

# OmniGAIA: Towards Native Omni-Modal AI Agents

Xiaoxi Li<sup>1\*</sup>, Wenxiang Jiao<sup>2</sup>, Jiarui Jin<sup>2</sup>, Shijian Wang<sup>3</sup>, Guanting Dong<sup>1</sup>, Jiajie Jin<sup>1</sup>, Hao Wang<sup>4</sup>,  
Yinuo Wang<sup>5</sup>, Ji-Rong Wen<sup>1</sup>, Yuan Lu<sup>2</sup>, Zhicheng Dou<sup>1†</sup>

<sup>1</sup>Renmin University of China, <sup>2</sup>Xiaohongshu Inc., <sup>3</sup>Southeast University, <sup>4</sup>Zhejiang University, <sup>5</sup>Tsinghua University

\*Work done during internship at Xiaohongshu, <sup>†</sup>Corresponding author

Human intelligence naturally intertwines omni-modal perception—spanning vision, audio, and language—with complex reasoning and tool usage to interact with the world. However, current multi-modal LLMs are primarily confined to bi-modal interactions (e.g., vision-language), lacking the unified cognitive capabilities required for general AI assistants. To bridge this gap, we introduce **OmniGAIA**, a comprehensive benchmark designed to evaluate omni-modal agents on tasks necessitating deep reasoning and multi-turn tool execution across video, audio, and image modalities. Constructed via a novel omni-modal event graph approach, OmniGAIA synthesizes complex, multi-hop queries derived from real-world data that require cross-modal reasoning and external tool integration. Furthermore, we propose **OmniAtlas**, a native omni-modal foundation agent under tool-integrated reasoning paradigm with active omni-modal perception. Trained on trajectories synthesized via a hindsight-guided tree exploration strategy and *OmniDPO* for fine-grained error correction, OmniAtlas effectively enhances the tool-use capabilities of existing open-source models. This work marks a step towards next-generation native omni-modal AI assistants for real-world scenarios.

**Contact:** [xiaoxi\\_li@ruc.edu.cn](mailto:xiaoxi_li@ruc.edu.cn), [dou@ruc.edu.cn](mailto:dou@ruc.edu.cn)

**Code & Demo:** <https://github.com/RUC-NLPIR/OmniGAIA>

**Dataset & Model:** <https://huggingface.co/collections/RUC-NLPIR/omnigaia>

**Leaderboard:** <https://huggingface.co/spaces/RUC-NLPIR/OmniGAIA-LeaderBoard>

## 1 Introduction

Human intelligence seamlessly intertwines language, vision, and audio with long-horizon reasoning and tool use to understand the world and take actions. Building general-purpose AI assistants therefore requires models that can *jointly* perceive across modalities, reason over long contexts, and interact with external tools for verification and knowledge acquisition. Yet, despite rapid progress, multimodal LLM research is still dominated by bi-modal settings (e.g., vision–language or audio–language), which limits their ability to handle truly interwoven real-world modalities.

Emerging omni-modal foundation models (e.g., Qwen3-Omni ([Xu et al., 2025b](#))) have begun to unify richer modalities, but most efforts primarily emphasize perception, leaving *tool-integrated, agentic reasoning* underexplored. Evaluation also lags behind: existing benchmarks are largely bi-modal and perception-centric (e.g., OmniBench ([Li et al., 2024](#)), WorldSense ([Hong et al., 2025](#)), UNO-Bench ([Chen et al., 2025a](#))), and thus do not adequately measure multi-hop omni-modal reasoning and multi-turn external tool use with verifiable open-form answers.

To bridge this gap, we introduce **OmniGAIA**, a challenging benchmark for native omni-modal agents. OmniGAIA comprises 360 tasks across 9 real-world domains, covering both video-with-audio and image+audio settings, and explicitly requires multi-turn tool use (e.g., web search/browsing and code) to produce verifiable *open-form* answers. To structure time-aligned multimodal cues and tool-related evidence for multi-hop reasoning, OmniGAIA is constructed via an *omni-modal event-graph-driven* pipeline ([Figure 2](#)): (1) we collect data and mine fine-grained signals from raw media; (2) we build an initial event graph that connects cross-modal entities/events and relations; (3) we expand the graph with *next-hop* evidence via cross-modal retrieval and external tools; and (4) we fuzzify key nodes/edges to generate multi-hop QA, followed by LLM screening and human verification for solvability and uniqueness.

Beyond benchmarking, we propose **OmniAtlas**, a native omni-modal foundation agent following the *Tool-Integrated Reasoning* (TIR) paradigm that naturally interleaves reasoning and tool calls. OmniAtlas further supports *active omni-modal perception* to selectively “look” or “listen” to the segments/regions in long media without blanket down-sampling. For training, we synthesize high-quality tool-integrated trajectories via *hindsight-guided tree exploration*,

### Example 1: Image + Audio Task

Calculate the number of months (rounded to the nearest whole number) between the large anti-cuts protest in London pictured and the official launch of the debt-relief campaign discussed in the audio.

 "... That's where the Rolling Jubilee comes in. It raises money to buy the debt ..."



#### Annotated Solution:

1. Analyze the image and identify it as an anti-cuts protest in London.
2. Search to confirm the protest date was *March 26, 2011*.
3. Analyze the audio and confirm it describes the "Rolling Jubilee" debt-relief campaign.
4. Search to confirm the campaign officially launched on *November 15, 2012*.
5. Calculate the date difference: 600 days. Convert to months:  $600 \div 30.436875 \approx 19.71$  months. Round to the nearest whole number: **20 months**.

Sources: ...Web URLs..., Level: Medium, Required Tools: web\_search

### Example 2: Video w/ Audio Task

During a visit to the Joliet Iron Works Historic Site as shown in the video, the speaker spots a movable bridge in the distance and remarks that it reminds him of a bridge featured in the movie *The Blues Brothers*. **What is the name of this bridge, and how many years had it been standing** when filming for *The Blues Brothers* began?

 "... See that bridge over there? That is the Ruby Street Bridge ..."



#### Annotated Solution:

1. Watch the video to confirm the location and note the mentioned movable bridge.
2. Watch the video / Search to confirm the nearby bridge as the **Ruby Street Bridge**.
3. Search to confirm the bridge's construction year: *1935*.
4. Search to confirm the filming start date for *The Blues Brothers*: *July 1979*.
5. Calculate the bridge's age at filming start:  $1979 - 1935 = 44$  years.

Sources: ...Web URLs..., Level: Medium, Required Tools: web\_search

**Figure 1 Short examples from OmniGAIA.** Two illustrative (image + audio and video w/ audio) instances showing omni-modal evidence integration and multi-step tool use (e.g., web search) to derive a verifiable final answer.

**Table 1 Comparison of OmniGAIA with existing benchmarks.** “Video”, “Image”, and “Audio” denote the supported modalities. “MC” indicates multiple-choice questions, whereas “Open” indicates open-form generation.

Benchmark	Video	Image	Audio	Multi-hop Reasoning	External Tools	Multi-Domain	Video Duration	Audio Duration	Answer Type	Qwen3-Omni Accuracy
GAIA (Mialon et al., 2024)	✗	✓	✗	✓	✓	✓	-	-	Open	-
AV-Odyssey (Gong et al., 2024)	✗	✓	✓	✗	✗	✓	-	3-364 s	MC	-
OmniBench (Li et al., 2024)	✗	✓	✓	✗	✗	✓	-	0.6-31 s	MC	58.4
Daily-Omni (Zhou et al., 2025)	✓	✗	✓	✗	✗	✗	30/60 s	30/60 s	MC	75.8
WorldSense (Hong et al., 2025)	✓	✗	✓	✗	✗	✓	15-656 s	15-656 s	MC	54.0
OmniVideoBench (Li et al., 2025b)	✓	✗	✓	✓	✗	✓	4-1955 s	4-1955 s	MC	38.4
VideoDR (Liu et al., 2026)	✓	✗	✗	✓	✓	✓	10-288 s	-	Open	37.0
UNO-Bench (Chen et al., 2025a)	✓	✓	✓	✓	✗	✓	0.7-641 s	1-641 s	MC/Open	42.1/37.1
<b>OmniGAIA (Ours)</b>	✓	✓	✓	✓	✓	✓	20-2352 s	20-657 s	Open	13.3

perform trajectory-level supervised learning, and further propose **OmniDPO** for fine-grained error correction.

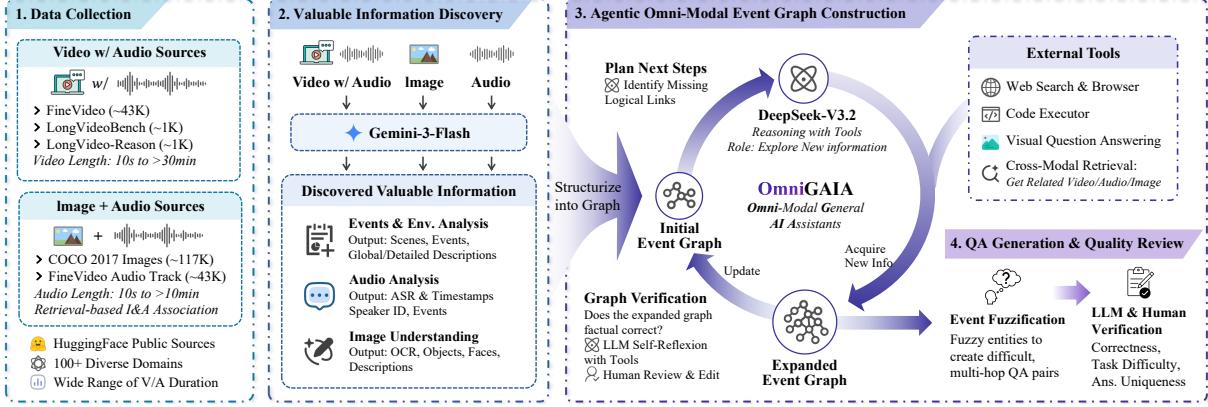
Experiments show that OmniGAIA is highly challenging: the strongest proprietary model (Gemini-3-Pro) reaches 62.5 Pass@1, while an open-source baseline (Qwen3-Omni) achieves 13.3. Our OmniAtlas recipe substantially improves open models (e.g., Qwen3-Omni: 13.3→20.8). Further analyses of fine-grained error types, tool-use behaviors, and perception strategies expose key limitations of current methods and point to promising directions for future omni-modal agents. Our main contributions are:

- We introduce **OmniGAIA, a challenging benchmark for native omni-modal agents**, featuring video/image/audio inputs, multi-domain coverage, multi-hop reasoning, multi-turn tool use, and open-form answers.
- We propose a scalable **Event-graph-driven Construction Pipeline** that systematically synthesizes hard yet solvable tasks from real-world data.
- We present **OmniAtlas, a native omni-modal foundation agent** with active perception and tool-integrated reasoning, together with a practical training recipe (trajectory synthesis, supervised learning, and **OmniDPO**) that significantly improves open-source backbones.
- We provide comprehensive evaluations and analyses, including category-wise results, fine-grained error breakdowns, and tool-use behavior studies that highlight key bottlenecks for omni-modal agents.

## 2 Related Work

### 2.1 Omni-Modal Foundation Models and Benchmarks

Building on advances in pure-text (Dubey et al., 2024), vision-language (Hurst et al., 2024), and audio-language (Chu et al., 2024) foundation models, recent omni-modal models seek to unify text, vision, and audio within a single LLM backbone. A common approach adopts a unified tokenization-and-projection interface that maps heterogeneous visual and acoustic inputs into a shared token space (Xu et al., 2025b; Liu et al., 2025a; Luo et al., 2025b; Ye et al., 2025).



**Figure 2 OmniGAIA construction pipeline.** From video w/ audio and image + audio data, we mine key signals, build and expand a tool-augmented event graph, and generate LLM and human-verified multi-hop QA via event fuzzification.

Concurrent work further strengthens omni-modal reasoning behaviors (Zhong et al., 2025; Long et al., 2025). For evaluation, existing benchmarks (e.g., OmniBench (Li et al., 2024), WorldSense (Hong et al., 2025), Daily-Omni (Zhou et al., 2025), UNO-Bench (Chen et al., 2025a)) largely emphasize short audios/videos and perception-centric tasks, leaving long-horizon reasoning and tool-integrated agency underexplored.

