ICLR 2026 · CONFERENCE PAPER

WebWatcher: Breaking New Frontiers of Vision-Language Deep Research Agent

Xinyu Geng[1,2*], Peng Xia[2,3*], Zhen Zhang[2], **Xinyu Wang**[2✉], Qiuchen Wang[2], Ruixue Ding[2], Chenxi Wang[2], Jialong Wu[2], Kuan Li[2], Yida Zhao[2], Huifeng Yin[2], **Yong Jiang**[2✉], Pengjun Xie[2], Fei Huang[2], Huaxiu Yao[3], Yi R. Fung[1✉], Jingren Zhou[2]

[1]The Hong Kong University of Science and Technology · [2]Tongyi Lab, Alibaba Group · [3]The University of North Carolina at Chapel Hill | github.com/Alibaba-NLP/DeepResearch/tree/main/WebAgent/WebWatcher

SOTA Open-Source VL Agent

Code Available on GitHub

Backbone: Qwen2.5-VL-7B / 32B
Training: SFT Cold Start + GRPO RL

MOTIVATION & PROBLEM

Deep research agents have shown **superhuman cognitive abilities** for complex information-seeking, yet existing work is overwhelmingly **text-centric**, overlooking the visual information pervasive in real-world web environments.

Key Challenges in Multimodal Deep Research:

- Requires stronger **perceptual + logical + knowledge** reasoning
- Demands proficiency with **sophisticated multi-modal tools**
- Current VQA data lacks **planning depth & multi-hop complexity**
- Existing agents rely on **rigid template-driven pipelines**

Chart Analysis · Visual Web UI · Scientific Diagrams · Multimodal Web

VL Agent	Search Agent	WebWatcher ✓
Shallow texture inference; fails on complex VL tasks	Text-only retrieval; poor cross-modal synthesis	Multi-tool + deep VL reasoning + cross-validation

KEY CONTRIBUTIONS

① **WebWatcher** — First multimodal VL deep research agent with joint visual-textual reasoning and 4 integrated tools

② **BrowseComp-VL** — Novel benchmark with 399 VQA pairs requiring complex cross-modal retrieval & synthesis

③ **Trajectory Pipeline** — Automated GPT-4o annotation pipeline generating high-quality tool-use trajectories

④ **SFT + GRPO Training** — Cold-start supervised fine-tuning followed by reinforcement learning

WEBWATCHER ARCHITECTURE

WebWatcher follows a **think-act-observe** (ReAct) loop, generating structured trajectories over multi-step tool interactions.

VL Input (Image + Query) → Thought <think> → Action <tool_call> → Observe <tool_response> → Finish <answer>

4 Integrated Multimodal Tools:

- **Web Image Search** — Google SerpApi · image retrieval with captions & URLs
- **Web Text Search** — Open-domain knowledge & information seeking
- **Visit (Jina AI)** — Navigate URLs · summarize webpage content
- **Code Interpreter** — Symbolic computation & numerical reasoning

BROWSECOMP-VL: TWO DIFFICULTY LEVELS

Level 1 — Explicit Multi-hop
Questions reference explicit entities; answers obtainable via iterative retrieval across multiple sources. Reasoning remains non-trivial due to cross-source integration.
Example: "Which game mode features a scenario absent from Skill Points Daily Scenarios but present in Challenge Road?"

Level 2 — Obfuscated Synthesis
Entities replaced with vague/ambiguous descriptions; dates fuzzed; names masked. Requires planning, comparison, and multi-modal synthesis rather than direct retrieval.
Example: "A significant railway station in northern India serves as a key junction... When did construction commence towards a nearby town?"

TRAINING PIPELINE: 3-STAGE FRAMEWORK

① Data Prep — 110K BrowseComp-VL pairs + 60K long-tail VQA + 40K hard VQA samples → 8K filtered SFT trajectories

② SFT Cold Start — Supervised fine-tuning on GPT-4o annotated tool-use trajectories.

