at most 0.4 H@10 points. We report this seed sensitivity separately from paired query-bootstrap intervals because they represent different sources of variation.

G.4 Alignment Transfer Beyond Shared Artists

The standard FS-COCO split separates target photographs but shares artists: each of 100 artists contributes 70 training and 30 evaluation sketches. A five-fold control trains on 80 artists and evaluates the held-out 20, pooling predictions over the same 3,000 evaluation targets. It reaches 21.7 sketch-only H@10 and 42.5 with ChatIR, compared with 22.3 and 43.3 for the standard map. Thus the gain persists when the alignment does not observe the evaluation artist's training drawings.

A separate map trained on Sketchy (Sangkloy et al., 2016), without FS-COCO training data, reaches 13.5 ± 0.2 sketch-only H@10 and 37.6 ± 0.5 when combined with ChatIR, across three seeds. The transferred map retains an 8.4-point gain over the same 29.2 dialogue baseline. Unaligned CLIP sketch features instead yield 18.1 under the same fusion weight. These results support the role of learned alignment and test training-corpus transfer. They do not introduce a new evaluation-image distribution or compare drawing effort with dialogue cost.

G.5 Gallery Scaling and Sketch Degradation

Method	10k	25k	50k	133k
Caption	31.3	24.6	20.1	13.6
+ sketch	51.0	43.0	36.7	27.6
ChatIR, turn 10	40.5	34.7	29.2	22.0
+ sketch	56.4	49.2	43.3	33.9
Weighted pooling, turn 10	44.2	37.2	31.8	24.4
+ sketch	**57.9**	**50.0**	**44.2**	**35.6**
Sketch only	34.7	27.2	22.3	16.1

Table 27: **Recorded sketch gains as gallery size increases.** Entries are current H@10. ChatIR uses sketch weight .5 and weighted pooling uses .3, so this table is a scaling diagnostic rather than a matched weight comparison.

The scaling table uses different sketch weights for ChatIR and weighted pooling, so it is supporting evidence rather than a matched comparison. The degradation table provides the more direct boundary condition used in the main paper. Once the sketch loses enough visual structure, adding it no longer helps and can hurt. We therefore do not interpret a positive sketch gain as a modality effect that should hold for arbitrary drawings.

Sketch condition	Ink	R@1	H@10	medR	Joint H@10
Original	100%	7.2	22.3	79	44.2
Components, 90%	96%	6.4	20.1	88	43.7
Components, 75%	89%	5.6	18.2	103	42.7
Components, 50%	81%	4.6	15.6	163	41.5
Components, 25%	65%	3.4	11.1	387	37.7
Components, 10%	47%	2.1	7.3	1928	33.8
Rendered at 112 px	101%	3.7	13.9	188	39.7
Rendered at 56 px	97%	0.5	2.3	2862	31.1
Dilated 5 px	210%	1.0	6.5	702	36.7
Eroded 5 px	5%	0.0	0.1	23703	19.5

Table 28: **Sketch degradation eventually removes the gain.** Joint H@10 uses weighted pooling with sketch weight .3. The table is used to locate a failure boundary, not to compare against the fixed $\alpha = .5$ main setting.

H Data and Experimental Details

H.1 Benchmark Construction

FS-COCO-Dialog extends FS-COCO with generated dialogue while preserving disjoint training and evaluation target images. Qwen3-VL-32B (Bai et al., 2025a) acts as an image-blind questioner, receiving the initial caption and dialogue history. MiniCPM-V-4.5 (Yu et al., 2026) answers from the target photograph and the current question. The open protocol forbids yes/no questions, while a second evaluation protocol permits them on the same target images. The two protocols are therefore paired interaction variants, not independent image test sets.

