WHEN HISTORY HELPS AND HURTS: SELECTIVE HISTORY USE ACROSS MULTIMODAL TURNS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable multimodal interaction depends on selective use of conversational history: an earlier question may remain relevant while its previous answer is outdated, whereas a current request may depend on historical evidence despite conflicting new observations. Existing multi-turn evaluations rarely separate these history-use demands from underlying question difficulty. To address this gap, we introduce ReTurn, a benchmark of 7,000 base tasks spanning visual and audio evidence for evaluating selective history use. For task-carrying history, Reconfirm/Reground require applying a historical question to current media while varying historical agreement; for evidence-carrying history, Retrieve/Rebind require answering a current question using historical media while varying current-media competition. Each pair preserves the target question, media, and answer. Tasks support open-ended and multiple-choice evaluation, with matched single-turn counterparts serving as answerability references. Across 13 omni-modal, vision-language, and audio-language models, median model-level open-ended accuracy falls from 93.7% with direct input to 72.3% in conversation. Behavioral probes show that high question recall can coexist with weaker task application, while competing media can redirect answers away from historical targets. Supervised adaptation yields only partial gains. ReTurn provides a controlled framework for assessing whether multimodal models select and use the historical information required by each request.

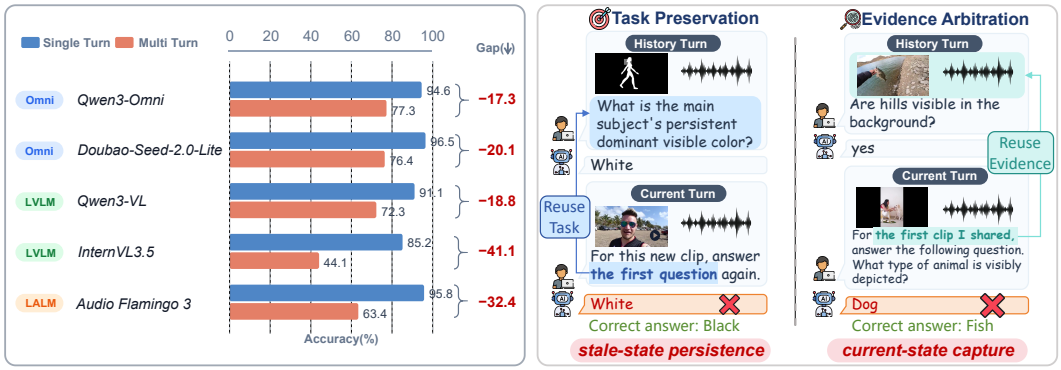
1 INTRODUCTION

Multimodal large language models (MLLMs) increasingly support sustained interaction across visual and audio inputs. Vision-language, audio-language, and omni-modal systems allow users to discuss successive observations, revisit earlier media, and continue a task without restating it (Team, 2025b; Ghosh et al., 2025; Team, 2025a; Tong et al., 2025). Evaluation has accordingly expanded to multi-image dialogue and streaming audiovisual interaction (Liu et al., 2024b; Wang et al., 2025b). In these settings, conversational history supplies questions and evidence that may be absent from the current turn but remain essential to answering it.

However, history is not uniformly beneficial. When a user applies an earlier question to new media, the question remains necessary, but its previous answer may be outdated. Conversely, a request about earlier media requires using historical evidence despite conflicting current observations. These complementary demands call for **selective history use** (Figure 1). Existing evaluations examine conversational memory, information updating, and multi-turn performance (Wu et al., 2025; Maharana et al., 2024; Xue et al., 2025; Laban et al., 2026). Yet assessing this ability requires distinguishing history-use difficulties from underlying question difficulty, while testing how misleading context affects the use of necessary history.

