RACE: LEARNING MULTIMODAL SEARCH THROUGH REFERENTIAL ABSTRACTION AND COMPOSITIONAL EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Training multimodal search agents requires tasks that make visually grounded search necessary for determining the answer. Yet synthetic questions can admit correct answers while bypassing the image or ignoring evidence from intended targets, weakening the resulting supervision. We propose **RACE**, an image-conditioned task synthesis framework that constructs dependencies from images to targets and from target-specific evidence to answers. *Referential Abstraction* (RA) replaces identity-revealing text with specific visual references or ambiguous relational references resolved by searching among depicted candidates. *Compositional Evidence* (CE) links retrieved facts about multiple entities in one image or entities from different images into joint questions whose answers require every designated contribution. Both operations also apply to images found during search, allowing new visual observations to determine subsequent search targets or supply answer evidence. Fine-tuning Qwen3.5-27B on only 12K teacher-generated trajectories from RACE raises average accuracy across ten multimodal benchmarks from 53.5 to 56.0, a relative improvement of approximately 4.7% over the matched base agent. Our code, data, and model will be publicly available.

1 INTRODUCTION

Multimodal deep search requires agents to resolve targets in images, search for information about them, and combine evidence. When search reveals further targets in new images, resolving these references determines what to investigate next. Recent systems support multi-turn exploration, search guided by intermediate images, and concurrent search across entities (Huang et al., 2026b; Zhang et al., 2026b; Li et al., 2026). Synthetic-data pipelines link evidence across search steps or entities to construct challenging questions, then collect successful teacher trajectories for fine-tuning. Answer correctness alone does not establish which visual search skills a teacher trajectory exercises.

Our central insight is that task synthesis should make visually grounded search necessary for determining the answer. This requires a dependency from images to targets and from target-specific evidence to the answer. Textual identity clues let a solver bypass the first dependency by recovering the target without the image (Figure 1(a)). A task intended to train multi-target reasoning can still mention an entity whose evidence is unnecessary for its answer (Figure 1(b)). Both failures allow correct answers without the corresponding search process. In our constructed example, resolving the ambiguous reference assigns the two women's facts to a directed comparison, whose answer requires both contributions (Figure 1(c)).

We introduce RACE (Referential Abstraction and Compositional Evidence), a framework that constructs these dependencies jointly from real images and retrieved evidence. *Referential Abstraction* (RA) ties target resolution to the image through specific visual references or ambiguous relational references verified by search. The latter requires searching every depicted candidate, supporting one match, ruling out alternatives, and rejecting conditions that identify the target without the image.

Compositional Evidence (CE) makes the answer depend on every selected target's evidence. It connects retrieved facts through supported relations to construct a joint question about multiple entities in one image or one entity from each of several images. The facts must jointly determine one answer, while removing any designated contribution leaves it undetermined. RA therefore establishes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

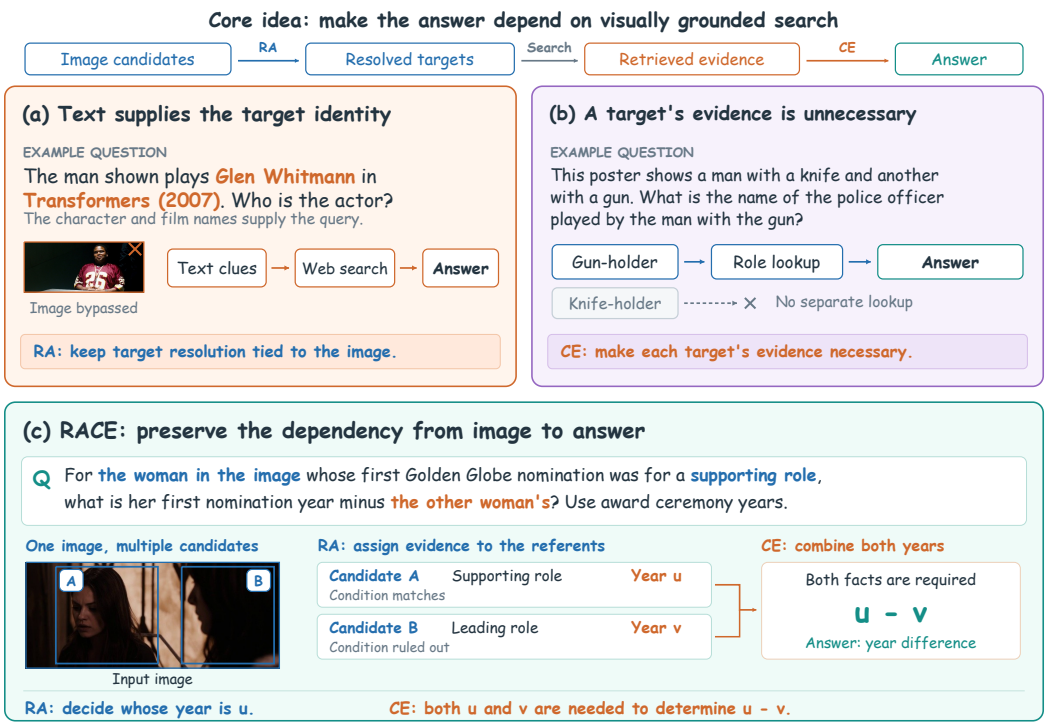


Figure 1: **Making visually grounded search necessary.** (a) Textual identity clues allow the target to be identified without the image. (b) The question uses the gun-holder's evidence while leaving the knife-holder's evidence unused. (c) RA resolves the relational reference and assigns the two candidates' years to u and v; CE combines both years to answer the directed comparison $u - v$. Arrows denote information dependencies.

whose evidence is relevant, and CE specifies how those contributions determine the answer. This also supports multi-image questions in which every image supplies a necessary contribution.

The two operations recur as text and image searches reveal further targets or evidence, linking visual target resolution to subsequent search and the final answer. A teacher solves the resulting tasks using only their images and questions. We retain only trajectories whose final answers are judged correct against the reference answers for supervised fine-tuning (SFT).

We make three contributions:

- We frame task synthesis around making visually grounded search necessary for the answer. This perspective unifies identity leakage and unused entity evidence as breaks in the dependency from images, through targets and their evidence, to answers.

- We introduce RACE, which constructs this dependency through Referential Abstraction and Compositional Evidence: the former ties target resolution to images, while the latter makes every selected target's evidence necessary for joint questions within or across images.

- We show that SFT on only 12K synthesized trajectories raises Qwen3.5-27B's average accuracy from 53.5 to 56.0 across ten benchmarks (Table 1) and increases visual-tool use beyond the first round on all four primary benchmarks.

2 Related Work

2.1 Multimodal Deep Search Agents

Multimodal search agents learn to use image and text tools through supervised training or reinforcement learning (RL) (Geng et al., 2025; Wu et al., 2025; Narayan et al., 2025; Zhang et al., 2026c). OpenSearch-VL develops fine-grained visual tools (Chen et al., 2026), Vision-DeepResearch supports multi-turn, multi-entity, and multi-scale retrieval (Huang et al., 2026b), and HyperEyes optimizes parallel grounding and search (Li et al., 2026). Other methods address visual evidence reuse (Du et al., 2026), context compression by rendering historical text observations as images (Liu et al., 2026), reasoning over retrieved images (Zhang et al., 2026d), and adaptive task generation (Huang et al., 2026a). DeepVoyager-VL combines intermediate visual dependencies with selective acquisition of visual evidence for long-horizon search (Zhang et al., 2026b). RACE focuses on how task structure determines the reasoning taught by successful trajectories. It treats visual target discrimination and joint evidence use as coupled requirements on a solution, specifying the intended supervision before teacher execution.