## 2.2 Autonomous Agents

LLM-driven autonomous agents tackle real-world tasks by reasoning and acting through external tools that interface with their environment (Wang et al., 2024b; Luo et al., 2025a). Existing approaches broadly fall into workflow-based paradigms (Yao et al., 2022; Wang et al., 2023, 2024d; Li et al., 2025h) and native agentic reasoning methods (Li et al., 2025g; Qian et al., 2025; Feng et al., 2025; Jiang et al., 2025), and have shown strong performance on text-only tasks. Moving beyond text, recent studies investigate vision-language agents for multimodal web search (Li et al., 2025c; Wu et al., 2025b; Geng et al., 2025), long-form video understanding (Wang et al., 2024c; Zhang et al., 2025b; Yin et al., 2025), and GUI navigation (Xie et al., 2024; Zhang et al., 2025a; Wang et al., 2024a). However, *omni-modal* foundation agents that natively fuse audio, vision, and language while performing long-horizon agentic reasoning remain underexplored. Such capabilities are essential for building general-purpose AI assistants in real-world scenarios.

## 3 OmniGAIA: Benchmarking Omni-Modal General AI Assistants

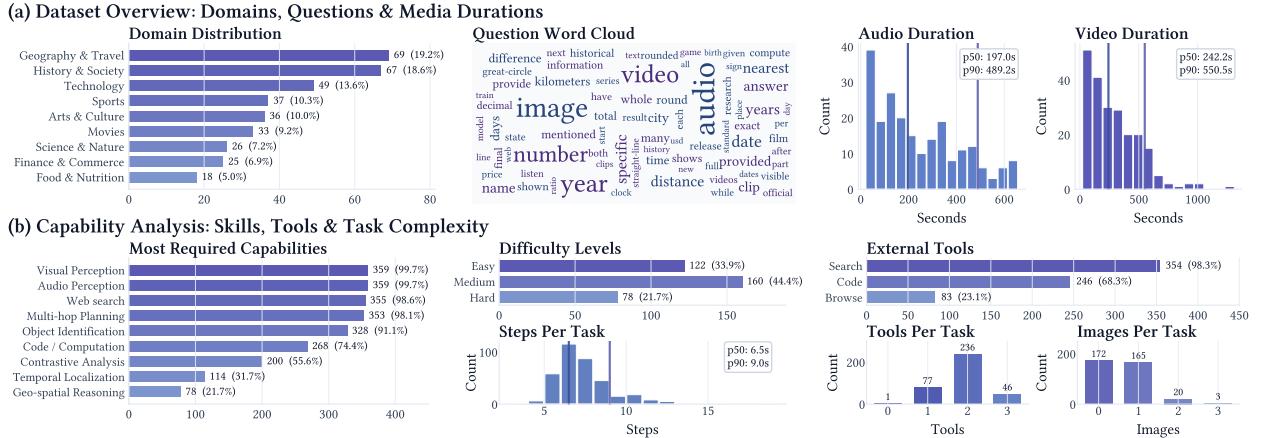
OmniGAIA is a benchmark of challenging *omni-modal agentic* tasks designed to stress-test unified perception over vision, audio, and language, together with long-horizon reasoning and multi-turn tool use in realistic scenarios.

### 3.1 Data Collection

To reflect the complexity of real-world omni-modal interactions, we construct OmniGAIA from two complementary settings: (i) **video with audio**, and (ii) **image + audio** pairs.

For the video setting, we aggregate high-quality videos from multiple sources to ensure diversity in both content and duration. We include **FineVideo** (Farré et al., 2024) (43K videos spanning broad domains; average length 4 minutes). To evaluate long-context reasoning, we further incorporate **LongVideoBench** (Wu et al., 2024) (~1K videos) and **LongVideo-Reason** (Chen et al., 2025c) (~1K videos), both containing videos around 10 minutes.

For the image + audio setting, we use audio tracks from FineVideo to provide diverse acoustic environments, and draw images from **COCO 2017** (Lin et al., 2014), which contains 122K complex everyday-scene images with object detection and segmentation annotations.



**Figure 3 OmniGAIA statistics.** This figure presents a detailed breakdown of domain distributions, required capabilities, and task attributes, underscoring the complex demands placed on omni-modal perception, reasoning, and tool utilization.

### 3.2 Discovering Valuable Information

We employ a strong omni-modal model (Gemini-3-Flash) to extract fine-grained, time-aware signals from each modality for task construction. For videos, we split each video into clips of at most 60 seconds to capture subtle temporal details, and generate both clip-level and full-video descriptions covering scenes, events, and non-speech ambient sounds. For audio, we run timestamped automatic speech recognition (ASR), speaker diarization, and audio event detection; we also tag non-speech acoustic environments (e.g., street, indoor, stadium, nature) and produce global audio summaries. For images, we apply optical character recognition (OCR), recognize objects and faces, and generate a holistic caption to summarize visual content.

### 3.3 Omni-modal Event Graph Construction

To reliably synthesize complex multi-hop tasks, we build an *omni-modal event graph* that structures the discovered information into an explicit graph for each sample. This graph serves as the backbone of our event-graph-driven construction pipeline, enabling systematic evidence expansion and controllable information fuzzification for QA generation.

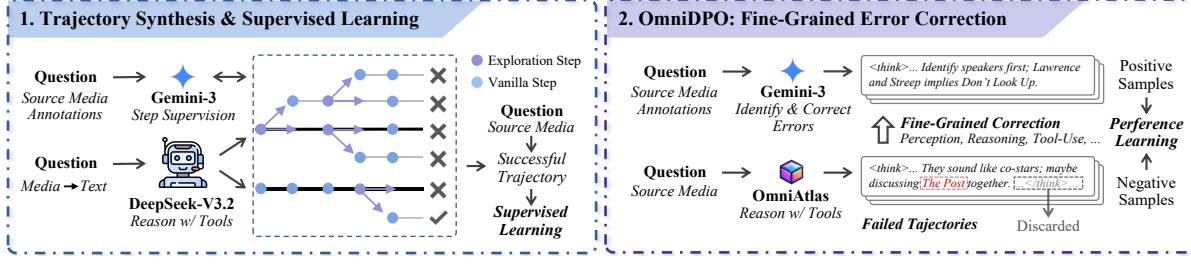
Using the extracted information, we leverage a strong reasoning agent DeepSeek-V3.2 to automatically build an event graph that represents entities/events and their cross-modal relations. Importantly, real-world logic is rarely a simple linear chain; it often exhibits branching (one-to-many), cascading (sequential), and mixed topologies. The graph representation captures such structures and supports reliable synthesis of logically consistent, challenging tasks.

### 3.4 Agentic Omni-modal Event Graph Expansion

Given an initial event graph, we introduce **Agentic Event Graph Expansion** to proactively discover missing evidence and create tasks that truly require cross-modal association and external tool use. Following the Tool-Integrated Reasoning (TIR) paradigm, we use a strong reasoning model (DeepSeek-V3.2) as an exploration agent that searches for *next-hop valuable information* and links it back to the graph.

**Functionality for the event exploration agent.** We equip the agent with a set of omni-modal and external tools:

- **Cross-modal sources linking:** The agent can call `search_related_{video/audio/image}_info` to retrieve context-related multi-modal materials from our database. This is crucial when current graph is insufficient for a tightly-coupled multi-hop question. For the image + audio setting, we pre-retrieve the initial related audios candidates to encourage explicit cross-modal reasoning.
- **Web knowledge integration:** With `web_search` and `page_browser`, the agent can retrieve top web pages and read detailed content, enabling time-sensitive, verifiable external knowledge beyond the original media.
- **External visual exploration:** Using `web_image_search` and `visual_question_answering`, the agent can search web images and query their content, expanding task construction to scenarios requiring external visual evidence.



**Figure 4** **OmniAtlas training strategy.** Left, we synthesize tool-integrated trajectories via step-level supervision and guided tree exploration, selecting successful runs for supervised fine-tuning; right, **OmniDPO** locates the first error in a failed trajectory and generates a corrected prefix, forming positive/negative preference pairs for fine-grained correction.

- **Computation:** The code\_executor tool supports complex computations (e.g., arithmetic, statistics), enabling tasks that require reliable multi-step numerical reasoning.

During task generation, these tools are embedded in the prompting interface, and the agent autonomously decides whether and how to invoke them to expand the information boundary of the current graph, producing complex QA pairs enriched with next-hop evidence.

### 3.5 QA Pairs Generation via Event Fuzzification

To convert expanded graphs into truly challenging tasks, we propose **QA generation via event fuzzification**. Directly querying a graph node often reduces to trivial fact lookup. Instead, we select specific nodes/edges along long reasoning paths and apply *fuzzy entities* (e.g., replacing a specific entity with its type, or masking key attributes) to mask or abstract key information. This forces models to traverse the full logical path and integrate multi-source, multi-modal evidence to derive a unique answer.

### 3.6 Quality Inspection

To ensure rigor and high quality, we apply an inspection pipeline with *LLM screening* and *human verification*, with an optional difficulty expansion step in between.

**1. LLM screening:** We form a review committee with DeepSeek-V3.2 and Gemini-3-Pro to automatically evaluate each QA pair across multiple criteria: (i) naturalness and clarity of the question; (ii) indispensability of omni-modal perception and tool use (filtering out unimodal or trivial cases); and (iii) answer correctness and uniqueness.

**2. Difficulty expansion:** For preliminarily qualified samples, we optionally increase difficulty by linking additional data sources, mining deeper evidence, or introducing more complex computation steps.

**3. Human review:** Finally, we invite three graduate-level computer science reviewers to verify each QA pair against the underlying media. They check question soundness, annotation correctness, and answer correctness/uniqueness, and fix minor issues to ensure each test case is reliably solvable and high-quality.

### 3.7 Statistics

As shown in Figure 3, OmniGAIA comprises 360 omni-modal agentic tasks across 9 real-world domains, intentionally designed to stress *long-horizon* perception and *tool-integrated* reasoning. Tasks often require grounding evidence from both vision and audio over minutes-long media, planning multi-step solution paths, and verifying or extending information via external tools (primarily web search, and occasionally code/computation). The statistics highlight that performance hinges not only on native perception, but also on reliable multi-hop planning and effective tool use under long contexts.

## 4 OmniAtlas: Omni-Modal Foundation Agent

In this section, we introduce **OmniAtlas**, a native omni-modal foundation agent that unifies vision, audio, and language perception with long-horizon reasoning and autonomous tool use. To overcome the key weaknesses of current open-source omni-modal models in perception and tool-integrated reasoning, we present a comprehensive training and optimization recipe.

### 4.1 Autonomous Tool-Integrated Reasoning

To enable OmniAtlas to acquire external knowledge and handle complex tasks, we integrate tools like *web search*, *page browser*, and *code executor*. The agent adopts a *tool-integrated reasoning* paradigm, autonomously switching between internal reasoning and tool usage as needed.

Formally, we define an agent trajectory as  $\tau = [(s_t, a_t, o_t)]_{t=0}^T$ , where  $s_t$  denotes the reasoning thought at step  $t$ ,  $a_t$  the action (either a tool call or a final response), and  $o_t$  the observation returned by the tool (empty if no tool is invoked). The model generates the next thought and action conditioned on the interaction history:

$$p_\theta(\tau \mid \mathbf{x}) = \prod_{t=0}^T p_\theta(s_t, a_t \mid \mathbf{x}, s_{<t}, a_{<t}, o_{<t}) \quad (1)$$

Here,  $\mathbf{x}$  denotes the user instruction and omni-modal inputs. When tool-call tokens are detected, generation is paused, the corresponding tool is executed, and the returned observation  $o_t$  is appended to the context so the model can continue. This design preserves intermediate reasoning states and supports coherent long-horizon problem solving, aligning with the tool-integrated generation philosophy of DeepSeek-V3.2 (DeepSeek-AI, 2025).

**Active Omni-Modal Perception.** For long videos or high-resolution images, naively ingesting all media is token-expensive and often requires aggressive downsampling that can discard critical details (Li et al., 2025a). To mitigate this, OmniAtlas supports *active* omni-modal perception: the agent can selectively request the specific segments or regions it needs via operations such as `read_video(video_id, t_start, t_end)`, `read_audio(audio_id, t_start, t_end)`, and `read_image(image_ids, crop_box)`. When invoked, the corresponding raw media content is loaded into the model context, enabling “look-where-needed” perception without blanket downsampling.