	Eval: yes/no	**Eval: open**	**Train: open**
Dialogues	3,000	3,000	7,000
Caption words, mean	6.6	6.6	6.5
Question words, mean	9.6	11.4	11.4
Answer words, mean / median	1.8 / 1	1.8 / 1	1.8 / 1
Yes/no-form questions	95.5%	0.0%	0.0%
Bare yes/no answers	30.2%	0.1%	0.0%
“Not shown” answers	0.3%	0.3%	0.3%
Repeated-question dialogues	2.4%	1.8%	2.0%

Table 29: **The open protocol removes binary questioning without lengthening answers.**

The questioner prompt used for the open protocol is:

> You are looking for one specific image in a database of millions. You cannot see it; you only know what the dialogue below tells you. Ask ONE short question whose answer would best narrow down which image it is. Ask about a single concrete visual detail that is still unknown. Never repeat a question or ask about something already stated.

The open-protocol suffix additionally forbids questions answerable by yes or no. The reference to “millions” is part of the generation prompt and does not describe the evaluation gallery size.

H.2 Dialogue Fidelity

Two independent vision-language models were used as automated fidelity judges on sampled QA pairs. Their disagreement is substantial (Table 30).

Because the judges disagree, we do not treat either model’s labels as ground truth. In particular, weak association between a model-judged label and retrieval rank would not show that incorrect answers are harmless. The reweighting experiment below tests sensitivity to those labels only.

Split	Judge	correct	partial	wrong	agreement	κ
Yes/no	InternVL3.5-8B	39.8	0.1	59.0	51.2%	0.24
Yes/no	GLM-4.6V	77.1	8.0	14.8		
Open	InternVL3.5-8B	52.8	0.4	46.3	57.8%	0.36
Open	GLM-4.6V	74.0	14.6	11.5		

Table 30: **Agreement between independent fidelity judges.** Correct, partial, and wrong entries are percentages. A further 1.1% (yes/no) and 0.5% (open) of InternVL3.5 replies could not be parsed into a grade and are excluded from the three categories. The GLM-4.6V rows differ from 100% by rounding only.

The claims of the main paper are stated for the recorded trajectories and do not depend on either judge's labels. The same evidence-route protocol is also replicated on raw human-written VisDial dialogue in Appendix D.4.

The answerer uses temperature zero and the instruction: "Answer the question about this image truthfully in at most 12 words. If the image does not show it, say so plainly. No preamble." Both fidelity judges receive the image and QA pair with the instruction: "You grade question-answer pairs about an image. Look at the image. Reply with exactly one word: correct, partial, or wrong."

Turn weighting	Text H@10	Joint H@10
Original weighting	31.8	44.2
Soft fidelity weighting	32.0	43.8
Hard fidelity weighting	30.2	42.5
Drop wrong turns	30.2	42.4

Table 31: **Retrieval under model-judged turn weighting.** Weighted pooling with sketch weight .3. These rows reweight the original generated answers using model labels.

H.3 Model and Optimization Details

The split file (5,000 fitting, 1,000 selection, and 1,000 held-out training targets), the 50k gallery manifest, the target IDs of every cohort, all prompts, and the configuration of every reported comparison are recorded with the per-query rank files. The enlarged galleries used for the scaling diagnostic are auxiliary, and the main conclusions do not depend on their being perfectly deduplicated.

H.4 Input Provenance and Adapter Selection

In the preprocessed VisDial (Das et al., 2017) setting, the inputs are LLM-preprocessed standalone captions distributed with DAR, while in the raw setting (Appendix D.4) they are the original caption and human question–answer rounds. Concatenation joins captions $0, \ldots, t$ with a comma and space, latest-turn retrieval encodes caption t, and pooling averages their embeddings. The released ChatIR-FT BLIP checkpoint is distinct from the local ChatIR-FT adaptation trained on the 7k FS-COCO-Dialog dialogues.

The production sketch comparison includes a caption–photo text map alongside the sketch map, with the same linear-map recipe. In that broader pipeline–backbone matrix, both checkpoints were selected using single-modality H@10 on the evaluation split among six evaluations. Those results are therefore conditional on evaluation-based checkpoint selection and are used as a breadth analysis rather than an untouched-test estimate.