2.2 Datasets and Benchmarks

Knowledge-based VQA evaluates knowledge beyond image content (Marino et al., 2019; Schwenk et al., 2022) and fine-grained entity information (Chen et al., 2023). MMSearch and Vision-DeepResearch Benchmark evaluate multimodal retrieval and grounding (Jiang et al., 2024; Zeng et al., 2026). We evaluate on MMSearch-Plus, MM-BrowseComp, BrowseComp-V^3, and VisBrowse-Bench, which emphasize provenance, multimodal browsing, and visual evidence verification (Tao et al., 2025; Li et al., 2025; Zhang et al., 2026a;e). To generate training data for these agents, existing synthesis pipelines introduce images into textual tasks or knowledge paths (Geng et al., 2025; Chen et al., 2026), construct questions from complex images (Huang et al., 2026b), and incorporate visual-detail steps or webpage images (Zhang et al., 2026d; Du et al., 2026). Structured approaches use scene graphs (Hudson & Manning, 2019; Yang et al., 2026) and image-rooted evidence hypergraphs (Yao et al., 2026). DeepVoyager-VL composes event-graph programs with necessary intermediate visual evidence and minimal constraint sets for answer determination (Zhang et al., 2026b), while HyperEyes constructs mosaic questions that require information from every selected entity (Li et al., 2026). RACE formulates synthesis over visually grounded targets, with separate conditions for unique reference resolution and each target's necessary contribution to a joint answer. Coupling these conditions makes the reasoning requirements explicit at the target level, within scenes, across images, and throughout search.

3 Method

RACE synthesizes tasks around required dependencies from images to targets and from target-specific evidence to answers (Figure 2). Referential Abstraction ties target resolution to image-grounded references; Compositional Evidence makes each selected target's evidence necessary for the joint answer. These are information dependencies: ambiguous references may require retrieving candidate evidence before the target is resolved. Grounding and dependency feedback guide revisions within these operations, including when search introduces new images. Teacher trajectories with correct final answers provide supervision for SFT.

3.1 Image-Conditioned Task Synthesis

We start from real images to link target identification in natural scenes to verifiable external evidence. Sources include social-media photographs, Bilibili video frames, and MovieNet film imagery (Huang et al., 2020). Existing search questions and solution trajectories are not required. For unannotated images, lightweight filtering, visual analysis, and retrieval establish target identities. MovieNet annotations provide identity-to-image correspondences. Both routes supply grounded candidates and identity provenance to the generator.

For input images $\mathcal{I} = \{I_1, \ldots, I_m\}$, let E_i denote the candidates in image I_i. The generator selects source targets $S \subseteq \bigcup_i E_i$, constructs references $R = \{\rho_e : e \in S\}$, explores textual and visual

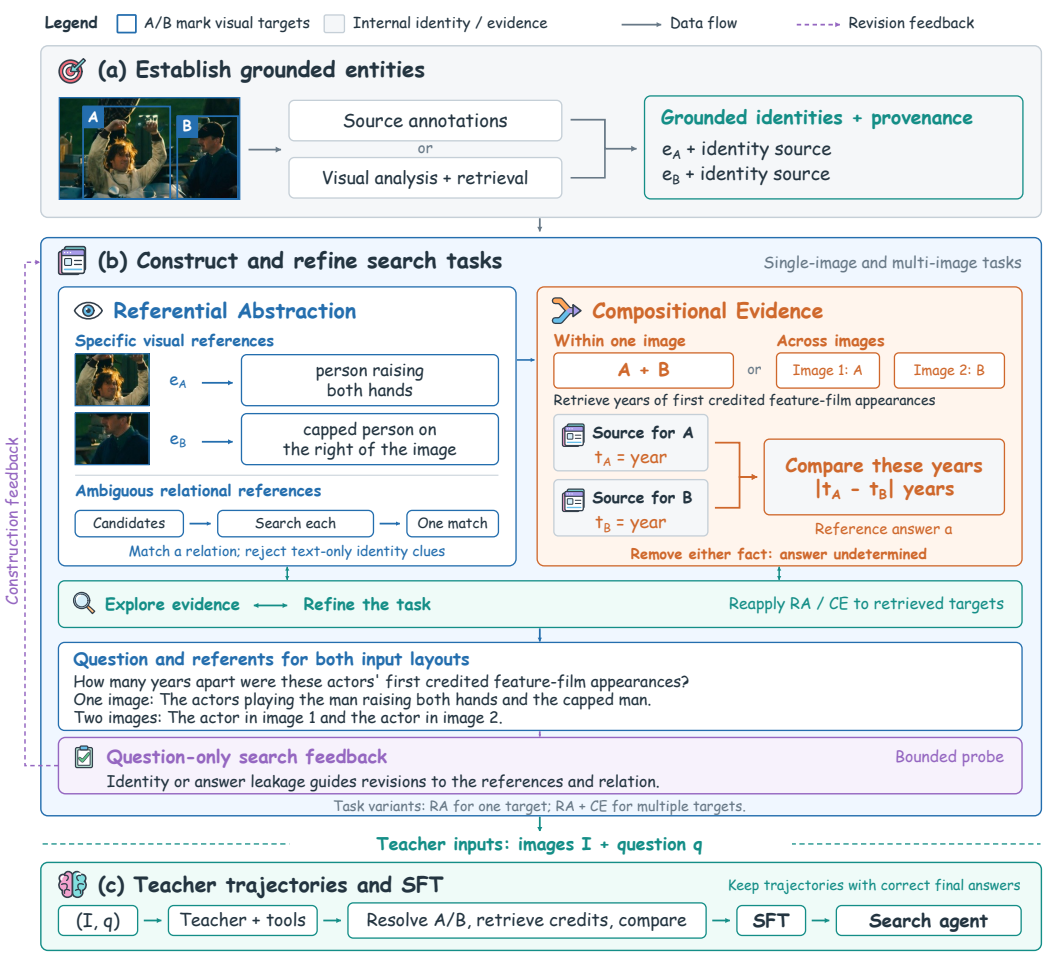


Figure 2: **RACE task synthesis and training.** (a) Source annotations or visual analysis and retrieval establish target identities and provenance. (b) RA expresses targets through specific visual or ambiguous relational references. CE combines evidence from multiple targets within or across images. The example compares the years of the actors' first credited feature-film appearances, yielding $|t_A - t_B|$. Evidence exploration and question-only search feedback refine the task. (c) The teacher receives $(\mathcal{I}, q)$ and uses tools to solve the task; trajectories with correct final answers are retained for SFT.

evidence F, and defines an information requirement c. These inputs determine the generated question and reference answer through

$$(q, a) = G(\mathcal{I}, S, R, c, F), \qquad T = (\mathcal{I}, q, a), \tag{1}$$

where q is the question and a its reference answer. Evidence exploration and question construction can alternate as supported requirements emerge. Search begins from the source targets and can reveal intermediate targets along the resulting evidence paths. The solver receives only $(\mathcal{I}, q)$. Target identities, their reference correspondences, construction evidence, and the reference answer remain internal. A question-only search probe supplies additional feedback on identity or answer leakage beyond the recorded solution. The configurations, refinement procedure, and synthesis algorithm are detailed in Appendix C.

3.2 Referential Abstraction

RA supports two forms of reference for candidates $E(V)$ in an initial or retrieved image V. For $\rho = (\nu, \phi)$, ν specifies visible cues and ϕ is an optional condition checked against retrieved candidate

evidence F_e. Define the candidate set and resolved matches as

$$\begin{aligned} \mathcal{C}(V, \nu) &= \{e \in E(V) : e \models_V \nu\}, \\ \mathcal{M}(V, \rho; F) &= \{e \in \mathcal{C}(V, \nu) : \phi(e; F_e) = 1\}. \end{aligned} \tag{2}$$

Both forms require a unique match: $\mathcal{M}(V, \rho; F) = \{e^\star\}$.

Specific visual references. Here $\phi \equiv 1$ and $\mathcal{C}(V, \nu) = \{e^\star\}$: appearance, position, order, spatial relations, or occlusion locate the target before retrieval. Examples include "the third person from the left," "the person to the left of the one in blue," and "the person partly hidden behind the post."

Ambiguous relational references. Here $|\mathcal{C}(V, \nu)| > 1$, and ϕ selects among visually plausible candidates using retrieved relations or histories. For example, "the man in the image who appeared in a crime film in year y" requires checking the candidates' filmographies. Construction must search every candidate, support one match, and rule out all alternatives; unresolved cases are rejected. The condition must remain non-unique outside the image; a question-only search probe checks for identity or answer leakage.

RA excludes direct identity disclosures and pairs the resolved reference with a supported information request. Its candidate searches establish the referent; CE combines facts about resolved targets into a joint answer. Construction identities and evidence remain internal.