### 4.2 Trajectory Synthesizing via Guided Tree Exploration

Our preliminary experiments on OmniGAIA show that open-source omni-modal models still lag behind in both omni-modal perception and tool-integrated reasoning. To internalize these capabilities, we synthesize high-quality agent trajectories via a two-stage pipeline: (i) we use Gemini-3-Flash to convert raw multi-modal inputs into detailed textual descriptions; (ii) we then generate tool-augmented solution trajectories using **Hindsight-Guided Tree Exploration**.

Concretely, since proprietary Gemini models do not expose raw reasoning traces, we use strong reasoning agent DeepSeek-V3.2 to synthesize tool-integrated trajectories. Starting from the root state, we sample  $k = 3$  candidate continuations (reasoning + tool actions) at each step and use a verifier (Gemini-3-Flash), conditioned on the ground-truth answer, to prune incorrect or redundant branches; we keep only successful trajectories for training (Figure 4).

### 4.3 Trajectory-Level Supervised Fine-Tuning

We perform *trajectory-level* supervised fine-tuning (SFT) to teach the model effective perception, reasoning, and tool-use behaviors. We use standard teacher forcing, but apply *masked* supervision: we compute loss only on tokens generated by the agent (reasoning and tool-call tokens), while masking out tool observations to prevent memorizing environment feedback.

Let the input sequence be  $\mathbf{y} = [y_1, y_2, \dots, y_L]$  with a mask  $\mathbf{m} \in \{0, 1\}^L$ , where  $m_i = 1$  iff  $y_i$  belongs to the agent’s thoughts or actions. The masked SFT objective is:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{\sum_{i=1}^L m_i} \sum_{i=1}^L m_i \log p_\theta(y_i \mid y_{<i}, \mathbf{x}) \quad (2)$$

This encourages the model to learn *how to think and act* without fitting the noisy tool observation tokens.

**Table 2 Main results on the OmniGAIA benchmark.** The Pass@1 metric is reported for all tasks. Best and second-best scores are highlighted in **bold** and underlined respectively, shown separately for proprietary and open-source models.

Method	# Params	Category-Wise Breakdown										Difficulty Levels			Overall
		Geo.	Tech.	Hist.	Fin.	Sport	Art	Movie	Sci.	Food	Easy	Med.	Hard		
<i>Proprietary Omni-Modal Models</i>															
Gemini-2.5-Flash-Lite	-	5.8	8.2	14.9	4.0	10.8	8.3	6.1	3.9	11.1	9.8	8.1	7.7	8.6	
Gemini-2.5-Pro	-	23.2	28.6	32.8	20.0	32.4	41.7	42.4	26.9	33.3	41.8	26.9	21.8	30.8	
Gemini-3-Flash	-	<u>50.7</u>	<u>57.1</u>	<u>44.8</u>	<u>48.0</u>	<u>59.5</u>	<u>55.6</u>	<u>54.6</u>	<u>38.5</u>	<u>61.1</u>	<u>67.2</u>	<u>46.9</u>	<u>37.2</u>	<u>51.7</u>	
Gemini-3-Pro	-	<b>65.2</b>	<b>59.2</b>	<b>62.1</b>	<b>72.0</b>	<b>78.4</b>	<b>52.8</b>	<b>48.5</b>	<b>42.3</b>	<b>88.9</b>	<b>78.7</b>	<b>61.9</b>	<b>38.5</b>	<b>62.5</b>	
<i>Open-Source Omni-Modal Models</i>															
Qwen-2.5-Omni	3B	0.0	2.0	4.5	0.0	0.0	0.0	0.0	3.9	0.0	1.6	1.9	0.0	1.4	
Qwen-2.5-Omni	7B	1.5	4.1	7.5	4.0	0.0	2.8	0.0	7.7	5.6	8.2	1.3	1.3	3.6	
Baichuan-Omni-1.5	8B	2.9	4.1	3.0	4.0	2.7	0.0	3.0	3.8	0.0	4.9	2.5	0.0	2.8	
MiniCPM-O-2.6	8B	2.9	2.0	1.5	0.0	2.7	8.3	3.0	3.8	5.6	3.3	2.5	3.8	3.1	
Ming-Lite-Omni-1.5	20B-A3B	2.9	6.1	1.5	4.0	5.4	2.8	6.1	7.7	5.6	4.9	3.8	2.6	3.9	
Qwen-3-Omni	30B-A3B	<u>8.7</u>	14.3	11.9	<u>28.0</u>	10.8	13.9	<u>9.1</u>	<b>15.4</b>	<u>22.2</u>	19.7	10.6	<b>9.0</b>	<u>13.3</u>	
Ming-Flash-Omni	100B-A6B	5.8	8.2	10.4	12.0	8.1	5.6	6.1	<u>11.5</u>	11.1	12.3	7.5	3.8	8.3	
LongCat-Flash-Omni	560B-A27B	<u>8.7</u>	10.2	16.4	12.0	10.8	8.3	6.1	<u>11.5</u>	16.7	16.4	9.4	<u>6.4</u>	11.1	
OmniAtlas-Qwen-2.5	3B	4.4	12.2	<u>16.7</u>	4.0	<u>16.2</u>	11.1	3.0	<u>11.5</u>	11.1	13.9	10.0	5.1	10.3	
OmniAtlas-Qwen-2.5	7B	<u>8.7</u>	<u>18.4</u>	16.4	4.0	<u>16.2</u>	<u>22.2</u>	3.0	7.7	<u>22.2</u>	<u>22.1</u>	<u>11.3</u>	3.9	<u>13.3</u>	
OmniAtlas-Qwen-3	30B-A3B	<b>10.1</b>	<b>30.6</b>	<b>29.9</b>	<b>32.0</b>	<b>18.9</b>	<u>16.7</u>	<b>12.1</b>	<u>11.5</u>	<b>27.8</b>	<b>31.1</b>	<b>18.8</b>	<b>9.0</b>	<b>20.8</b>	

#### 4.4 OmniDPO: Fine-Grained Error Correction

Omni-modal agentic tasks require multiple tightly-coupled capabilities (e.g., visual/audio perception, reasoning, and tool use), and full-trajectory SFT alone is often insufficient to correct fine-grained mistakes. We propose **OmniDPO**, which performs preference optimization on *fine-grained segments* aligned with failure modes, including perception, reasoning, tool use or other specific types of errors.

Specifically, we let the SFT model explore on the training set. For each failed trajectory, Gemini-3-Flash (with access to the annotated solution and answer) identifies the *first* erroneous step and generates a corrected prefix up to that point. This approach enables the training process to concentrate on rectifying a single error per optimization. We denote the original (incorrect) prefix as  $\tau_{\text{lose}}$  and the corrected prefix as  $\tau_{\text{win}}$ , and optimize a masked DPO objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) = -\mathbb{E}_{(\tau_{\text{win}}, \tau_{\text{lose}}) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(\tau_{\text{win}})}{\pi_{\text{ref}}(\tau_{\text{win}})} - \beta \log \frac{\pi_\theta(\tau_{\text{lose}})}{\pi_{\text{ref}}(\tau_{\text{lose}})} \right) \right] \quad (3)$$

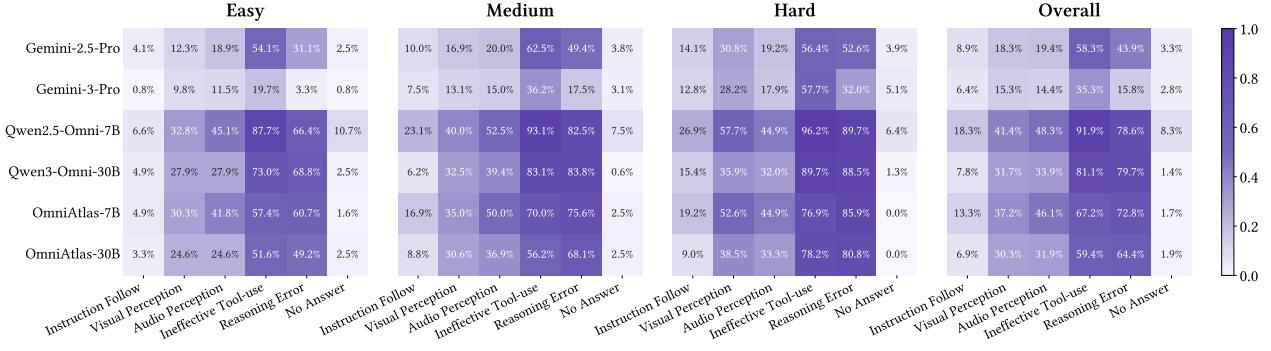
Here  $\pi_{\text{ref}}$  is a reference policy (typically the SFT model). As in Section 4.3, we compute log-probabilities only on agent-generated tokens, focusing correction on the specific module where the error appears.

## 5 Experiments

### 5.1 Experimental Settings

**Evaluation** We employ LLM-as-a-Judge based on DeepSeek-V3.2 (DeepSeek-AI, 2025) to evaluate answer equivalence, considering that answers may appear in diverse forms. Pass@1 is reported, where a trial is considered correct if the model’s final answer is judged equivalent to the ground truth. The judging prompt is detailed in Appendix B. All models are provided with the same external tools, including web search, browser, and code executor.

**Models** We evaluate omni-modal foundation models: proprietary models Gemini-2.5-[Flash-Lite, Pro] (Team, 2025a) and Gemini-3-[Flash, Pro] (Google, 2025); and open-source models Qwen2.5-Omni-[3B,7B] (Xu et al., 2025a), Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025b), Baichuan-Omni-1.5 (Xu et al., 2025a), MiniCPM-O-2.6 (Yao et al., 2024), Ming-Lite-Omni-1.5 (AI et al., 2025), Ming-Flash-Omni (AI and Group, 2025), and LongCat-Flash-Omni (Team, 2025b).



**Figure 5 Fine-grained error analysis.** These heatmaps illustrate the frequency of specific error types—including failures in instruction following, visual/audio perception, tool usage, reasoning, and absence of an answer—across six different models.

## 5.2 Main Results

Table 2 summarizes the Pass@1 performance on OmniGAIA under the unified tool setting. The benchmark proves highly challenging: while the state-of-the-art proprietary model, Gemini-3-Pro, achieves 62.5, the strongest open-source baseline, Qwen-3-Omni, reaches only 13.3.

**(1) Substantial proprietary-open gap:** A stark performance disparity exists between Gemini-3-Pro and Qwen-3-Omni ( $\sim 4.7 \times$ , 62.5 vs. 13.3). This underscores the critical need for advancements in both native omni-modal perception and robust tool-integrated reasoning within the open-source community.

**(2) Scaling parameters alone is insufficient:** Merely increasing model size yields diminishing returns. For instance, the massive LongCat-Flash-Omni (560B) underperforms the smaller Qwen-3-Omni (30B) (11.1 vs. 13.3). This suggests that agentic capabilities—specifically tool-use policies—rather than raw parameter count, are the primary bottleneck.

**(3) OmniAtlas delivers consistent improvements:** Our approach significantly boosts Qwen-3-Omni from 13.3 to 20.8 (+7.5 absolute). Notably, the gains are even more pronounced on smaller backbones (e.g., Qwen-2.5-Omni-7B improves  $\sim 3.7 \times$  from 3.6 to 13.3), demonstrating the efficacy of OmniAtlas in unlocking agentic potential across varying model sizes.

**(4) Hard tasks remain the main challenge:** Performance degrades sharply as task difficulty increases (e.g., Gemini-3-Pro drops from 78.7 on Easy to 38.5 on Hard). While OmniAtlas improves performance on Easy and Medium tasks, the “Hard” subset—requiring deep multi-hop reasoning—remains a formidable challenge, highlighting significant opportunities for future research.

## 5.3 Fine-Grained Error Analysis

Figure 5 breaks down fine-grained error types by difficulty.

**(1) Tool-use and reasoning failures predominate:** Ineffective tool usage and reasoning errors represent the most prevalent failure modes (35.3%–91.9% and 15.8%–79.7%, respectively), significantly outpacing instruction-following issues (6.4%–18.3%) and “No Answer” cases (1.4%–8.3%).

**(2) Hard tasks reveal cascading failure modes:** On hard tasks, open-source models exhibit near-saturated tool misuse ( $\sim 90\%-96\%$ ) alongside high reasoning error rates ( $\sim 80\%-90\%$ ). This suggests that initial failures in evidence acquisition via tools propagate downstream, inevitably leading to reasoning collapse.

**(3) Proprietary models demonstrate superior robustness:** Gemini-3-Pro significantly outperforms Qwen-3-Omni, exhibiting much lower error rates in visual/audio perception (15.3%/14.4% vs. 31.7%/33.9%) and particularly in tool-use/reasoning (35.3%/15.8% vs. 81.1%/79.7%), reflecting its more mature planning and verification capabilities.