③ GRPO RL — Group-Relative Policy Optimization refines decision-making. Reward: $R = 0.2 \cdot r_f + 0.8 \cdot r_a$. N=16 rollouts.

PERFORMANCE SUMMARY (WEBWATCHER-32B)

- **HLE (Avg)** — Humanity's Last Exam — Best Open-Source Agent — **13.6%** (Bio: 33.8%↑)
- **BrowseComp-VL** — Avg L1+L2 — New SOTA — **27.0%** (L1: 28.4 / L2: 25.0)
- **LiveVQA** — Real-time Visual QA — New SOTA — **58.7%** (+8.7% vs OmniSearch)
- **MMSearch** — Multimodal Search — New SOTA — **55.3%** (+6.9% vs OmniSearch)

HLE BENCHMARK RESULTS (%)

Model	Bio	Chem	CS/AI	Math	Physics	Avg
— Direct Inference —						
GPT-4o	13.8	0.0	3.9	6.8	7.1	6.5
Qwen2.5-VL-72B	3.4	8.0	0.0	8.0	0.0	4.9
— Prompt Workflow —						
Gemini-2.5-flash	25.9	3.2	7.1	8.0	3.5	11.4
Qwen2.5-VL-72B	15.8	10.3	8.1	8.0	6.8	8.6
— Reasoning Models —						
o4-mini	12.1	23.7	17.7	0.0	33.3	16.0
Gemini-2.5-Pro	23.7	17.7	13.3	8.0	14.3	15.8
— Open Source Agents —						
OmniSearch (GPT-4o)	15.5	8.2	0.0	6.8	21.4	9.3
WebWatcher-7B	18.6	6.5	6.7	4.0	7.1	10.6
WebWatcher-32B	**33.8**	**9.7**	0.0	**8.9**	14.3	**13.6**

BROWSECOMP-VL, LIVEVQA & MMSEARCH RESULTS (%)

Model	BC-L1	BC-L2	BC-Avg	LiveVQA	MMSearch
— Direct Inference —					
GPT-4o	6.4	4.0	5.5	29.7	18.7
Qwen2.5-VL-72B	9.2	3.0	7.1	30.3	11.7
— Prompt Workflow —					
o3	26.7	23.0	24.9	50.0	54.3
Claude-3.7-Sonnet	13.9	6.0	11.2	30.3	32.7
— Agents —					
OmniSearch (GPT-4o)	19.7	10.0	16.3	40.9	49.7
WebWatcher-7B	23.6	17.0	21.2	51.2	49.1
WebWatcher-32B	**28.4**	**25.0**	**27.0**	**58.7**	**55.3**

HUMAN VS. AGENT (BROWSECOMP-VL)

Setting	Acc%	I_u	T_s(m)	T_u(m)
L1 Human	33.2	42	35	59
L2 Human	18.0	144	109	116
L1 WebWatcher	**28.4**	1	**0.3**	2.5
L2 WebWatcher	**25.0**	3	**0.8**	2.5

WebWatcher solves tasks in **~0.3–0.8 min** vs humans at 35–109 min. Near-human accuracy at 100× speed.

TOOL USAGE ACROSS BENCHMARKS

Web Text Search dominates (48.4% overall):

- Text Search — 62% (BC-VL)
- Image Search — 39% (SimpleVQA)
- Visit — 27% (BC-VL)
- Code Interp — 8.5% (overall)

Tool distribution mirrors benchmark demands — WebWatcher is **context-aware and cost-conscious**, not reliant on any single tool.

ANALYSIS & ABLATION

Ablation: Optimal Tool Call Count (HLE)

- =1 call — 8.79%
- =2 calls — 10.61%
- ⩾3 calls — 12.12%
- =4 calls — 10.00%

Cold Start vs. Instruct Initialization:

✗ Instruct Init	✓ SFT Cold Start
Stalls near zero; strict grader suppresses partial answers; frequent tool-call format errors	Lifts initial scores; GRPO trends upward; 0.06–0.18 margin advantage across all benchmarks