3.3 Compositional Evidence

Compositional Evidence constructs joint questions from retrieved facts about multiple targets. For source images $I_1, \ldots, I_m$, the selected targets are

$$S_i \subseteq E_i, \qquad S = \bigcup_{i=1}^{m} S_i, \qquad |S| \geq 2. \tag{3}$$

Within-image composition uses $m = 1$ and $|S_1| \geq 2$. Cross-image composition uses $m \geq 2$ and $S_i = \{e_i\}$, with a distinct grounded target from each image and $\big|\{i : S_i \neq \varnothing\}\big| \geq 2$. The generator searches each target and records its facts, sources, and image provenance. It then connects the evidence through compatible quantities, shared relations, or intersecting constraints. If no supported joint relation can be constructed, the target or image selection is revised or discarded.

Let $z_1, \ldots, z_k$ denote the selected evidence contributions, with each target supplying at least one. CE constructs

$$a = c(z_1, \ldots, z_k), \qquad k \geq 2. \tag{4}$$

For example, the years of two actors' first credited feature-film appearances support a year-difference question whether they are referenced in one image or separate images. The generated question refers to every required target and image, and the recorded solution links each contribution to the answer.

Let $[k] = \{1, \ldots, k\}$ and let $\mathcal{Y}_c(U)$ denote answers consistent with the question and designated facts indexed by $U \subseteq [k]$. The construction requires joint sufficiency and individual necessity:

$$\mathcal{Y}_c([k]) = \{a\}, \qquad \big|\mathcal{Y}_c([k] \setminus \{j\})\big| > 1 \quad \text{for every } j \in [k]. \tag{5}$$

All designated facts determine one answer; removing any one leaves it undetermined under the constructed requirement. Each image must supply necessary evidence through its target, so removing an image's contributions must also leave the answer undetermined. CE checks source support and recorded dependencies, revising or rejecting unsupported relations and redundant contributions. These conditions concern the constructed evidence requirement; bounded shortcut probes assess additional solution paths (Appendix C). Both configurations also apply to targets introduced by retrieved images.

3.4 Applying the Operations during Search

A retrieved image can supply answer evidence or determine the next search target. To obtain answer evidence, the generator retrieves an image J associated with a target and identifies a relevant visual attribute $v(J)$, such as a color, spatial configuration, or inscription. The question identifies J through

its relation to the target and asks the solver to acquire the attribute. Its value can answer the question directly or contribute to a joint answer through

$$a = v(J) \qquad \text{or} \qquad a = c\big(v(J), z_2, \ldots, z_k\big). \tag{6}$$

When a retrieved scene determines the next search target, the generator applies Referential Abstraction again. A reference ρ_W selects a target W in an input image. Retrieval supplies a scene J, in which a reference $\rho_{W'}$ selects a further target W' for subsequent search:

$$W \xrightarrow{\text{retrieve}} J \xrightarrow{\rho_{W'}} W' \xrightarrow{\text{search}} z(W'). \tag{7}$$

The retrieved fact $z(W')$ can answer a single-target question or contribute to a composed answer $a = c\big(z(W'), z_2, \ldots, z_k\big)$. The reference in J specifies which depicted entity supplies the next information requirement. The operations also assess whether each required image can be located from preceding evidence and what it establishes for the answer or a later step. As requirements expand, an earlier answer can become an intermediate fact. A retrieved image can serve both roles, and information from several newly observed targets can be composed.

3.5 Teacher Trajectories and Supervised Fine-Tuning

Qwen3.7-Plus solves retained tasks with search tools, receiving only their images and questions. For N retained tasks $T_n = (\mathcal{I}_n, q_n, a_n)$, teacher execution and trajectory selection give

$$\tau_n \sim \pi_{\mathrm{T}}(\cdot \mid \mathcal{I}_n, q_n), \qquad \mathcal{D} = \{(\mathcal{I}_n, q_n, \tau_n) : g_{\mathrm{traj}}(\tau_n, a_n) = 1\}. \tag{8}$$

Here $g_{\mathrm{traj}}(\tau_n, a_n) = 1$ requires a protocol-valid trajectory whose final answer is judged correct against a_n. Each trajectory records assistant outputs and tool observations in execution order, capturing how visual references guide retrieval and how evidence supports the answer. We train Qwen3.5-27B on 12K synthesized trajectories in one SFT stage. For trajectory tokens $u_1, \ldots, u_L$, we minimize the loss over assistant-output positions $\mathcal{A}(\tau)$,

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(\mathcal{I},q,\tau)\sim\mathcal{D}} \left[\sum_{t \in \mathcal{A}(\tau)} \log p_\theta(u_t \mid \mathcal{I}, q, u_{<t}) \right]. \tag{9}$$

Tool observations provide context, while assistant tool calls and responses provide supervised outputs. The questions and demonstrations in $\mathcal{D}$ carry the construction requirements into training.

4 Experiments

We evaluate RACE under a shared student model and interaction environment. We first compare answer accuracy before and after SFT, then analyze how training changes tool use and visual revisiting.

4.1 Experimental Setup

Training and evaluation. We train Qwen3.5-27B in one SFT stage using only 12K trajectories generated by Qwen3.7-Plus from our synthesized tasks. Assistant tool calls and responses are supervised, while tool observations provide context. The primary evaluation covers MMSearch-Plus (Tao et al., 2025), MM-BrowseComp (Li et al., 2025), BrowseComp-V^3 (Zhang et al., 2026a), and VisBrowse-Bench (Zhang et al., 2026e). These benchmarks assess fine-grained visual-cue use, multimodal retrieval and reasoning, and evidence acquisition and verification during browsing. We report final-answer accuracy under each benchmark's scoring protocol, separately from behavioral measures. Evaluation sets are specified in Appendix A. Supplementary comparisons cover BrowseComp-VL, VDR, MMSearch, SimpleVQA (Cheng et al., 2025), LiveVQA (Fu et al., 2025), and FVQA (Wang et al., 2016) (Table 1). Avg. weights all ten benchmarks equally and is reported only for rows with complete aggregate scores.

Table 1: **Search accuracy and factual generalization** (%, ↑). Results cover four primary search benchmarks and six supplementary benchmarks. Models are grouped by role. The final block compares Qwen3.5-27B before and after SFT under matched settings. Avg. is the unweighted mean over all ten benchmarks.

	Multimodal search						Factual generalization				
Model / setting	MM-BC	MS+	BC-V^3	VisB	BC-VL	VDR	MM Search	Simple VQA	Live VQA	FVQA	**Avg.**
Frontier API and large-backbone references											
Qwen3.5-397B-A17B	46.0	42.1	31.0	52.7	49.4	—	78.3	76.4	85.2	77.1	—
Qwen3.7-Plus	37.1	46.3	42.3	59.8	55.6	30.2	80.7	77.7	84.6	79.7	59.4
Qwen3.8-Max	53.2	59.1	56.3	66.9	—	33.2	84.0	83.8	88.8	85.6	—
Gemini-3.1-Pro	48.7	51.5	50.0	62.1	65.9	40.8	86.1	81.0	87.3	84.0	65.7
Specialized multimodal search agents (30B–32B)											
WebWatcher-32B	—	—	8.7	—	27.0	—	55.3	59.0	58.7	—	—
SenseNova-MARS-32B	—	—	—	—	—	—	74.3	74.1	60.8	72.6	—
Vision-DeepResearch-30B-A3B	—	28.5	—	—	53.7	37.8	69.6	—	77.6	74.2	—
OpenSearch-VL-30B-A3B	—	—	—	—	41.1	33.5	68.7	74.9	67.4	73.2	—
REDSearcher-MM-RL-30B-A3B	23.5	26.6	—	—	57.2	—	72.9	—	79.3	—	—
Skywork-R1V4-30B-A3B	—	—	—	—	38.4	—	66.1	—	—	67.2	—
MM-DeepResearch-32B	—	—	—	—	43.0	—	69.0	67.6	68.0	70.1	—
LMM-Searcher-30B-A3B (30 turns)	22.3	32.9	—	42.0	—	—	71.0	—	—	—	—
same model, 100 turns	30.1	34.8	—	48.3	—	—	72.3	—	—	—	—
SimpleSearch-VL-30B-A3B	—	34.4	—	—	55.9	—	83.6	79.6	81.1	79.0	—
Our study with the Qwen3.5-27B backbone under identical settings											
Qwen3.5-27B (base agent)	31.3	37.6	31.7	48.5	47.9	24.4	79.3	74.5	84.5	75.7	53.5
RACE-27B (ours)	**33.5**	**40.2**	**36.7**	**51.5**	**54.2**	**27.4**	78.7	**75.5**	**85.0**	**76.8**	**56.0**
Δ	+2.2	+2.6	+5.0	+3.0	+6.3	+3.0	-0.6	+1.0	+0.5	+1.1	+2.5

Notes. Sources and evaluation settings are detailed in Appendix A. Abbreviations are MM-BC (MM-BrowseComp), MS+ (MMSearch-Plus), BC-V^3 (BrowseComp-V^3), VisB (VisBrowse-Bench), and BC-VL (BrowseComp-VL). Avg. is the unweighted mean of all ten aggregate scores. Published scores retain their source-specific evaluation settings. Here — denotes unavailable scores. LMM-Searcher uses Thinking initialization and the MS+ single-image subset. Turn budgets are listed by row. Δ denotes SFT minus matched base-agent performance, computed from displayed scores, in percentage points.