**(4) OmniAtlas enhances tool policy, yet perception remains a bottleneck:** While OmniAtlas effectively reduces tool misuse (e.g., 81.1%→59.4%) and reasoning errors (79.7%→64.4%), visual and audio perception errors remain high ( $\sim 30\%-50\%$ ). This indicates that the fundamental perception capability of omni-modal foundation models is a persistent bottleneck requiring further attention. Representative success/failure trajectories are analyzed in Appendix D.

## 5.4 Tool Call Distribution Analysis

Figure 6 illustrates the distribution of tool calls per task run, highlighting successful runs in color.

**(1) External tools are indispensable:** Models exhibiting minimal tool usage (e.g., Qwen-3-Omni-30B, concentrated near 0 calls) achieve negligible success rates. This confirms that native perception alone is insufficient for many OmniGAIA tasks, necessitating external evidence gathering.

**(2) More tool calls do not guarantee better performance:** A high volume of tool calls (long tails reaching  $> 10\text{--}20$ ) does not guarantee success. A substantial fraction of such runs still fail, indicative of inefficient exploration or “thrashing” behaviors where models repeatedly invoke tools without resolving underlying uncertainties.

**(3) OmniAtlas shifts from under-calling to more active tool use:**

In contrast to the passive Qwen-3-Omni-30B, OmniAtlas-30B exhibits a much higher and broader tool-call distribution, aligning with its improvements in ineffective tool-use and overall Pass@1, while leaving new opportunities for more efficient and effective tool-use policies.

## 5.5 Native Perception vs. Tool-based Perception

Do we really need native omni-modal agents, or can perception tools substitute for them? Table 3 offers a controlled ablation under matched model families.

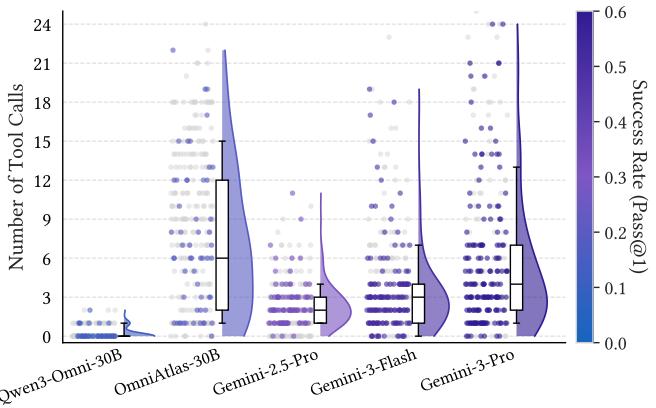
**(1) Native perception is optimal for strong agents:**

For Gemini-3-Flash, native perception achieves the best Avg. score (51.7) with fewer tool calls (4.4). Replacing native channels with perception tools lowers Avg. to 50.0/43.3/46.4 while increasing calls to 7.6/6.8/9.4, yielding no accuracy-cost benefit.

**(2) Perception tools help weak agents on Easy and Medium but not Hard:** For Qwen-3-Omni, tools improve Easy/Med. performance (19.7→24.6; 10.6→15.0/11.9) but consistently reduce Hard performance (9.0→3.9/5.1/7.7). This suggests tool outputs can patch missing low-level signals, but cannot replace native cross-modal integration for long-horizon reasoning.

**(3) Tool perception consistently increases interaction cost:**

Adding perception tools increases the call budget across settings (Qwen-3-Omni: 0.2→0.5–2.0; Gemini-3-Flash: 4.4→6.8–9.4), implying higher latency and deployment cost.



**Figure 6 Tool call distribution analysis.** Dots indicate individual task runs (grey: failed; colored: successful). The box and half-violin plots visualize the tool call frequency for successful runs.

**Table 3 Performance analysis with tool-based perception.** All Qwen-3 models use the 30B-A3B version for fair comparison. Best results within the Gemini-3 and Qwen-3 groups are in **bold**.

Method	Perception Model	Difficulty Levels			Avg.	Tool Calls
		Easy	Med.	Hard		
<i>Native Omni-Modal Perception (Input All Media)</i>						
Gemini-3-Flash	No Need	<b>67.2</b>	<b>46.9</b>	<b>37.2</b>	<b>51.7</b>	4.4
Qwen-3-Omni	No Need	19.7	10.6	<b>9.0</b>	13.3	<b>0.2</b>
<i>Audio Perception Model as a Tool (Input Only Vision)</i>						
Gemini-3-Flash	Gemini-3-Flash	60.7	48.8	35.9	50.0	7.6
Qwen-3-Omni	Qwen-3-Omni	24.6	15.0	3.9	15.8	0.8
Qwen-3-VL	Qwen-3-Omni	24.6	<b>18.1</b>	7.7	<b>18.1</b>	2.8
<i>Visual Perception Model as a Tool (Input Only Audio)</i>						
Gemini-3-Flash	Gemini-3-Flash	50.0	43.1	33.3	43.3	6.8
Qwen-3-Omni	Qwen-3-Omni	18.0	11.3	5.1	12.2	0.5
<i>Audio and Visual Perception Models as Tools (Input No Media)</i>						
Gemini-3-Flash	Gemini-3-Flash	52.5	<b>46.9</b>	<b>35.9</b>	46.4	9.4
Qwen-3-Omni	Qwen-3-Omni	23.8	11.9	7.7	15.0	2.0
Qwen-3	Qwen-3-Omni	<b>32.8</b>	10.6	6.4	17.2	2.3

Therefore, native perception should be the default for capable omni-modal agents to achieve higher performance ceilings, while tool-based perception is best treated as a fallback for weaker agents or missing-modality scenarios.

## 5.6 Training Effectiveness of OmniAtlas

**Table 4** quantifies how OmniAtlas-SFT and OmniDPO affect error rates and performance.

### (1) OmniAtlas-SFT contributes most of the gains:

It drives the majority of improvements by boosting Pass@1 and reducing the ineffective tool-use rate (Qwen-3-Omni-30B: 13.3→18.9, 81.1%→65.3%).

### (2) OmniDPO further delivers across-the-board gains:

**It provides additional improvements (to 13.3→20.8)** and continues to lower perception, tool-use, and reasoning errors, which verifies the effectiveness of the fine-grained error correction.

**Table 4 Training effectiveness of OmniAtlas.** We report four primary error types (↓) and the overall performance (↑). For each model group, the best result in each column is highlighted in **bold**.

Method	Visual Percept.	Audio Percept.	Ineffect. Tool-Use	Reason. Error	Perform.
Qwen-2.5-Omni-7B	41.4	48.3	91.9	78.6	3.6
+ OmniAtlas-SFT	38.9	49.7	69.2	75.0	11.4
+ OmniDPO	<b>37.2</b>	<b>46.1</b>	<b>67.2</b>	<b>72.8</b>	<b>13.3</b>
Qwen-3-Omni-30B	31.7	33.9	81.1	79.7	13.3
+ OmniAtlas-SFT	32.2	35.8	65.3	68.1	18.9
+ OmniDPO	<b>30.3</b>	<b>31.9</b>	<b>59.4</b>	<b>64.4</b>	<b>20.8</b>

## 6 Conclusion and Future Work

We introduce **OmniGAIA**, a benchmark for *native omni-modal agents* that requires multi-hop reasoning and multi-turn tool use over video-with-audio and image+audio inputs. OmniGAIA is built with an event-graph pipeline that aligns and expands cross-modal evidence with tools, then synthesizes verifiable multi-hop questions via controllable event fuzzification and human-vetted screening. We further propose **OmniAtlas**, a native omni-modal foundation agent that follows tool-integrated reasoning with *active* perception, trained via hindsight-guided tree exploration, trajectory-level masked SFT, and *OmniDPO* for fine-grained error correction. Experiments show OmniGAIA remains challenging for current models, and that effective tool-use and long-horizon reasoning—rather than parameter scaling alone—are decisive bottlenecks; our OmniAtlas recipe improves Qwen3-Omni from 13.3 to 20.8 Pass@1 while reducing tool-use and reasoning failures.

Looking ahead, we see three promising directions: **(1) Omni-modal Agentic RL** to directly optimize long-horizon agentic policies under omni-modal feedback; **(2) Omni-modal MCP Services** with scalable tools for broader omni-modal tasks; and **(3) Omni-modal Embodied Agents** benchmarks and foundation models in physical world, advancing LLM-brained AI assistants for real-world task completion.

## 7 Impact Statement

This work advances research on *native omni-modal agents* by introducing OmniGAIA, a benchmark for long-horizon multi-hop reasoning with multi-turn tool use over video-with-audio and image+audio inputs, and by proposing OmniAtlas, a practical recipe for improving such tool-integrated behaviors in open models. These contributions may enable more reliable cross-modal grounding and verification in assistive applications (e.g., education and accessibility) and help standardize evaluation of tool-augmented omni-modal agents. We emphasize that any omni-modal agent should respect data provenance and licensing and prioritize privacy-preserving practices when handling audio/visual inputs.

## References

Inclusion AI and Ant Group. Ming-flash-omni: A sparse, unified architecture for multimodal perception and generation. *CoRR*, abs/2510.24821, 2025. doi: 10.48550/ARXIV.2510.24821. URL <https://doi.org/10.48550/arXiv.2510.24821>.

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, Guangming Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, Kaiyou Song, Kaimeng Ren, Libin Wang, Lixiang Ru, Lele Xie, Longhua Tan, Lyuxin Xue, Lan Wang, Mochen Bai, Ning Gao, Pei Chen, Qingpei Guo, Qinglong Zhang, Qiang Xu, Rui Liu, Ruijie Xiong, Sirui Gao, Tinghao Liu, Taisong Li, Weilong Chai, Xinyu Xiao, Xiaomei Wang, Xiaoxue Chen, Xiao Lu, Xiaoyu Li, Xingning Dong, Xuzheng Yu, Yi Yuan, Yuting Gao, Yunxiao Sun, Yipeng Chen, Yifei Wu, Yongjie Lyu, Zipeng Ma, Zipeng Feng, Zhijiang Fang, Zhihao Qiu, Ziyuan Huang, and Zhengyu He. Ming-omni: A unified multimodal model for perception and generation. *CoRR*, abs/2506.09344, 2025. doi: 10.48550/ARXIV.2506.09344. URL <https://doi.org/10.48550/arXiv.2506.09344>.

- Chen Chen, Zeyang Hu, Fengjiao Chen, Liya Ma, Jiaxing Liu, Xiaoyu Li, Ziwen Wang, Xuezhi Cao, and Xunliang Cai. Uno-bench: A unified benchmark for exploring the compositional law between uni-modal and omni-modal in omni models. *CoRR*, abs/2510.18915, 2025a. doi: 10.48550/ARXIV.2510.18915. URL <https://doi.org/10.48550/arXiv.2510.18915>.
- Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, and Boqing Gong. Omnimxr: Evaluating omni-modality language models on reasoning across modalities. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL <https://openreview.net/forum?id=jki6EFsZLw>.
- Yifei Chen, Guanting Dong, and Zhicheng Dou. Et-agent: Incentivizing effective tool-integrated reasoning agent via behavior calibration, 2026a. URL <https://arxiv.org/abs/2601.06860>.
- Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025c. URL <https://openreview.net/forum?id=wCXAlfvCy6>.
- Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, Biao Wang, Qinglin Lu, and Ruqi Huang. Omnidvideo-r1: Reinforcing audio-visual reasoning with query intention and modality attention, 2026b. URL <https://arxiv.org/abs/2602.05847>.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *CoRR*, abs/2407.10759, 2024. doi: 10.48550/ARXIV.2407.10759. URL <https://doi.org/10.48550/arXiv.2407.10759>.
- DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL <https://arxiv.org/abs/2512.02556>.
- Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, Yuanxing Zhang, Jiaheng Liu, Qiang Liu, Pengfei Wan, and Liang Wang. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models, 2026. URL <https://arxiv.org/abs/2602.04804>.
- Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic entropy-balanced policy optimization, 2025a. URL <https://arxiv.org/abs/2510.14545>.
- Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. Tool-star: Empowering llm-brained multi-tool reasoner via reinforcement learning. *CoRR*, abs/2505.16410, 2025b. doi: 10.48550/ARXIV.2505.16410. URL <https://doi.org/10.48550/arXiv.2505.16410>.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. <https://huggingface.co/datasets/HuggingFaceFV/finevideo>, 2024.
- Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in llms, 2025. URL <https://arxiv.org/abs/2504.11536>.
- 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 frontier of vision-language deep research agent. *CoRR*, abs/2508.05748, 2025. doi: 10.48550/ARXIV.2508.05748. URL <https://doi.org/10.48550/arXiv.2508.05748>.
- Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhouran Yang, and Xiangyu Yue. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *CoRR*, abs/2412.02611, 2024. doi: 10.48550/ARXIV.2412.02611. URL <https://doi.org/10.48550/arXiv.2412.02611>.
- Google. A new era of intelligence with gemini 3, 2025. URL <https://blog.google/products/gemini/gemini-3>.
- Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *CoRR*, abs/2502.04326, 2025. doi: 10.48550/ARXIV.2502.04326. URL <https://doi.org/10.48550/arXiv.2502.04326>.
- Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 32779–32798. Association for Computational Linguistics, 2025. URL <https://aclanthology.org/2025.acl-long.1575/>.