The matched representation-versus-evidence comparison in Appendix D.2, by contrast, uses the held-out-selection protocol described there: maps are fit on 5,000 training targets and selected on a disjoint 1,000-target training split, with a further 1,000-target cohort left untouched by this protocol. The same qualitative interaction is observed under both protocols.

Component	Setting
Retrieval backbones	CLIP ViT-L/14, BLIP-ITC base, SigLIP-2 SO400M/16@384, and OpenCLIP ViT-H/14 (Cherti et al., 2023)
Sketch and text alignment	Bias-free linear map $W \in \mathbb{R}^{d\times d}$ initialized as identity. Symmetric InfoNCE with logit scale 100. AdamW, learning rate 10^{-4}, weight decay 10^{-2}, batch 256, 60 epochs. Dimension $d = 768$ (CLIP), 256 (BLIP-ITC), 1152 (SigLIP-2), 1024 (ViT-H/14). 7k pairs in the breadth matrix, 5k fit / 1k selection in the held-out protocol
Sketch fusion	Fixed score fusion with $\alpha = .5$ and normalized mapped sketch features
ChatIR-FT (local)	BLIP text encoder fine-tuned on the 7k FS-COCO-Dialog dialogues with random prefix lengths. Symmetric InfoNCE, temperature 0.07. AdamW, learning rate 10^{-5}, weight decay 0.05, batch 128, 5 epochs, seed 0
CAST	Guided transformation with frozen retrieval towers. AdamW, learning rate 5×10^{-4}, weight decay 10^{-4}, exponential decay 0.93 per epoch, temperature 0.03, batch 256, 50 epochs
Learned pooling	Attention pooler (tanh scorer with hidden width 128 plus a learned position bias) trained for 600 steps, batch 256, AdamW, learning rate 10^{-3}, weight decay 10^{-2}, logit scale 100
Question generation	Qwen3-VL-32B with temperature .7 and a 40-token generation limit
PlugIR reformulation	Qwen3-VL-32B with deterministic decoding on the shared benchmark dialogue
DAR	SDXL-Turbo (Sauer et al., 2024) at 512^2 with one denoising step and generated-image weight .3
CIReVL-style	MiniCPM-V-4.5 sketch caption followed by Qwen3-VL-32B query composition
Reranking	Qwen3-VL-32B over a fixed top-20 candidate set
Privileged text	Target-grounded Qwen3-VL-32B descriptions with separate spatial and appearance prompts
Statistics	2,000 paired bootstrap resamples over matched episode IDs (seed 0), percentile 95% intervals

Table 32: **Implementation summary.** Settings used for the controlled comparisons.

H.5 Statistical Comparisons

All paired differences in the main paper are computed on shared episode IDs. Paired bootstrap intervals therefore resample the same targets jointly across the compared methods. Random perturbation or alignment seeds are a separate source of variation and are summarized across repeated runs rather than folded into the query bootstrap. We avoid using an across-query interval as evidence about training-seed variability.

The evidence-access experiments also have different inferential scopes. The DAR-missed description cohort is a selected diagnostic subset, whereas the sketch table uses the full 3,000-query FS-COCO-Dialog evaluation cohort. Claims in the main paper preserve that distinction rather than pooling the two analyses into a single estimate.

H.6 Scope of the Evidence-Access Comparisons

The acquisition probes change access to target-grounded observations and, in some cases, how evidence is expressed. A description generated from the target can add facts, rephrase known facts, or correct noisy answers. A human sketch can similarly provide overlapping and complementary content. The matched dialogue-only rewrite in Figure 2 controls target access while providing a compact linguistic representation. It does not match every semantic detail or establish which individual recovered queries require novel facts. Target-mismatched controls provide an additional boundary condition: category-matched and randomly permuted sketches reverse the gain under the same fusion rule, as does permuted target-informed appearance text (Appendix D.5). These controls show that the benefit depends on target-specific evidence rather than merely on introducing another source. They still do not isolate which recovered queries require facts absent from the dialogue, because the correct sketch may provide new content, re-express known content, or correct noisy dialogue evidence.