Comparisons. Baselines include frontier API models, large backbones, and agents trained for multimodal search. Our Qwen3.5-27B base and SFT agents share test sets, tool access, decoding settings, inference budgets, and scoring settings. This matched comparison isolates the gain from SFT. Published agents, including Vision-DeepResearch (Huang et al., 2026b), retain their reported model sizes and training regimes. The comparison combines our evaluations with published scores. The Gemini-3.1-Pro row contains both. Score sources and evaluation variants are recorded in Appendix A.

4.2 Main Results

RACE consistently improves performance over the base agent. SFT improves Qwen3.5-27B on all four primary benchmarks by 2.2–5.0 percentage points, averaging 3.2 points under matched evaluation settings (Table 1). The gains extend to nine of the ten evaluated benchmarks, indicating that the benefit of RACE supervision is not confined to the primary evaluation.

Gains are larger on search-oriented benchmarks. The largest improvements occur on BrowseComp-VL and BrowseComp-V^3, at 6.3 and 5.0 points, respectively, compared with 0.5–1.1 points on SimpleVQA, LiveVQA, and FVQA (Table 1). This pattern is consistent with RACE's focus on supervision for visual search and evidence integration. The benefit varies across search benchmarks: MMSearch is the only benchmark where performance decreases, by 0.6 points.

4.3 Training-Data Comparison

RACE supervision outperforms the compared open-source training mixture. With the same Qwen3.5-27B backbone, SFT on only 12K RACE trajectories exceeds SFT on only 28K open-source trajectories by 0.4–3.0 percentage points across all four primary benchmarks (Table 2). The 28K open-source training mixture combines data from OpenSearch-VL, Vision-DeepResearch, and REDSearcher-MM.

4.4 Training-Trajectory Characteristics

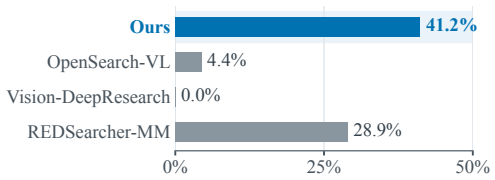
Visual tools comprise image search, text-to-image search, and image cropping; calls within one assistant message form a tool round.

RACE trajectories more often use multiple input images. At least two images occur in the initial input of 41.2% of RACE trajectories, compared with at most 28.9% across the three training-source baselines (Figure 3).

Table 2: **Training-data comparison** (%, ↑). Each SFT variant uses only the indicated training data.

Training setting	MM-BC	MS+	BC-V^3	VisB
Qwen3.5-27B (Agentic)	31.3	37.6	31.7	48.5
Open-source only (28K)	32.1	37.9	36.3	48.5
Ours only (12K)	**33.5**	**40.2**	**36.7**	**51.5**

Figure 3: **Multi-image input.** Trajectory proportions across training sources.



Visual calls occur more often after the first round. Visual-tool use extended beyond the first round in 34.6% of Ours trajectories, compared with 6.0%, 5.6%, and 1.7% for OpenSearch-VL, Vision-DeepResearch, and REDSearcher-MM, respectively (Figure 4(c)). This higher rate persisted within each of four round-count ranges (Appendix B.4).

Visual–web–visual sequences are more frequent. A visual call followed by web search or page fetching and another visual call occurred in 28.8% of Ours trajectories, compared with 4.8%, 4.5%, and 1.4% for the same baselines (Figure 4(b)). This visual–web–visual pattern follows serialized call order and allows intervening calls. Ours also had a larger share of trajectories with multiple visual calls (Figure 4(a)) and used more distinct tool types per trajectory on average: 3.16 versus 2.86–2.93 (Figure 4(c)). Metric definitions and supplementary visual-span statistics are provided in Appendices B.1 and B.5.

4.5 Evaluation-Time Search Behavior

SFT increases visual-call share and total tool use. After SFT, the student used visual tools beyond the first round more often on all four held-out benchmarks, with increases of 12.5–19.5 points (Figure 5(c)). Visual-call share, the within-trajectory fraction of visual calls averaged across trajectories, also increased on every benchmark, by 5.4–12.3 points (Figure 5(b)). These changes accompanied an increase of 3.0–20.4 tool calls per trajectory, largest on MM-BrowseComp (Figure 5(a)).

Visual use after the first round increases across task structures. Across 790 matched question pairs grouped by dependency depth, checklist length, and input-image count, the SFT agent has higher point estimates of post-first-round visual use in all nine strata (Tables 4–6). The increases are not monotonic in these structural measures: on MMSearch-Plus, the difference is 24.1 points for two-image questions and 5.7 points for questions with three or more images. Paired 95% confidence intervals include zero in three strata (Figure 6, Appendix B.2).

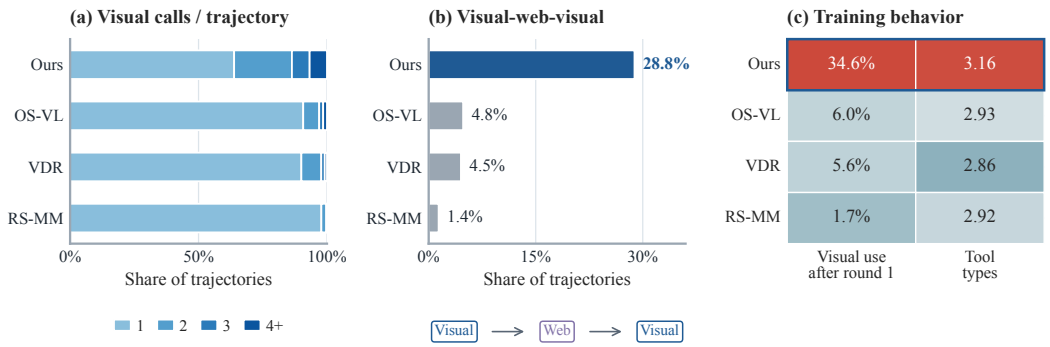


Figure 4: **Training trajectory distributions and behavior.** (a) Visual calls per trajectory. (b) Proportion of trajectories containing a visual–web–visual pattern: a visual call, web search or page fetching, and another visual call in serialized order, with intervening calls allowed. (c) The proportion of trajectories with visual calls after the first tool round and mean number of distinct tool types. Colors are standardized within each metric, and the blue outline marks Ours. Abbreviations: OS-VL, OpenSearch-VL; VDR, Vision-DeepResearch; RS-MM, REDSearcher-MM.

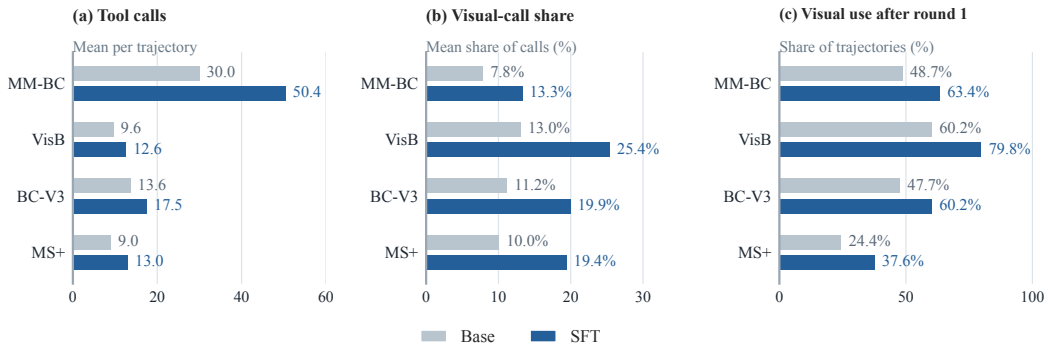


Figure 5: **Evaluation-time search behavior before and after SFT.** Gray and blue bars give base-agent and SFT-agent point estimates, respectively. (a) Mean tool calls per trajectory. (b) Mean within-trajectory fraction of tool calls that use visual tools. (c) Fraction of trajectories that use visual tools after the first tool round. Panels (b) and (c) show percentages. Abbreviations: MM-BC, MM-BrowseComp; VisB, VisBrowse-Bench; BC-V3, BrowseComp-V^3; MS+, MMSearch-Plus.