Yuyang Hu, Jiongnan Liu, Jiejun Tan, Yutao Zhu, and Zhicheng Dou. Memory matters more: Event-centric memory as a logic map for agent searching and reasoning, 2026. URL <https://arxiv.org/abs/2601.04726>.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Burette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024. doi: 10.48550/ARXIV.2410.21276. URL <https://doi.org/10.48550/arXiv.2410.21276>.

Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, Tianyu Pang, and Wenhua Chen. Verltool: Towards holistic agentic reinforcement learning with tool use, 2025. URL <https://arxiv.org/abs/2509.01055>.

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Zhao Yang, Hongjin Qian, and Zhicheng Dou. Decoupled planning and execution: A hierarchical reasoning framework for deep search. *CoRR*, abs/2507.02652, 2025a. doi: 10.48550/ARXIV.2507.02652. URL <https://doi.org/10.48550/arXiv.2507.02652>.

Zhuoran Jin, Hongbang Yuan, Kejian Zhu, Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Omni-reward: Towards generalist omni-modal reward modeling with free-form preferences. *CoRR*, abs/2510.23451, 2025b. doi: 10.48550/ARXIV.2510.23451. URL <https://doi.org/10.48550/arXiv.2510.23451>.

Zicheng Kong, Dehua Ma, Zhenbo Xu, Alven Yang, Yiwei Ru, Haoran Wang, Zixuan Zhou, Fuqing Bie, Liuyu Xiang, Huijia Wu, Jian Zhao, and Zhaofeng He. Omni-rrm: Advancing omni reward modeling via automatic rubric-grounded preference synthesis, 2026. URL <https://arxiv.org/abs/2602.00846>.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025a. URL <https://openreview.net/forum?id=zKv8qULV6n>.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, Ying He, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, Zhaohui Wang, Jiayi Tian, Yanghai Wang, Zhe Cao, Minxin Dai, Ke Wang, Runzhe Wen, Yinghao Ma, Yaning Pan, Sungkyun Chang, Termeh Taheri, Haiwen Xia, Christos Plachouras, Emmanouil Benetos, Yizhi Li, Ge Zhang, Jian Yang, Tianhao Peng, Zili Wang, Minghao Liu, Junran Peng, Zhaoxiang Zhang, and Jiaheng Liu. Omnidvideobench: Towards audio-visual understanding evaluation for omni mllms. *CoRR*, abs/2510.10689, 2025b. doi: 10.48550/ARXIV.2510.10689. URL <https://doi.org/10.48550/arXiv.2510.10689>.

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. *CoRR*, abs/2508.13186, 2025c. doi: 10.48550/ARXIV.2508.13186. URL <https://doi.org/10.48550/arXiv.2508.13186>.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *CoRR*, abs/2501.05366, 2025d. doi: 10.48550/ARXIV.2501.05366. URL <https://doi.org/10.48550/arXiv.2501.05366>.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Guanting Dong, Jiajie Jin, Yinuo Wang, Hao Wang, Yutao Zhu, Ji-Rong Wen, Yuan Lu, and Zhicheng Dou. Deepagent: A general reasoning agent with scalable toolsets. *CoRR*, abs/2510.21618, 2025e. doi: 10.48550/ARXIV.2510.21618. URL <https://doi.org/10.48550/arXiv.2510.21618>.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *CoRR*, abs/2504.21776, 2025f. doi: 10.48550/ARXIV.2504.21776. URL <https://doi.org/10.48550/arXiv.2504.21776>.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. Torl: Scaling tool-integrated RL. *CoRR*, abs/2503.23383, 2025g. doi: 10.48550/ARXIV.2503.23383. URL <https://doi.org/10.48550/arXiv.2503.23383>.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnidbench: Towards the future of universal omni-language models. *CoRR*, abs/2409.15272, 2024. doi: 10.48550/ARXIV.2409.15272. URL <https://doi.org/10.48550/arXiv.2409.15272>.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. In-the-flow agentic system optimization for effective planning and tool use. *CoRR*, abs/2510.05592, 2025h. doi: 10.48550/ARXIV.2510.05592. URL <https://doi.org/10.48550/arXiv.2510.05592>.

Huawei Lin, Yunzhi Shi, Tong Geng, Weijie Zhao, Wei Wang, and Ravender Pal Singh. Agent-omni: Test-time multimodal reasoning via model coordination for understanding anything, 2025. URL <https://arxiv.org/abs/2511.02834>.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1\_48. URL [https://doi.org/10.1007/978-3-319-10602-1\\_48](https://doi.org/10.1007/978-3-319-10602-1_48).

Che Liu, Yingji Zhang, Dong Zhang, Weijie Zhang, Chenggong Gong, Haohan Li, Yu Lu, Shilin Zhou, Yue Lu, Ziliang Gan, Ziao Wang, Junwei Liao, Haipang Wu, Ji Liu, André Freitas, Qifan Wang, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. Nexus-o: An omni-perceptive and -interactive model for language, audio, and vision. *CoRR*, abs/2503.01879, 2025a. doi: 10.48550/ARXIV.2503.01879. URL <https://doi.org/10.48550/arXiv.2503.01879>.

Chengwen Liu, Xiaomin Yu, Zhuoyue Chang, Zhe Huang, Shuo Zhang, Heng Lian, Kunyi Wang, Rui Xu, Sen Hu, Jianheng Hou, Hao Peng, Chengwei Qin, Xiaobin Hu, Hong Peng, Ronghao Chen, and Huacan Wang. Watching, reasoning, and searching: A video deep research benchmark on open web for agentic video reasoning, 2026. URL <https://arxiv.org/abs/2601.06943>.

Kai Liu, Jungang Li, Yuchong Sun, Shengqiong Wu, Jianzhang Gao, Daoan Zhang, Wei Zhang, Sheng Jin, Sicheng Yu, Geng Zhan, Jiayi Ji, Fan Zhou, Liang Zheng, Shuicheng Yan, Hao Fei, and Tat-Seng Chua. Javisgpt: A unified multi-modal LLM for sounding-video comprehension and generation. *CoRR*, abs/2512.22905, 2025b. doi: 10.48550/ARXIV.2512.22905. URL <https://doi.org/10.48550/arXiv.2512.22905>.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *CoRR*, abs/2508.09736, 2025. doi: 10.48550/ARXIV.2508.09736. URL <https://doi.org/10.48550/arXiv.2508.09736>.

Xudong Lu, Huankang Guan, Yang Bo, Jinpeng Chen, Xintong Guo, Shuhan Li, Fang Liu, Peiwen Sun, Xueying Li, Wei Zhang, Xue Yang, Rui Liu, and Hongsheng Li. Phostream: Benchmarking real-world streaming for omnimodal assistants in mobile scenarios, 2026. URL <https://arxiv.org/abs/2601.22575>.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiao Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Meng Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, Yiqiao Jin, Fan Zhang, Xian Wu, Hanqing Zhao, Dacheng Tao, Philip S. Yu, and Ming Zhang. Large language model agent: A survey on methodology, applications and challenges. *CoRR*, abs/2503.21460, 2025a. doi: 10.48550/ARXIV.2503.21460. URL <https://doi.org/10.48550/arXiv.2503.21460>.

Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching. *CoRR*, abs/2510.13721, 2025b. doi: 10.48550/ARXIV.2510.13721. URL <https://doi.org/10.48550/arXiv.2510.13721>.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=fibxvahvs3>.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. Toolrl: Reward is all tool learning needs, 2025. URL <https://arxiv.org/abs/2504.13958>.

Jiejun Tan, Zhicheng Dou, Yan Yu, Jiehan Cheng, Qiang Ju, Jian Xie, and Ji-Rong Wen. Hiersearch: A hierarchical enterprise deep search framework integrating local and web searches, 2025. URL <https://arxiv.org/abs/2508.08088>.

Keda Tao, Wenjie Du, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. Omniagent: Audio-guided active perception agent for omnimodal audio-video understanding. *CoRR*, abs/2512.23646, 2025. doi: 10.48550/ARXIV.2512.23646. URL <https://doi.org/10.48550/arXiv.2512.23646>.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025a. doi: 10.48550/ARXIV.2507.06261. URL <https://doi.org/10.48550/arXiv.2507.06261>.

Meituan LongCat Team. Longcat-flash-omni technical report. *CoRR*, abs/2511.00279, 2025b. doi: 10.48550/ARXIV.2511.00279. URL <https://doi.org/10.48550/arXiv.2511.00279>.

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *CoRR*, abs/2401.16158, 2024a. doi: 10.48550/ARXIV.2401.16158. URL <https://doi.org/10.48550/arXiv.2401.16158>.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 2609–2634. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.147. URL <https://doi.org/10.18653/v1/2023.acl-long.147>.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers Comput. Sci.*, 18(6):186345, 2024b. doi: 10.1007/S11704-024-40231-1. URL <https://doi.org/10.1007/s11704-024-40231-1>.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, and Chao Zhang. End-to-end listen, look, speak and act. *CoRR*, abs/2510.16756, 2025. doi: 10.48550/ARXIV.2510.16756. URL <https://doi.org/10.48550/arXiv.2510.16756>.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and GÜl Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXX*, volume 15138 of *Lecture Notes in Computer Science*, pages 58–76. Springer, 2024c. doi: 10.1007/978-3-031-72989-8\4. URL [https://doi.org/10.1007/978-3-031-72989-8\\_4](https://doi.org/10.1007/978-3-031-72989-8_4).

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better LLM agents. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024d. URL <https://openreview.net/forum?id=jJ9BoXAfFa>.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL [http://papers.nips.cc/paper\\_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets\\_and\\_Benchmarks\\_Track.html](http://papers.nips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html).

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webdancer: Towards autonomous information seeking agency. *CoRR*, abs/2505.22648, 2025a. doi: 10.48550/ARXIV.2505.22648. URL <https://doi.org/10.48550/arXiv.2505.22648>.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. *CoRR*, abs/2506.20670, 2025b. doi: 10.48550/ARXIV.2506.20670. URL <https://doi.org/10.48550/arXiv.2506.20670>.

Zijian Wu, Xiangyan Liu, Xinyuan Zhang, Lingjun Chen, Fanqing Meng, Lingxiao Du, Yiran Zhao, Fanshi Zhang, Yaoqi Ye, Jiawei Wang, Zirui Wang, Jinjie Ni, Yufan Yang, Arvin Xu, and Michael Qizhe Shieh. Mcpmark: A benchmark for stress-testing realistic and comprehensive MCP use. *CoRR*, abs/2509.24002, 2025c. doi: 10.48550/ARXIV.2509.24002. URL <https://doi.org/10.48550/arXiv.2509.24002>.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL [http://papers.nips.cc/paper\\_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets\\_and\\_Benchmarks\\_Track.html](http://papers.nips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html).

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215, 2025a. doi: 10.48550/ARXIV.2503.20215. URL <https://doi.org/10.48550/arXiv.2503.20215>.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *CoRR*, abs/2509.17765, 2025b. doi: 10.48550/ARXIV.2509.17765. URL <https://doi.org/10.48550/arXiv.2509.17765>.