5 Conclusion

We introduced RACE, an image-conditioned task synthesis framework designed to make visually grounded search necessary for determining the answer. Referential Abstraction ties target resolution to image-grounded references, and Compositional Evidence connects the resolved targets' evidence to a joint answer that requires every designated contribution. Both operations apply to initial and retrieved images, so new observations can establish further targets or contribute answer evidence. Grounding and dependency feedback refine the two operations during synthesis, with question-only probes identifying additional shortcuts. Fine-tuning Qwen3.5-27B on only 12K teacher-generated trajectories raises average accuracy across ten multimodal benchmarks from 53.5 to 56.0, a gain of 2.5 points over the matched base agent. The framework shapes supervision by constructing dependencies between images, targets, evidence, and answers. Its effectiveness depends on resolvable references, accessible evidence, and teacher and tool capabilities. Shortcut probes cannot exclude every alternative solution.

AI USE STATEMENT

Generative AI assisted task synthesis and manuscript preparation, including drafting, language editing, and reference lookup. In the synthesis pipeline, models ground visual targets, construct questions and reference answers, and support quality checks and shortcut probes (Section 3 and Appendix C). Qwen3.7-Plus generates the search trajectories used for supervised fine-tuning. The authors take responsibility for the manuscript, synthesized data, and reported findings.

ETHICS STATEMENT

RACE uses real images and web evidence, which may contain copyrighted material or information about identifiable people. Their use and redistribution must respect source licenses, platform terms, and privacy. Synthetic questions, reference answers, and trajectories may inherit factual errors or biases from the source material and generative models. The construction criteria and probes address task grounding, answer support, and shortcuts. They do not constitute a privacy or safety audit. Applications involving personal information or consequential decisions require additional review.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix C describe the synthesis procedure, screening, and SFT objective. Section 4.1 and Appendix A specify the evaluated models, benchmark sets, score sources, and decoding and iteration-limit labels from archived runs. Appendix B documents trajectory selection, behavioral metrics, and uncertainty estimation. We plan to release the synthesis and analysis code, prompts, and redistributable task and trajectory data, together with the data and scripts underlying the reported tables and figures. Re-running synthesis and evaluation requires access to the corresponding models, image assets, and search services. Results may vary with changes in web content, retrieval results, and hosted models.

REFERENCES

Shuang Chen, Kaituo Feng, Hangting Chen, Wenxuan Huang, Dasen Dai, Quanxin Shou, Yunlong Lin, Xiangyu Yue, Shenghua Gao, and Tianyu Pang. OpenSearch-VL: An Open Recipe for Frontier Multimodal Search Agents, 2026. URL https://arxiv.org/abs/2605.05185. arXiv:2605.05185.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions?, 2023. URL https://arxiv.org/abs/2302.11713.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, Yutao Zeng, Zhoufutu Wen, Ke Jin, Baorui Wang, Weixiao Zhou, Yunhong Lu, Hangyuan Ji, Tongliang Li, Wenhao Huang, and Zhoujun Li. SimpleVQA: Multimodal Factuality Evaluation for Multimodal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4637–4646, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cheng_SimpleVQA_Multimodal_Factuality_Evaluation_for_Multimodal_Large_Language_Models_ICCV_2025_paper.html.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, Gao Huang, Dahua Lin, and Lewei Lu. SenseNova-MARS: Empowering Multimodal Agentic Reasoning and Search via Reinforcement Learning, 2025. URL https://arxiv.org/abs/2512.24330.

Zheng Chu, Xiao Wang, Jack Hong, Huiming Fan, Yuqi Huang, Yue Yang, Guohai Xu, Chenxiao Zhao, Cheng Xiang, Shengchao Hu, Dongdong Kuang, Ming Liu, Bing Qin, and Xing Yu. REDSearcher: A Scalable and Cost-Efficient Framework for Long-Horizon Search Agents, 2026. URL https://arxiv.org/abs/2602.14234.

Ming Dai, Zhihong Lu, Jinjie Gu, Jiedong Zhuang, Yefeng Liu, Wankou Yang, Jian Wang, and Chunhua Shen. SimpleSearch-VL: A Simple Recipe for Multimodal Agentic Deep Search, 2026. URL https://arxiv.org/abs/2606.31504.

Yifan Du, Zikang Liu, Jinbiao Peng, Jie Wu, Junyi Li, Jinyang Li, Wayne Xin Zhao, and Ji-Rong Wen. Towards Long-horizon Agentic Multimodal Search, 2026. URL https://arxiv.org/abs/2604.12890v1. arXiv:2604.12890v1.

Mingyang Fu, Yuyang Peng, Benlin Liu, Yao Wan, and Dongping Chen. LiveVQA: Live Visual Knowledge Seeking, 2025. URL https://arxiv.org/abs/2504.05288v1.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. WebWatcher: Breaking New Frontiers of Vision-Language Deep Research Agent, 2025. URL https://arxiv.org/abs/2508.05748v1. arXiv:2508.05748v1.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *The European Conference on Computer Vision (ECCV)*, 2020.

Shijue Huang, Hangyu Guo, Guanting Dong, Chenxin Li, Junting Lu, Xinyu Geng, Zhaochen Su, Zhenyu Li, Shuang Chen, Hongru Wang, and Yi R. Fung. Towards On-Policy Data Evolution for Visual-Native Multimodal Deep Search Agents, 2026a. URL https://arxiv.org/abs/2605.10832. arXiv:2605.10832.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, Zehui Chen, Xu Tang, Yao Hu, Shaohui Lin, Philip Torr, Feng Zhao, and Wanli Ouyang. Vision-DeepResearch: Incentivizing DeepResearch Capability in Multimodal Large Language Models, 2026b. URL https://arxiv.org/abs/2601.22060. arXiv:2601.22060.

Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering, 2019. URL https://arxiv.org/abs/1902.09506. arXiv:1902.09506.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. MMSearch: Benchmarking the Potential of Large Models as Multi-modal Search Engines, 2024. URL https://arxiv.org/abs/2409.12959.

Guankai Li, Jiabin Chen, Yi Xu, Xichen Zhang, and Yuan Lu. HyperEyes: Dual-Grained Efficiency-Aware Reinforcement Learning for Parallel Multimodal Search Agents, 2026. URL https://arxiv.org/abs/2605.07177v2. arXiv:2605.07177v2.

Shilong Li, Xingyuan Bu, Wenjie Wang, Jiaheng Liu, Jun Dong, Haoyang He, Hao Lu, Haozhe Zhang, Chenchen Jing, Zhen Li, Chuanhao Li, Jiayi Tian, Chenchen Zhang, Tianhao Peng, Yancheng He, Jihao Gu, Yuanxing Zhang, Jian Yang, Ge Zhang, Wenhao Huang, Wangchunshu Zhou, Zhaoxiang Zhang, Ruizhe Ding, and Shilei Wen. MM-BrowseComp: A Comprehensive Benchmark for Multimodal Browsing Agents, 2025. URL https://arxiv.org/abs/2508.13186v1. arXiv:2508.13186v1.

Yikun Liu, Yuan Liu, Haicheng Wang, Zhongyin Zhao, Le Tian, Xiao Zhou, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. POINTS-Seeker: An Open Recipe for Multimodal Search Agents with Visual Memory Management, 2026. URL https://arxiv.org/abs/2604.14029.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge, 2019. URL https://arxiv.org/abs/1906.00067.

Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M. Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. DeepMMSearch-R1: Empowering Multimodal LLMs in Multimodal Web Search, 2025. URL https://arxiv.org/abs/2510.12801.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge, 2022. URL https://arxiv.org/abs/2206.01718.