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

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A GPT-4V level MLLM on your phone. *CoRR*, abs/2408.01800, 2024. doi: 10.48550/ARXIV.2408.01800. URL <https://doi.org/10.48550/arXiv.2408.01800>.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, Yuming Lou, Dong Yang, Zhijian Liu, Yukang Chen, Ambrish Dantrey, Ehsan Jahangiri, Sreyan Ghosh, Daguang Xu, Ehsan Hosseini-Asl, Danial Mohseni-Taheri, Vidya Murali, Sifei Liu, Yao Lu, Oluwatobi Olabiyi, Yu-Chiang Frank Wang, Rafael Valle, Bryan Catanzaro, Andrew Tao, Song Han, Jan Kautz, Hongxu Yin, and Pavlo Molchanov. Omnivinci: Enhancing architecture and data for omni-modal understanding LLM. *CoRR*, abs/2510.15870, 2025. doi: 10.48550/ARXIV.2510.15870. URL <https://doi.org/10.48550/arXiv.2510.15870>.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. *CoRR*, abs/2512.12360, 2025. doi: 10.48550/ARXIV.2512.12360. URL <https://doi.org/10.48550/arXiv.2512.12360>.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. Videoexplorer: Think with videos for agentic long-video understanding, 2025. URL <https://arxiv.org/abs/2506.10821>.

Zhenrui Yue, Kartikeya Upasani, Xianjun Yang, Suyu Ge, Shaoliang Nie, Yuning Mao, Zhe Liu, and Dong Wang. Dr. zero: Self-evolving search agents without training data. *CoRR*, abs/2601.07055, 2026. doi: 10.48550/ARXIV.2601.07055. URL <https://doi.org/10.48550/arXiv.2601.07055>.

Chi Zhang, Zhao Yang, Jiaxuan Liu, Yanda Li, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. In Naomi Yamashita, Vanessa Evers, Koji Yatani, Sharon Xianghua Ding, Bongshin Lee, Marshini Chetty, and Phoebe O. Toups Dugas, editors, *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pages 70:1–70:20. ACM, 2025a. doi: 10.1145/3706598.3713600. URL <https://doi.org/10.1145/3706598.3713600>.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *CoRR*, abs/2505.18079, 2025b. doi: 10.48550/ARXIV.2505.18079. URL <https://doi.org/10.48550/arXiv.2505.18079>.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyuan Luo, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *CoRR*, abs/2403.13372, 2024. doi: 10.48550/ARXIV.2403.13372. URL <https://doi.org/10.48550/arXiv.2403.13372>.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *CoRR*, abs/2505.20256, 2025. doi: 10.48550/ARXIV.2505.20256. URL <https://doi.org/10.48550/arXiv.2505.20256>.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *CoRR*, abs/2505.17862, 2025. doi: 10.48550/ARXIV.2505.17862. URL <https://doi.org/10.48550/arXiv.2505.17862>.

Yifan Zhu, Xinyu Mu, Tao Feng, Zhonghong Ou, Yuning Gong, and Haoran Luo. Omnidrag-agent: Agentic omnimodal reasoning for low-resource long audio-video question answering, 2026a. URL <https://arxiv.org/abs/2602.03707>.

Yutao Zhu, Xingshuo Zhang, Maosen Zhang, Jiajie Jin, Liancheng Zhang, Xiaoshuai Song, Kangzhi Zhao, Wencong Zeng, Ruiming Tang, Han Li, Ji-Rong Wen, and Zhicheng Dou. Gisa: A benchmark for general information-seeking assistant, 2026b. URL <https://arxiv.org/abs/2602.08543>.

## Appendix

### A Implementation Details

#### A.1 Training Details

We implement omni-modal agentic SFT and DPO training based on the LlamaFactory codebase (Zheng et al., 2024). Following Section 4, we first perform supervised fine-tuning for 2 epochs on 2,156 synthesized high-quality trajectories, and then continue training with OmniDPO for another 2 epochs to obtain the final OmniAtlas models. We train three backbone scales: Qwen2.5-Omni-[3B,7B] (Xu et al., 2025a), and Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025b). All model parameters are updated during training, including the vision tower, multi-modal projector, and language model. The training experiments were conducted on four nodes of 8 NVIDIA H20-141GB GPUs.

#### A.2 Evaluation Details

We evaluate models using a two-stage procedure that combines exact match with an LLM-as-a-Judge fallback. Given a question, we first attempt to extract the model predicted answer enclosed by `<answer>` and `</answer>`. If an extracted answer exists, we perform an exact string match against the labeled answer. If it matches exactly, the prediction is marked as correct and no LLM judging is used. If an extracted answer exists but does not exactly match, or if no `<answer>...</answer>` span can be extracted, we ignore the extracted span (if any), take the last 20 words of the model output (split by spaces) as the predicted answer, and use LLM-as-a-Judge to determine whether it is equivalent to the labeled answer.

## B Instruction Templates

### B.1 Evaluation Prompt

This prompt implements our LLM-as-a-Judge step for answer equivalence when exact match is insufficient. We constrain the judge to output a single binary label (Correct/Incorrect) to make Pass@1 computation deterministic and avoid leaking intermediate reasoning. The judge takes only the question, the normalized prediction string (the last 20 words), and the labeled answer.

#### LLM-as-a-Judge Prompt (DeepSeek-V3.2)

```
Please determine if the model correctly predicted the answer.  
Question: {question}  
Model Predicted Answer: {predicted}  
Labeled Answer: {standard}  
Return 'Correct' if the model's prediction is completely accurate, otherwise return 'Incorrect'. Provide only this single  
word response.
```

### B.2 System Prompts

This is the unified system prompt used for all base agents in our evaluation to standardize instruction-following, tool usage, and answer formatting across models. We explicitly require the final answer to be wrapped by `<answer>...</answer>` so it can be reliably extracted for exact matching and judging, while leaving the model free to use tools and multi-step reasoning internally.

#### Base Agent System Prompt

```
You are an omni-modal general AI assistant. Please answer the question provided to you based on the input image, audio, or  
video content.
```

```
You should think step by step to answer the question. You may use available tools to assist with your analysis if needed.
```

```
Please provide your final answer using this format: <answer>YOUR_ANSWER</answer>.
```

This prompt equips OmniAtlas with *active omni-modal perception*: when the model is uncertain about specific regions/segments, it can explicitly request additional evidence by calling perception tools. The added note encourages “look/listen-where-needed” behavior (Section 4) instead of passively relying on a single lossy media ingestion, which is critical for long videos and high-resolution images.

### OmniAtlas System Prompt

```
You are an omni-modal general AI assistant. Please answer the question provided to you based on the input image, audio, or video content.
```

```
You should think step by step to answer the question. You may use available tools to assist with your analysis if needed.
```

**\*\*Note:\*\***

- If there are segments in the input image/audio/video that are unclear to you, you should use the "read\_image/read\_audio/read\_video" tool to examine them carefully to ensure you have correctly perceived the input media.

```
Please provide your final answer using this format: <answer>YOUR_ANSWER</answer>.
```

## B.3 Active Omni-Modal Perception Tool Schemas

This schema defines the `read_video` tool used by OmniAtlas to retrieve a specific time window from a long video for higher-fidelity inspection. Exposing `t_start/t_end` enables targeted evidence acquisition and reduces unnecessary context/cost versus loading the entire video.

### Function Schema: `read_video`

```
def get_function_schema_read_video():
    return {
        "type": "function",
        "function": {
            "name": "read_video",
            "description": "Reads a specific time segment of a video file to examine details.",
            "parameters": {
                "type": "object",
                "properties": {
                    "video_id": {"type": "string", "description": "The video identifier or filename."},
                    "t_start": {"type": "integer", "description": "Start time in seconds."},
                    "t_end": {"type": "integer", "description": "End time in seconds."},
                },
                "required": ["video_id", "t_start", "t_end"],
            },
        },
    }
```

This schema defines the `read_audio` tool for selectively listening to a specific time segment. Segment-level access supports pinpointing key speech/non-speech cues and mitigates information loss from global summaries.

### Function Schema: `read_audio`

```
def get_function_schema_read_audio():
    return {
        "type": "function",
        "function": {
            "name": "read_audio",
            "description": "Reads a specific time segment of an audio file to listen to details.",
            "parameters": {
                "type": "object",
                "properties": {
                    "audio_id": {"type": "string", "description": "The audio identifier or filename."},
                    "t_start": {"type": "integer", "description": "Start time in seconds."},
                    "t_end": {"type": "integer", "description": "End time in seconds."},
                },
                "required": ["audio_id", "t_start", "t_end"],
            },
        },
    }
```

```

        },
    },
}
```

This schema defines the `read_image` tool to re-examine images, optionally with a crop box. Cropping enables fine-grained verification of small objects/text without downampling the entire image, aligning with our active perception principle.

#### Function Schema: `read_image`

```

def get_function_schema_read_image():
    return {
        "type": "function",
        "function": {
            "name": "read_image",
            "description": "Reads specific images to view them in detail. Optionally crop the image by providing a crop box [left, top, right, bottom].",
            "parameters": {
                "type": "object",
                "properties": {
                    "image_ids": {"type": "array", "items": {"type": "string"}, "description": "List of image identifiers or filenames."},
                    "crop_box": {
                        "type": "array",
                        "items": {"type": "integer"},
                        "minItems": 4,
                        "maxItems": 4,
                        "description": "Optional. A 4-element list [left, top, right, bottom] specifying the cropping rectangle."
                    },
                    "required": ["image_ids"],
                },
            },
        },
    }
```

#### B.4 Tool-based Perception: Tool Schemas and System Prompts

This schema defines the `audio_qa` perception tool used in our tool-based perception ablations (Table 3) to answer sub-questions from audio only. By wrapping a perception model behind a tool interface, we can isolate whether deficiencies come from perception versus agentic planning/tool use.

#### Perception Tool Schema: `audio_qa`

```

def get_openai_function_audio_qa() -> Dict[str, Any]:
    return {
        "type": "function",
        "function": {
            "name": "audio_qa",
            "description": "Answer the question using audio from audio_path or video_path.",
            "parameters": {
                "type": "object",
                "properties": {
                    "question": {"type": "string", "description": "The question to answer."},
                    "audio_path": {"type": "string", "description": "Audio file path."},
                    "video_path": {"type": "string", "description": "Video file path (audio will be used.)"},
                },
                "required": ["question"],
            },
        },
    }
```

This schema defines the `vision_qa` perception tool for answering sub-questions from visual content only. Together

with `audio_qa`, it enables controlled settings where the base agent can delegate missing modalities to specialized tools.

#### Perception Tool Schema: `vision_qa`

```
def get_openai_function_vision_qa() -> Dict[str, Any]:
    return {
        "type": "function",
        "function": {
            "name": "vision_qa",
            "description": "Answer the question using visual content from image_path or video_path.",
            "parameters": {
                "type": "object",
                "properties": {
                    "question": {"type": "string", "description": "The question to answer."},
                    "image_path": {"type": "string", "description": "Image file path."},
                    "video_path": {"type": "string", "description": "Video file path."},
                },
                "required": ["question"],
            },
        },
    }
```

This is the system prompt for the `audio_qa` tool backend. The prompt strictly restricts the tool to audio evidence and allows abstention (“cannot determine”) to prevent hallucinated cross-modal guesses.

#### System Prompt: Audio QA Prompt

```
You are an audio perception assistant. Answer the question using only the provided audio. If the audio does not contain enough information, say you cannot determine.
```

This is the system prompt for the `vision_qa` tool backend. It enforces a vision-only evidence policy and abstention when visual information is insufficient, making the ablation faithful and verifiable.

#### System Prompt: Vision QA Prompt

```
You are a visual perception assistant. Answer the question using only the provided image or video. If the visual content does not contain enough information, say you cannot determine.
```

## B.5 Perception Analysis Prompts for Data Construction

This prompt converts raw images into a high-density, structured JSON report (OCR, objects, faces, global summary) used as intermediate signals for event graph construction. The “certainty-first” constraint reduces noise and hallucination in downstream graph reasoning, while the structured fields make evidence retrieval and linking explicit.