Xijia Tao, Yihua Teng, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. MMSearch-Plus: Benchmarking Provenance-Aware Search for Multimodal Browsing Agents, 2025. URL https://arxiv.org/abs/2508.21475.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. FVQA: Fact-based Visual Question Answering, 2016. URL https://arxiv.org/abs/1606.05433.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. MMSearch-R1: Incentivizing LMMs to Search, 2025. URL https://arxiv.org/abs/2506.20670.

Zhiwei Yang, Yuanchen Wu, Nan Zhang, Yucong Meng, Ke Yan, and Shouhong Ding. Scene Graph Thinking: Reinforcing Structured Visual Reasoning for Multimodal Large Language Models, 2026. URL https://arxiv.org/abs/2607.05716. arXiv:2607.05716.

Huanjin Yao, Qixiang Yin, Min Yang, Ziwang Zhao, Yibo Wang, Haotian Luo, Jingyi Zhang, and Jiaxing Huang. MM-DeepResearch: A Simple and Effective Multimodal Agentic Search Baseline, 2026. URL https://arxiv.org/abs/2603.01050. arXiv:2603.01050.

Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, Shiting Huang, Yiming Zhao, Xu Tang, Yao Hu, Philip Torr, Wanli Ouyang, and Shaosheng Cao. Vision-DeepResearch Benchmark: Rethinking Visual and Textual Search for Multimodal Large Language Models, 2026. URL https://arxiv.org/abs/2602.02185.

Huanyao Zhang, Jiepeng Zhou, Bo Li, Bowen Zhou, Yanzhe Shan, Haishan Lu, Zhiyong Cao, Jiaoyang Chen, Yuqian Han, Zinan Sheng, Zhengwei Tao, Hao Liang, Jialong Wu, Yang Shi, Yuanpeng He, Jiaye Lin, Qintong Zhang, Guochen Yan, Runhao Zhao, Zhengpin Li, Xiaohan Yu, Lang Mei, Chong Chen, Wentao Zhang, and Bin Cui. BrowseComp-V^3: A Visual, Vertical, and Verifiable Benchmark for Multimodal Browsing Agents, 2026a. URL https://arxiv.org/abs/2602.12876.

Huanyao Zhang, Jiepeng Zhou, Runhao Zhao, Yanzhe Shan, Jiaoyang Chen, Bowen Zhou, Bo Li, Fang Wang, Jialong Wu, Zhengwei Tao, Lang Mei, Xiaohan Yu, Liyan Liu, Chong Chen, and Wentao Zhang. DeepVoyager-VL: Incentivizing Vision-in-the-Loop Search for Long-Horizon Multimodal Agents, 2026b. URL https://arxiv.org/abs/2608.01827. arXiv:2608.01827.

Ruiyang Zhang, Qianguo Sun, Chao Song, Yiyan Qi, and Zhedong Zheng. VSearcher: Long-Horizon Multimodal Search Agent via Reinforcement Learning, 2026c. URL https://arxiv.org/abs/2603.02795.

Yifan Zhang, Liang Hu, Haofeng Sun, Peiyu Wang, Yichen Wei, Shukang Yin, Jiangbo Pei, Wei Shen, Peng Xia, Yi Peng, Tianyidan Xie, Eric Li, Yang Liu, Xuchen Song, and Yahui Zhou. Skywork-R1V4: Toward Agentic Multimodal Intelligence through Interleaved Thinking with Images and DeepResearch, 2025. URL https://arxiv.org/abs/2512.02395.

Zhengbo Zhang, Changtao Miao, Jinbo Su, Zhaowen Zhou, Chunxia Zhang, Xukai Wang, Ruiqi Liu, Kaiyuan Zheng, Jiansheng Cai, Bo Zhang, Zhe Li, Shiming Xiang, and Ying Yan. Visual-Seeker: Towards Visual-Native Multimodal Agentic Search via Active Visual Reasoning, 2026d. URL https://arxiv.org/abs/2606.15231. arXiv:2606.15231.

Zhengbo Zhang, Jinbo Su, Zhaowen Zhou, Changtao Miao, Yuhan Hong, Qimeng Wu, Yumeng Liu, Feier Wu, Yihe Tian, Yuhao Liang, Zitong Shan, Wanke Xia, Yi-Fan Zhang, Bo Zhang, Zhe Li, Shiming Xiang, and Ying Yan. VisBrowse-Bench: Benchmarking Visual-Native Search for Multimodal Browsing Agents, 2026e. URL https://arxiv.org/abs/2603.16289v1. arXiv:2603.16289v1.

A EXPERIMENTAL PROTOCOLS

The main agent is trained exclusively on 12K synthesized trajectories.

A.1 TRAINING AND EVALUATION CONFIGURATION

Teacher trajectories. Each synthesized task contains images, a question, and a supported reference answer. Qwen3.7-Plus receives only the images and question and uses search tools to solve the task. The reference answer and synthesis records remain hidden. Only protocol-valid trajectories whose final answers are judged correct against the reference answers are retained.

Supervised fine-tuning. We train Qwen3.5-27B in one SFT stage on the 12K retained trajectories. The objective minimizes negative log-likelihood over assistant tool calls and responses (Equation 9). Tool observations provide context and are excluded from the supervised token positions.

Table 3: **Training setup and archived evaluation settings.** Decoding and iteration-limit labels refer to the archived runs used for behavioral analysis.

Field	Specification
Teacher and student	Qwen3.7-Plus and Qwen3.5-27B
Main training data	12K synthesized trajectories only
Training stage and supervision	One SFT stage over assistant tool calls and responses
Trajectory-run decoding labels	`temp0.7`, `top-p0.9`, `pp1.5`
Trajectory-run iteration-limit labels	MM-BC (100), MS+ (20), BC-V^3 (20), VisB (15)
Primary benchmark questions	MM-BC (224), MS+ (311), BC-V^3 (300), VisB (169), using full sets

Benchmark versions and scoring. The base and SFT agents use the original 224-question MM-BrowseComp release (Li et al., 2025) and the full evaluation sets of MMSearch-Plus (311 questions), BrowseComp-V^3 (300), and VisBrowse-Bench (169) (Tao et al., 2025; Zhang et al., 2026a;e). MMSearch-Plus includes single-image and multi-image questions. Both agents use the same questions and scoring settings, with final-answer accuracy measured under each benchmark's protocol. BrowseComp-VL, VDR, MMSearch, SimpleVQA (Cheng et al., 2025), LiveVQA (Fu et al., 2025), and FVQA (Wang et al., 2016) provide supplementary comparisons. Avg. in Table 1 is the unweighted mean over all ten benchmarks and requires complete aggregate scores. The training-data comparison uses the four primary scores without an average (Table 2). The mean gain in Section 4.2 weights these benchmarks equally. Published results retain their original protocol labels. Missing scores are not imputed, and averages use a common benchmark set.

Score sources. We report our evaluations of Qwen3.5-397B-A17B, Qwen3.7-Plus, and Qwen3.8-Max. Our Gemini-3.1-Pro evaluations cover MM-BrowseComp, MMSearch, and FVQA. WebWatcher, SenseNova-MARS, Vision-DeepResearch, OpenSearch-VL, and MM-DeepResearch use their reported agent scores (Geng et al., 2025; Chng et al., 2025; Huang et al., 2026b; Chen et al., 2026; Yao et al., 2026), as do Skywork-R1V4, REDSearcher-MM-RL, and SimpleSearch-VL (Zhang et al., 2025; Chu et al., 2026; Dai et al., 2026). LMM-Searcher uses Thinking initialization with separate 30- and 100-turn settings. Its MMSearch-Plus scores cover the single-image subset (Du et al., 2026). Seven Gemini-3.1-Pro scores (MS+, BC-V^3, VisB, BC-VL, VDR, SimpleVQA, and LiveVQA) come from the agentic-workflow evaluation of Zhang et al. (2026b). That compilation also supplies WebWatcher's BC-V^3 score.

Archived inference runs. The base and SFT runs used for behavioral analysis share the per-benchmark decoding and iteration-limit labels in Table 3 and the system template `SYSTEM_TEMPLATE_RED_RESEARCHER_MM`. We use `pass0` and normalize tool aliases to web search, page fetching and summarization, image search, text-to-image search, and image cropping.

B Dataset and Behavioral Analysis

B.1 Metric Computation and Aggregation

Tool normalization. Tool aliases are mapped to web search, web fetching and summarization, image search, text-to-image search, and image cropping. For the measures in Sections 4.4 and 4.5, calls within an assistant message are counted individually and retain their serialized order.

Per-trajectory statistics. Tool-call count is the total number of calls C, and visual-call share is the number of visual calls divided by C for $C > 0$. Tool diversity counts the distinct tool types used. Reported means average the defined per-trajectory values within each population.

Trajectory indicators. Post-first-round visual use indicates whether a trajectory contains a visual call after its first tool round. For the visual–web–visual pattern in Section 4.4, we test whether serialized call indices $i < j < k$ exist such that calls i and k use visual tools and call j uses web search or page fetching and summarization. Each trajectory contributes a binary indicator, regardless of how many times the pattern occurs; trajectories without visual calls contribute zero. For the supplementary visual–search–fetch pattern in Appendix B.2, the three ordered calls correspond to a visual tool, web search, and page fetching. Both patterns allow intervening calls and calls within the same assistant message; serialized order does not establish that a later call uses an earlier call's result. Multi-image input indicates at least two image placeholders in the initial user message. We report each indicator as the proportion of trajectories satisfying it.

B.2 Task-Stratified Search Behavior

Population and task strata. We match pass0 evaluation trajectories by benchmark and query identifier, retaining one saved successful execution per model and question. Each bin therefore contains the same questions for both models. Success refers to execution status. Answer correctness is not a selection criterion. BrowseComp-V^3 contributes 284 pairs, grouped by the longest dependency chain in its annotated subgoal graph, counting nodes. The depth bins are 2–3 ($n = 107$), 4–5 ($n = 123$), and 6 or more ($n = 54$). MM-BrowseComp contributes 195 displayed pairs with 2 ($n = 63$), 3 ($n = 76$), or at least 4 ($n = 56$) checklist items. The seven pairs with one item are omitted. MMSearch-Plus contributes 311 pairs with 1 ($n = 222$), 2 ($n = 54$), or at least 3 ($n = 35$) input images. These labels come from benchmark metadata, independently of agent tool use. VisBrowse is omitted because its available metadata provide domain labels without comparable structural annotations.

Table 4: **BrowseComp-V^3 search behavior** (%). Base versus trained by dependency depth.

	Post-first-round visual use		Visual–search –fetch	
Dependency depth	Base	Trained	Base	Trained
2–3	43.9	57.0	61.7	95.3
4–5	46.3	61.0	52.0	93.5
6+	57.4	64.8	51.9	92.6

Table 5: **MM-BrowseComp search behavior** (%). Base versus trained by checklist length.

	Post-first-round visual use		Visual–search –fetch	
Checklist items	Base	Trained	Base	Trained
2	57.1	77.8	42.9	74.6
3	43.4	55.3	51.3	71.1
4+	46.4	62.5	53.6	75.0

Table 6: **MMSearch-Plus search behavior** (%). Base versus trained by input-image count.

	Post-first-round visual use		Visual–search –fetch	
Input images	Base	Trained	Base	Trained
1	22.5	34.2	42.3	83.3
2	27.8	51.9	44.4	98.1
3+	31.4	37.1	34.3	88.6

Tables 4–6 compare matched questions. Visual use follows the first tool round. Ordered visual–search–fetch allows intervening calls. Sampling and uncertainty estimation are described in Appendix B.2.

Post-first-round visual use. Post-first-round visual use had positive SFT–base point differences in all nine strata (Tables 4–6). Paired 95% intervals included zero for dependency depth 6 or more in BrowseComp-V^3, three checklist items in MM-BrowseComp, and three or more images in MMSearch-Plus (Figure 6). For MMSearch-Plus, the difference was 24.1 points with two images and 5.7 points with three or more; the latter had a paired 95% interval of $[-17.1, 28.6]$ points.

Supplementary call pattern. The visual–search–fetch pattern occurred in 92.6–95.3% of trained-agent trajectories on BrowseComp-V^3, compared with 51.9–61.7% for the base agent (Table 4). Occurrence was also higher in every MM-BrowseComp and MMSearch-Plus stratum, with paired differences of 19.7–31.7 and 41.0–54.3 points, respectively (Tables 5 and 6). Paired 95% intervals excluded zero in all nine strata (Figure 6). Longer trajectories offer more opportunities for this pattern to occur.

Uncertainty estimation. For each bin, we compute model-specific proportions and their paired difference. Percentile bootstrap 95% intervals use 4,000 resamples (seed 20260924). Difference intervals resample within-question SFT–base differences to preserve pairing. The intervals are pointwise and capture question-sampling uncertainty conditional on the retained runs.

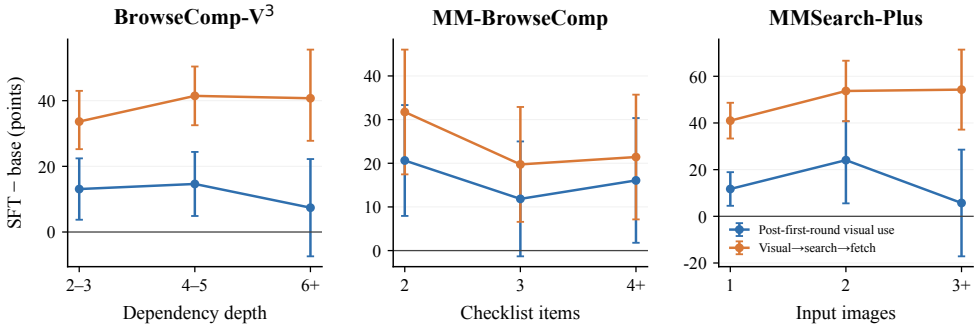


Figure 6: **Paired differences across task strata.** SFT minus base-agent trajectory proportions, in points, for the question pairs in Tables 4–6. Error bars are pointwise paired-bootstrap 95% intervals. The horizontal line marks zero.

B.3 Visual Structure of Training Data

The trajectory analysis compares the analyzed Ours trajectories with 10,752 OpenSearch-VL, 1,393 Vision-DeepResearch, and 16,053 REDSearcher-MM trajectories. Each trajectory contributes once, with generation and filtering settings retained from its source.

The analyzed Ours trajectories exceed the highest baseline rate by 12.2 points for multi-image input (Figure 3).

B.4 Continued Visual Use across Interaction Lengths

We compute continued visual-use rates within four tool-round ranges of 1–5, 6–10, 11–15, and 16 or more. Rates are reported separately for each source and range. Ours exceeds the highest baseline rate by 16.8, 34.3, 32.2, and 26.1 points, respectively (Figure 7).

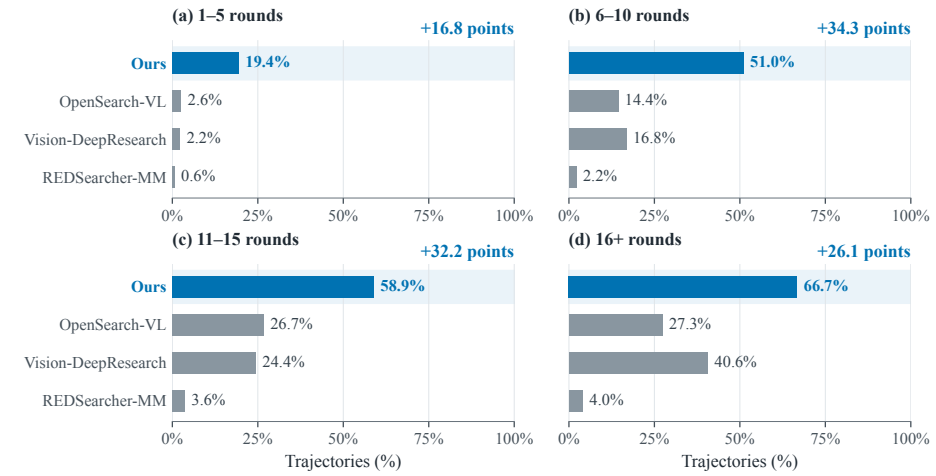


Figure 7: **Continued visual use across interaction lengths.** Bars give the proportion of trajectories with a visual-tool call after the first round, within each round-count range. All panels share a scale. Blue bars denote Ours, and blue annotations give the difference from the highest baseline rate.