#### Image Analysis Prompt

```
Please analyze this image and provide a comprehensive structured report in strictly JSON format.
```

##### \*\*Crucial Guidelines:\*\*

1. **\*\*CERTAINTY FIRST\*\*:** Only provide information you are absolutely certain about. Do not guess or hallucinate details from blurry or ambiguous regions. If you are unsure, do not include it.
2. **\*\*COMPREHENSIVE & FACTUAL\*\*:** Focus on extracting strictly factual, objective information. Ensure high detail and density of information to support future analysis. Describe all visible text, objects, and people in detail.

```
The JSON object must contain the following fields:
```

```
'''json
{
    "ocr": [
```

```

        {
            "text": "detected text string", "detailed_features": "detailed features of the text (e.g. location, color, etc)" }
        ],
        "objects": [
            { "label": "object name", "confidence": 0.95, "detailed_features": "detailed features of the object (e.g. location, color, shape, texture, etc)" }
        ],
        "faces": [
            {
                "age": "estimated age range (e.g. 25-30)",
                "gender": "Male/Female",
                "expression": "detailed description of facial expression and emotion",
                "visual_attributes": "clothing, glasses, hair color, distinctive features",
                "activity": "what the person is specifically doing (e.g. reading a book, talking on phone, typing on laptop, cooking, exercising, etc)"
            }
        ],
        "global_summary": "A comprehensive, exhaustive, and highly detailed description of the image content, covering all visible elements, context, actions, and details."
    }
```

```

If a field is not applicable or nothing is detected, return an empty list for that field.

After thinking, output your final response as a JSON code block:

```
'''json
{...}
'''
```

This clip-level prompt produces fine-grained audio annotations for a specific time window, enabling time-aligned evidence mining for long recordings. Clip segmentation improves recall of transient cues (short utterances/events) and supports our pipeline's timestamped linking and multi-hop reasoning.

### Audio Clip Analysis Prompt

You are analyzing a specific audio clip segment from a longer audio/video.

**Clip Context:**

- Total Duration: {total\_duration:.2f} seconds
- Current Clip Range: {start\_time:.2f}s to {end\_time:.2f}s

Please analyze this short audio clip and provide a comprehensive structured report in strictly JSON format.

**Crucial Guidelines:**

1. **CERTAINITY FIRST**: Only provide information you are absolutely certain about. Do not guess unclear speech or ambiguous sounds. If you are unsure, do not include it.
2. **COMPREHENSIVE & FACTUAL**: Concentrate on factual information from both speech and non-speech sounds. Ensure high detail and density of information to support future analysis.

The JSON object must contain the following fields:

```
'''json
{
    "asr": [
        { "text": "transcribed text", "start": 0.0, "end": 2.5, "speaker": "speaker_1" }
    ],
    "speakers": {
        "speaker_1": {
            "gender": "Male/Female",
            "age_estimate": "Adult/Child/Elderly",
            "accent": "description of accent or dialect",
            "tone": "emotional tone (e.g. anxious, authoritative, calm)"
        }
    },
    "events": [
        {
            "label": "event name (e.g. dog barking, siren, applause)",
            "category": "environment/sound_effect/music/speech",
            "start": 1.2,
            "end": 3.5
        }
    ],
}
```

```

    "nonspeech_information": "detailed description of the non-speech information in this clip, focusing on factual and
    certain information",
    "global_summary": "A comprehensive, exhaustive, and highly detailed summary of the clip content, covering all speech,
    sound events, background noises, and emotional cues, focusing on factual and certain information"
}
```
If no speech is detected, 'asr' should be an empty list.
If no specific events are detected, 'events' should be an empty list.
After thinking, output your final response as a JSON code block:

```json
{...}
```

```

This global prompt produces an overall structured audio report for the entire recording, complementing clip-level details with global context. We keep the same schema as the clip prompt to make aggregation consistent, while allowing longer ASR segmentation and higher-level summaries for event graph nodes.

## Audio Global Analysis Prompt

Please analyze this audio and provide a comprehensive structured report in strictly JSON format.

**\*\*Crucial Guidelines:\*\***

1. **\*\*CERTAINTY FIRST\*\*:** Only provide information you are absolutely certain about. Do not guess unclear speech or ambiguous sounds. If you are unsure, do not include it.
2. **\*\*COMPREHENSIVE & FACTUAL\*\*:** Concentrate on factual information from both speech and non-speech sounds. Ensure high detail and density of information to support future analysis.

The JSON object must contain the following fields:

```

```json
{
    "asr": [
        // If a single person speaks for a long time, segment the speech into pieces, with each segment containing one piece
        // of information
        { "text": "transcribed text", "start": 0.0, "end": 2.5, "speaker": "speaker_1" }
    ],
    "speakers": {
        "speaker_1": {
            "gender": "Male/Female",
            "age_estimate": "Adult/Child/Elderly",
            "accent": "description of accent or dialect",
            "tone": "emotional tone (e.g. anxious, authoritative, calm)"
        }
    },
    "events": [
        {
            "label": "event name (e.g. dog barking, siren, applause)",
            "category": "environment/sound_effect/music/speech",
            "start": 1.2,
            "end": 3.5
        }
    ],
    "nonspeech_information": "detailed description of the non-speech information in the audio, focusing on factual and
    certain information",
    "global_summary": "A comprehensive, exhaustive, and highly detailed summary of the audio content, covering all speech,
    sound events, background noises, and emotional cues, focusing on factual and certain information"
}
```

```

If no speech is detected, 'asr' should be an empty list.  
If no specific events are detected, 'events' should be an empty list.  
After thinking, output your final response as a JSON code block:

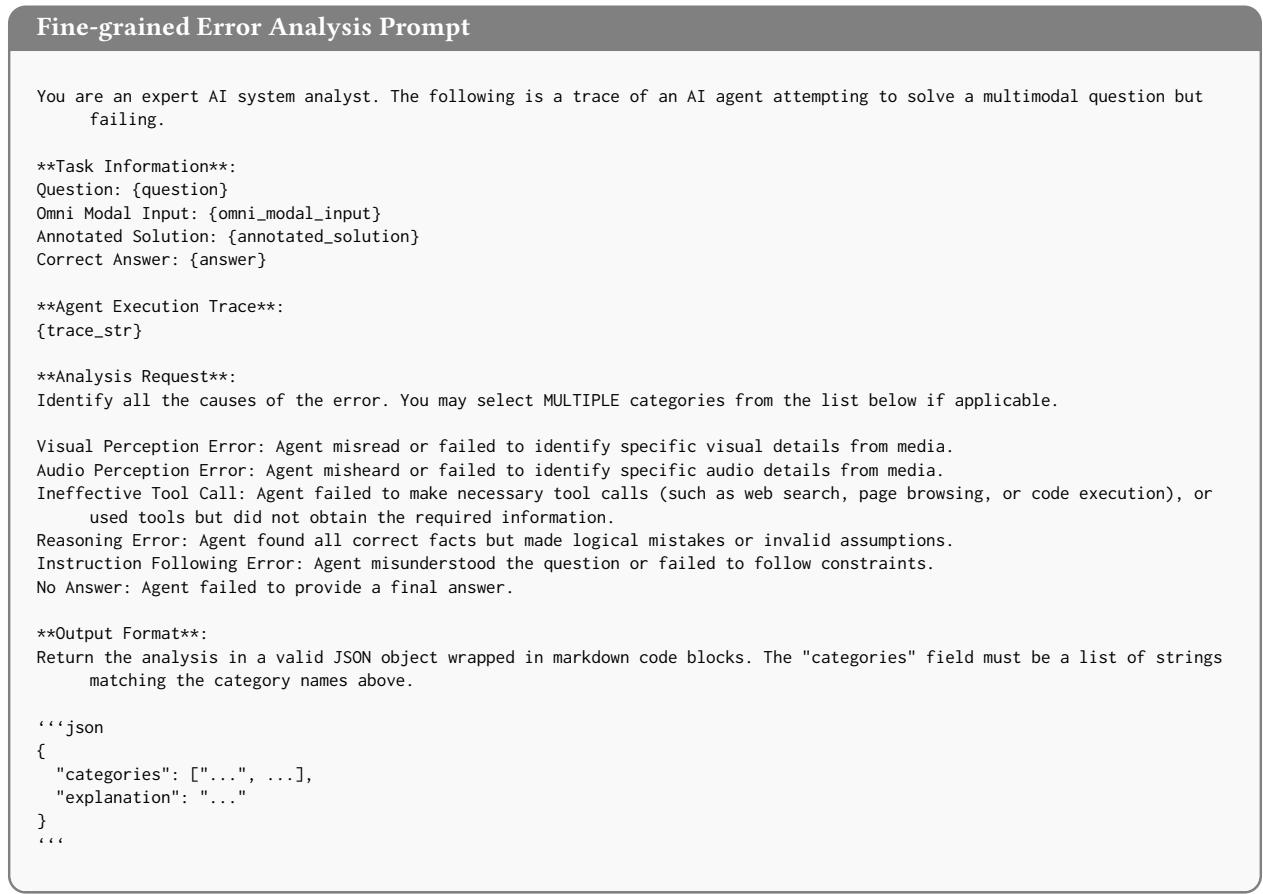
```

```json
{...}
```

```

## B.6 Error Analysis Prompt

This prompt supports our fine-grained error taxonomy analysis (Figure 5) by labeling failure causes from full execution traces. We allow multi-label categorization to capture cascade failures (e.g., tool misuse leading to reasoning errors) and require a JSON output for easy aggregation and reproducibility.



## C Detailed Related Work

### C.1 Omni-Modal Foundation Models and Benchmarks

Building on advances in pure-text (Dubey et al., 2024), vision-language (Hurst et al., 2024), and audio-language (Chu et al., 2024) foundation models, recent omni-modal models seek to unify text, vision, and audio within a single LLM backbone. A common approach adopts a unified tokenization-and-projection interface that maps heterogeneous visual and acoustic inputs into a shared token space (Xu et al., 2025b; Liu et al., 2025a; Luo et al., 2025b; Ye et al., 2025; Liu et al., 2025b). Concurrent work further strengthens omni-modal reasoning behaviors (Zhong et al., 2025; Chen et al., 2025b; Wang et al., 2025; Chen et al., 2026b), token compression (Ding et al., 2026), and reward modeling (Jin et al., 2025b; Kong et al., 2026). For evaluation, existing benchmarks (e.g., OmniBench (Li et al., 2024), WorldSense (Hong et al., 2025) and Daily-Omni (Zhou et al., 2025)) largely emphasize short audios/videos and perception-centric tasks, leaving long-horizon reasoning and tool-integrated agency underexplored. This gap hinders complex, interactive real-world applications.

### C.2 Autonomous Agents

LLM-driven autonomous agents tackle real-world tasks by reasoning and acting through external tools that interface with their environment (Wang et al., 2024b; Luo et al., 2025a; Zhu et al., 2026b; Wu et al., 2025c). Existing approaches broadly fall into workflow-based paradigms (Yao et al., 2022; Wang et al., 2023; Hu et al., 2025) and native agentic

**Table 5 Case Study I (Failure).** Qwen3-Omni-30B-A3B on an OmniGAIA video-with-audio question. The model fails to ground the bridge to the Joliet Iron Works context and does *not* invoke any external tools (0 tool calls), leading to an unverified Chicago-bridge prior and an incorrect final answer. ([ . . . ] denotes omitted trace content for brevity.)

| Example from OmniGAIA  |
|--|
| <b>Question:</b><br>During a visit to the Joliet Iron Works Historic Site as shown in the video, the speaker spots a movable bridge in the distance and remarks that it reminds him of a bridge featured in the movie <i>The Blues Brothers</i> . What is the name of this bridge, and how many years had it been standing when filming for <i>The Blues Brothers</i> began?   |
| <b>Labeled Answer:</b><br>Ruby Street Bridge; 44   |
| Model Output by Qwen3-Omni-30B-A3B   |
| <b>Tool Calls:</b> 0 (tools available but unused).<br>(Condensed) The model notes a potential mismatch between the question context ( <b>Joliet Iron Works</b> , the speaker pointing to “a bridge … it’s going down”) and its movie-location prior (well-known Chicago bridges in <i>The Blues Brothers</i> ). It briefly considers whether Joliet landmarks (e.g., Old Joliet Prison) could imply a local bridge, but ultimately treats the question as asking for the bridge <i>featured in the movie</i> and selects the <b>LaSalle Street Bridge</b> in Chicago. Without using tools to verify, it assumes a completion year of 1928, takes filming to begin in 1979, and computes $1979 - 1928 = 51$ years, yielding <answer>LaSalle Street Bridge, 51</answer>. [...] |
| Error Analysis   |
| <b>Error Categories:</b> Visual Perception Error; Ineffective Tool Call; Reasoning Error.<br><b>Why it fails:</b> (i) It does not use the video/audio context to anchor the bridge to Joliet; (ii) it under-calls tools despite tool availability; (iii) it relies on an unverified prior about a Chicago bridge and its construction year, producing an incorrect age.<br><b>Evaluation:</b> EM=0, LLM-as-a-Judge=Incorrect.  |

reasoning methods (Li et al., 2025d,g,f; Feng et al., 2025; Jiang et al., 2025; Dong et al., 2025b; Wu et al., 2025a; Jin et al., 2025a; Dong et al., 2025a; Li et al., 2025e; Chen et al., 2026a; Hu et al., 2026; Tan et al., 2025; Yue et al., 2026), and have shown strong performance on text-only tasks. Moving beyond text, recent studies investigate vision-language agents for multimodal web search (Li et al., 2025c; Wu et al., 2025b; Geng et al., 2025), long-form video understanding (Wang et al., 2024c; Yuan et al., 2025; Zhang et al., 2025b; Yin et al., 2025; Lin et al., 2025; Tao et al., 2025; Zhu et al., 2026a), and GUI navigation (Xie et al., 2024; Zhang et al., 2025a; Wang et al., 2024a; Lu et al., 2026). However, *omni-modal* foundation agents that natively fuse audio, vision, and language while performing long-horizon agentic reasoning remain underexplored. Such capabilities are essential for building general AI assistants in real-world scenarios.

## D Case Study

We analyze three execution traces on the *same* OmniGAIA instance (Tables 5, 6, and 7) to highlight a key lesson for omni-modal agents: tool access is necessary but not sufficient. The instance contains a deliberate distraction—the mention of *The Blues Brothers*—which can trigger a strong *Chicago-bridge prior*. The correct solution instead requires location-first grounding at Joliet Iron Works and then evidence-backed identification of the nearby movable bridge (Ruby Street Bridge, built 1935), followed by a simple computation for filming start in July 1979 ( $1979 - 1935 = 44$ ).

### D.1 What Capabilities Does This Instance Stress?

This instance stresses a tightly-coupled chain of capabilities:

- **Omni-modal grounding (location-first):** anchor the bridge to the Joliet Iron Works context, instead of following movie-location priors.
- **Tool planning & query formulation:** issue *entity- and location-specific* queries (e.g., “Joliet Iron Works” + “Ruby Street Bridge”), rather than underspecified Chicago-centric searches.
- **Hypothesis testing & verification:** treat early guesses as hypotheses, and actively seek disconfirming/local evidence before committing to a bridge identity and construction year.
- **Computation after verification:** use a calculator/code tool only after the facts are grounded (here,  $1979 - 1935$ ).

- **Answer normalization:** output a concise, extractable final answer aligned with the evaluation protocol.

## D.2 Case I: Failure by Under-Calling (No Tools)

In Case I (Table 5), the model fails early due to premature closure on a movie-driven prior. It does not use tools at all, so it never retrieves the decisive local evidence that ties the scene to the Ruby Street Bridge near Joliet Iron Works, nor does it verify the construction year and filming start date. As a result, it outputs a confident but unverified bridge name and an incorrect age.

## D.3 Case II: Failure by Tool-Query Drift (Tools Used, Wrong Hypothesis)

Case II (Table 6) shows a different and more subtle failure: the model *does* call tools, but its retrieval is locked onto the initial wrong hypothesis (a Chicago bridge). This produces confirmation bias: each search result reinforces the Chicago interpretation, and the agent never executes a corrective, location-specific query that could recover the Joliet bridge identity. The downstream computation is then correct arithmetic over wrong facts (e.g., subtracting from an incorrect construction year), illustrating that computation tools cannot compensate for upstream grounding/retrieval errors.

## D.4 Case III: Success via Location-First Disambiguation and Verification

In Case III (Table 7), the agent follows a robust pattern: it first disambiguates by location (Joliet Iron Works → nearby movable bridge), then performs targeted retrieval to verify the bridge identity and construction year (Ruby Street Bridge, 1935) and the filming start time (July 1979), and finally uses a computation tool for  $1979 - 1935 = 44$ . Notably, even when the generated response includes extra explanation and fails strict exact match, the answer remains semantically correct and is recovered by our LLM-as-a-Judge protocol.

**Takeaway.** Together, these three traces isolate two dominant failure modes for current omni-modal agents: (i) under-calling (insufficient evidence acquisition) and (ii) tool-query drift (evidence acquisition guided by an incorrect prior). The successful trace suggests a practical mitigation: location-first grounding + hypothesis testing + verification before computation, followed by strict answer normalization.

**Table 6 Case Study II (Failure).** OmniAtlas-Qwen3-30B-A3B on the same OmniGAIA instance. The model calls tools but exhibits query drift and confirmation bias, resulting in correct computation over incorrect facts. ([ . . . ] denotes omitted trace content for brevity.)

### Example from OmniGAIA

#### Question:

During a visit to the Joliet Iron Works Historic Site as shown in the video, the speaker spots a movable bridge in the distance and remarks that it reminds him of a bridge featured in the movie \*The Blues Brothers\*. What is the name of this bridge, and how many years had it been standing when filming for The Blues Brothers commenced in July 1979?

#### Labeled Answer:

Ruby Street Bridge; 44

### Model Output by OmniAtlas-Qwen3-30B-A3B

I can see in the video that the speaker is at the Joliet Iron Works Historic Site, and I can hear him talking about the surroundings. He mentions seeing a bridge in the distance that reminds him of one from The Blues Brothers movie. Based on what I hear, the bridge is the LaSalle Street Bridge in Chicago. However, I need to verify this and calculate how many years it had been standing when filming for The Blues Brothers started in July 1979. Let me search for more information about this bridge.

`<tool_call> "name": "web_search", "arguments": "query": "LaSalle Street Bridge Chicago The Blues Brothers filming July 1979 construction date" </tool_call>`

`<tool_response> [...] The Blues Brothers - Filming Locations of Chicago and Los Angeles [...] LaSalle Street Bridge, Chicago [...]`

`</tool_response>`

Let me get more specific information about the LaSalle Street Bridge and its construction date.

`<tool_call> "name": "web_search", "arguments": "query": "LaSalle Street Bridge construction date 1885 Chicago" </tool_call>`

`<tool_response> [...] Originally built as the Water Street Bridge in 1885, it was later renamed the LaSalle Street Bridge [...]`

`</tool_response>`

Now I need to calculate how many years the LaSalle Street Bridge had been standing when filming for The Blues Brothers commenced in July 1979. According to the search results, the bridge was originally built in 1885. Let me calculate this:

`<tool_call> "name": "code_executor", "arguments": "code": "# Calculate years from 1885 to July 1979\nconstruction_year = 1885\nfilming_year = 1979\nyears_standing = filming_year - construction_year\nprint(f'The LaSalle Street Bridge was built in {construction_year}\')\nprint(f'Filming for The Blues Brothers commenced in July {filming_year}\')\nprint(f'Number of years standing: {years_standing}')" </tool_call>`

`<tool_response> The LaSalle Street Bridge was built in 1885 Filming for The Blues Brothers commenced in July 1979 Number of years standing: 94 </tool_response>`

Let me verify the construction year with another source.

`<tool_call> "name": "web_search", "arguments": "query": "Water Street Bridge 1885 LaSalle Street Bridge Chicago" </tool_call>`

`<tool_response> [...] the Water Street Bridge in 1885, it was later renamed the LaSalle Street Bridge after the street name was changed [...]`

Now I have clear information. Let me also search for more specific information about the filming dates for The Blues Brothers in Chicago.

`<tool_call> "name": "web_search", "arguments": "query": "The Blues Brothers filming began July 1979 Chicago" </tool_call>`

`<tool_response> [...] Much of The Blues Brothers was shot on location in and around Chicago between July and October 1979 [...]`

Based on all the information gathered, I can now answer the question. Let me analyze what I've learned: 1. From the video: The speaker spots a movable bridge in the distance at the Joliet Iron Works Historic Site. 2. From the search results: The bridge referenced is the LaSalle Street Bridge (originally known as the Water Street Bridge) in Chicago. 3. From multiple sources: The bridge was constructed in 1885. 4. From the search results: Principal photography began in July 1979. Therefore, the bridge spotted by the speaker at the Joliet Iron Works Historic Site is the LaSalle Street Bridge (originally known as the Water Street Bridge) in Chicago. It was built in 1885. When filming for The Blues Brothers commenced in July 1979, the bridge had been standing for 94 years (1979 - 1885 = 94 years).

### Error Analysis

**Main Errors:** 1. **Incorrect Bridge Identification:** Model incorrectly identified the bridge as the **LaSalle Street Bridge** in Chicago, while the correct answer is the **Ruby Street Bridge** in Joliet (near the Joliet Iron Works Historic Site). 2. **Incorrect Construction Year:** Model used 1885 as construction year, but Ruby Street Bridge was actually built in 1935. 3. **Incorrect Age Calculation:** Based on wrong construction year, calculated 94 years instead of correct 44 years (1979 - 1935).

**Reason for Errors:** 1. Over-relied on general knowledge about *The Blues Brothers* movie (which features Chicago bridges) rather than focusing on the specific location mentioned in the question. 2. Failed to search for the specific bridge near Joliet Iron Works Historic Site, using instead generic Chicago bridge queries. 3. Did not verify local information about Joliet area bridges.

**Correct Information:** • **Bridge Name:** Ruby Street Bridge • **Location:** Near Joliet Iron Works Historic Site, over Des Plaines River • **Construction Year:** 1935 • **Filming Start:** July 1979 • **Age at Filming:** 44 years (1979 - 1935)

**Table 7 Case Study III (Success).** Gemini-3-Pro on the same OmniGAIA instance (condensed). The model grounds the location, retrieves the bridge construction year and filming start date with web search, uses code execution for arithmetic, and produces an answer judged equivalent to the label. ([ . . . ] denotes omitted trace content for brevity.)

| Example from OmniGAIA  |
|--|
| <p><b>Question:</b><br/>         During a visit to the Joliet Iron Works Historic Site as shown in the video, the speaker spots a movable bridge in the distance and remarks that it reminds him of a bridge featured in the movie <i>The Blues Brothers</i>. What is the name of this bridge, and how many years had it been standing when filming for <i>The Blues Brothers</i> commenced in July 1979?</p> <p><b>Labeled Answer:</b><br/>         Ruby Street Bridge; 44</p>  |
| <p><b>Model Output by Gemini-3-Pro</b></p> <p>The bridge is likely the Ruby Street Bridge in Joliet. I will verify the bridge name and its construction year, then compute its age in July 1979.</p> <pre> &lt;tool_call&gt; "name": "web_search", "arguments": "query": "Joliet Iron Works Historic Site Ruby Street Bridge Columbia Street east of Ruby Street Bridge" &lt;/tool_call&gt; &lt;tool_response&gt; [...] Joliet Iron Works Historic Site is located on Columbia Street, just east of the Ruby Street Bridge [...] &lt;/tool_response&gt; &lt;tool_call&gt; "name": "web_search", "arguments": "query": "Ruby Street Bridge Joliet year built 1935" &lt;/tool_call&gt; &lt;tool_response&gt; [...] Ruby Street Bridge [...] Year Built: 1935 [...] &lt;/tool_response&gt; &lt;tool_call&gt; "name": "web_search", "arguments": "query": "The Blues Brothers filming began July 1979" &lt;/tool_call&gt; &lt;tool_response&gt; [...] principal photography started in July 1979 [...] &lt;/tool_response&gt; &lt;tool_call&gt; "name": "code_executor", "arguments": "code": "1979 - 1935" &lt;/tool_call&gt; &lt;tool_response&gt; 44 &lt;/tool_response&gt;</pre> <p>Based on the video and the location of the <b>Joliet Iron Works Historic Site</b>, the movable bridge visible in the distance is the <b>Ruby Street Bridge</b> (also known as the <b>Ruby Street Bascule Bridge</b>, over the Des Plaines River).</p> <p>The speaker remarks that the bridge reminds him of a bridge featured in <i>The Blues Brothers</i>. While the famous bridge jump scene in the movie was filmed at the <b>East 95th Street Bridge</b> in Chicago, the Ruby Street Bridge in Joliet has a similar appearance (both are bascule bridges) and is located near the Joliet Correctional Center (Joliet Prison), where the movie's opening scenes were filmed.</p> <p>The Ruby Street Bridge was opened in <b>1935</b>. Filming for <i>The Blues Brothers</i> began in <b>July 1979</b>. Therefore, the bridge had been standing for <b>44</b> years when filming commenced.</p> <p>&lt;answer&gt;The bridge is the Ruby Street Bridge (or Ruby Street Bascule Bridge). It had been standing for 44 years when filming for The Blues Brothers commenced (1979 - 1935 = 44). &lt;/answer&gt;</p> |
| <p><b>Evaluation</b></p> <p>Exact match fails due to formatting differences, but LLM-as-a-Judge determines equivalence (Correct).</p>  |