B.5 Supplementary Visual-Span Statistics

Normalized visual span measures the separation between the first and last visual calls relative to the entire tool-call sequence. For a trajectory with C tool calls and first and last visual-call indices i_{first} and i_{last}, it is

$$S = \frac{i_{\text{last}} - i_{\text{first}}}{\max(C-1, 1)}.$$

A single visual call gives $S = 0$, and visual calls at both ends of a sequence with $C > 1$ give $S = 1$. The statistic is undefined for trajectories with no visual calls, which are excluded from both the proportions and means in Figure 8. It measures the distance between two calls, not the density of visual calls between them.

The first and last visual calls spanned more than half the tool-call sequence in 21.8% of Ours trajectories with visual calls, compared with 1.0–3.8% for the baselines. Mean normalized visual span was 22.5% for Ours, 4.7% for OpenSearch-VL, 4.9% for Vision-DeepResearch, and 1.0% for REDSearcher-MM.

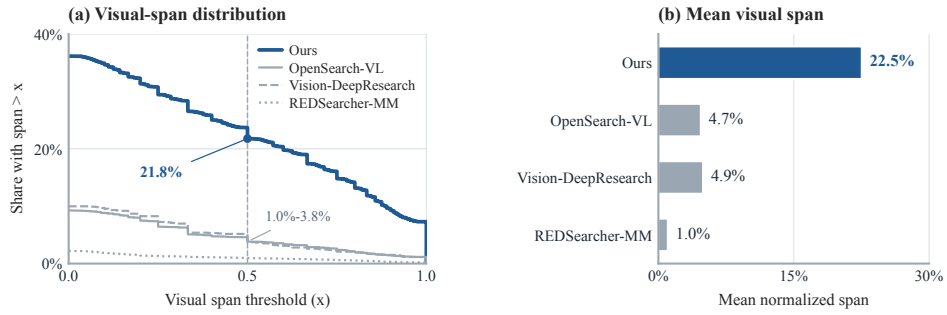


Figure 8: **Supplementary visual-span statistics for training trajectories.** (a) Proportion of trajectories whose normalized visual span exceeds each threshold x; higher curves indicate more trajectories with widely separated visual calls. (b) Mean normalized visual span for each source. Both panels summarize trajectories with at least one visual call; a single visual call has span zero. Baseline names follow Table 1.

C TASK SYNTHESIS DETAILS

Referential Abstraction and Compositional Evidence support single-target, within-image, and cross-image tasks (Table 7). Grounded construction, iterative refinement, and teacher execution form the synthesis and training procedure in Algorithm 1.

Source images include Weibo and X/Twitter photographs, Bilibili video frames, and MovieNet film imagery. Each candidate has a source image and a description locating it in the scene. Ambiguous relational references require coverage of every entity matching the coarse visual description; unrelated scene objects need not be annotated. MovieNet identity annotations can pair frames containing the same person. Cross-image target selection follows Equation 3. Reference objects provide visual context, while search targets initiate retrieval and intermediate entities can emerge during exploration. The task specifies required information without fixing the order of identification and retrieval.

Table 7: **Task configurations.** Referential Abstraction constructs target references, and Compositional Evidence adds joint information requirements. The same configurations apply to targets in retrieved images.

Construction	Target configuration	Required behavior
Single-target abstraction	One target among the candidates in an image	Resolve a specific visual or ambiguous relational reference, then retrieve the requested information.
Within-image composition	Multiple required targets in one image	Retrieve evidence for each target and derive an answer requiring all designated contributions.
Cross-image composition	One required target per image across multiple images	Resolve each image's target and combine its evidence with the other images' contributions.

Reference construction. Specific visual references use appearance, occlusion cues, or positional phrases computed from normalized bounding boxes: a 3×3 grid, left-to-right order within a row, or row membership. Each retained phrase must distinguish its target within the scene; a reference neighbor need not become a separate search target. For an ambiguous relational reference $\rho = (\nu, \phi)$, the construction record must cover every candidate in $\mathcal{C}(V, \nu)$ before selecting a unique match. For each candidate, it must record source evidence and whether the condition is satisfied, ruled out, or unresolved. Missing evidence is unresolved, not evidence of non-satisfaction. Retention requires one supported match and supported exclusion of all alternatives; otherwise the generator revises the condition, retrieves further evidence, or discards the task. Conditions concerning awards, roles, genres, or periods of activity are chosen after examining candidate evidence and must remain insufficient to identify the target through text-only search. These checks specify the intended candidate-elimination path; the bounded shortcut probe does not establish that every possible solver must follow the same search order or inspect every candidate.

Target and image constructors. When enabled, the within-image multi-anchor constructor selects two grounded targets. Cross-image construction grounds one target in each of two images and rejects pairs lacking either identity. The two identities seed separate opening searches. The generator receives both images and the combined evidence, then constructs a relation supported by both targets' facts. Both images remain attached to the final question, with one image placeholder next to each corresponding reference. A same-entity route also pairs two views for reference resolution. Scene records store the selected targets, image associations, spatial cues, and routing decisions. Evidence records store the query, source URL, title, acquisition turn, and modality.

Evidence and image necessity. CE selects the contributions required by the joint question from the retrieved evidence and records their target and image provenance. For every designated contribution z_j, the construction record must explain why the remaining contributions no longer determine a unique answer. For a required image I_i, the same check removes all contributions rooted in its target. For example, the difference between the years of two actors' first credited feature-film appearances cannot be determined from either year alone, regardless of whether the actors appear in one image or two. Image pairs without a supported relation, duplicated evidence, or a target whose contribution

can be omitted must be revised or rejected. These are checks of the recorded construction; generator rationales are not measurements of solver performance under evidence or image removal.

Algorithm 1 Image-conditioned task synthesis and agent training

Require: Image collections, generator G, teacher π_{T}, student π_θ
$\mathcal{D} \leftarrow \varnothing$
for each source image collection $\mathcal{I}$ **do**
  Ground candidate entities E_i in each source image I_i
  Choose S from one or multiple images and ground each selected target
  Explore supporting text and visual evidence F
  Apply RA to construct specific visual or ambiguous relational references R
  For ambiguous references, search all candidates and verify a unique supported match
  For joint tasks, use CE to connect supported facts through a relation c
  Check that removing any required fact or image leaves the answer undetermined
  Apply abstraction and composition as needed to targets in retrieved images
  Specify whether each retrieved image supplies answer evidence or a further search clue
  Generate or expand (q, a) from $(\mathcal{I}, S, R, c, F)$
  Probe for solver shortcuts and combine feedback with RA/CE construction criteria
  if feedback identifies an unresolved reference, support, dependency, or shortcut issue **then**
    Reapply RA/CE to revise (R, c, F) and reassess (q, a), or discard
  end if
  if a valid task is retained **then**
    Execute $(\mathcal{I}, q)$ with π_{T} to obtain trajectory τ
    if $g_{\mathrm{traj}}(\tau, a) = 1$ **then**
      Add $(\mathcal{I}, q, \tau)$ to $\mathcal{D}$
    end if
  end if
end for
Train π_θ on $\mathcal{D}$ using Equation 9

Construction feedback. RA refines references using scene cues and candidate evidence; CE refines relations using source support and the recorded dependencies. The generator returns a question, answer, visual step, and ordered reasoning steps with dependency links. For each input image, it also generates a rationale assessing whether the question would still identify the target or determine the answer after removing that image and its referring clause. Static checks enforce one image placeholder per attached image and reject identity or answer leakage. A question-only probe generates search queries and retrieves web evidence with all images withheld. A judge compares its output with the reference answer and grounded identities to flag answer recovery, identity recovery, or substantial candidate narrowing. Detected failures guide revisions to the references, evidence, or relation, followed by renewed checks. Wording revisions preserve the supported answer; changes to the information requirement trigger new evidence and answer checks. Let $g_{\mathrm{ref}}, g_{\mathrm{sup}}, g_{\mathrm{dep}}$ denote verdicts on reference grounding, answer support, and evidence and image necessity, and let $b(T)$ indicate a detected shortcut. The retention criterion is

$$\mathcal{T}_{\mathrm{keep}} = \big\{T : g_{\mathrm{ref}}(T) = g_{\mathrm{sup}}(T) = g_{\mathrm{dep}}(T) = 1,\ b(T) = 0\big\}. \tag{10}$$

Here $b(T) = 0$ means that no shortcut was detected within the probe budget; it does not exclude other solution paths.