To address this need, we introduce **ReTurn**, a benchmark for evaluating selective history use through two roles of history. *Task-carrying history* supplies a question to apply to current media; *Reconfirm/Reground* vary whether historical media supports the same answer or a different one. *Evidence-carrying history* supplies the media needed to answer the current question; *Retrieve/Rebind* vary competition from current non-target media. Each local pair preserves the target question, media, and correct answer, while a matched single-turn counterpart assesses answerability with the question and target media supplied directly. ReTurn comprises 7,000 base tasks spanning visual and

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



(a) Overall Conversational Degradation (b) Selective History Failures

Figure 1: **Selective history use in ReTurn.** (a) Single-turn versus multi-turn OpenQA accuracy across model families, evaluated on supported modalities. (b) Illustrative failures in applying a historical question to current media (Reground) and answering a current question from historical media despite competition (Rebind).

audio evidence, with paired open-ended (OpenQA) and multiple-choice (MCQ) views. Evaluation retains each model's historical replies, allowing earlier errors to remain part of the context.

Models often answer the underlying questions reliably with direct input yet struggle in conversation. Across 13 omni-modal models, large vision-language models (LVLMs), and large audio-language models (LALMs), median model-level OpenQA accuracy falls from 93.7% in Single-turn to 72.3% in Multi-turn. Long conversations expose complementary difficulties: applying a historical question despite its outdated answer (Reground versus Reconfirm), and preserving a historical target despite current competitors (Rebind versus Retrieve).

These findings motivate two hypothesized demands: *history access*, recovering required historical information, and *cross-turn binding*, pairing questions with target evidence across turns despite competing within-turn associations. Behavioral probes show high question recall alongside weaker application, while swapping current competitors can redirect answers away from unchanged historical targets. Inference-time assistance and supervised adaptation yield only partial gains, motivating evaluation of both historical information recovery and its use under interference.

Our contributions are threefold:

- We identify selective history use as a requirement for multimodal interaction, highlighting the dual role of history as essential information and a potential source of interference.
- We introduce ReTurn, an omni-modal benchmark that distinguishes task-carrying from evidence-carrying history, with matched pairs and single-turn references.
- We evaluate omni-modal models, LVLMs, and LALMs, revealing limitations in selective history use and investigating failures through behavioral diagnostics and mitigation studies.

2 Related Work

2.1 Multimodal Models for Interactive Systems

Multimodal models have broadened the range of evidence available to conversational assistants. Vision-language models such as Qwen3-VL and InternVL3.5 support image and video understanding (Team, 2025b; Wang et al., 2025a), while Audio Flamingo 3 and MiMo-Audio support speech and general audio (Ghosh et al., 2025; Xiaomi, 2025). Alongside these modality-specific systems, omni-modal models integrate visual, acoustic, and textual processing, while InteractiveOmni explicitly targets audiovisual multi-turn dialogue (Team, 2025a; Cui et al., 2026; Tong et al., 2025). These interfaces allow successive requests to build on previously supplied media rather than isolated inputs. As interaction extends across turns, these capabilities also require models to reuse earlier questions and observations according to the user's current request.

Table 1: Representative multi-turn benchmarks. Scale retains native units (q.: questions; dlg.: dialogues; vid.: videos), so counts are not directly comparable. Checks denote explicit designs: **Update**, revised information; **Direct**, matched direct-input reference; **Role**, separate task-/evidence-carrying history; **Conflict**, target-preserving reference/conflict pairs.

Benchmark	Scope		Structure		History Design	
	Modality	Scale	Update	Direct	Role	Conflict
LongMemEval	T	500 q.	✓	✗	✗	✗
Lost in Conversation	T	600 tasks	✗	✓	✗	✗
MMDU	V+T	110 dlg.	✗	✗	✗	✗
MultiVerse	V+T	647 dlg.	✗	✗	✗	✗
MMRC	V+T	5,120 dlg.	✓	✗	✗	✗
MT-Video-Bench	V+T	1,000 dlg.	✗	✗	✗	✗
Audio MultiChallenge	A+T	452 dlg.	✓	✗	✗	✗
OmniMMI	V+A+T	1,121 vid.	✗	✗	✗	✗
ReTurn (Ours)	**V+A+T**	**7,000 tasks**	✓	✓	✓	✓

2.2 Multi-turn Evaluation and History Use

ConvBench, MMDU, and MultiVerse evaluate multi-turn visual dialogue (Liu et al., 2024a;b; Lee et al., 2025), while MT-Video-Bench, Audio MultiChallenge, and OmniMMI cover video, audio, and streaming audiovisual interaction (Pan et al., 2026; Gosai et al., 2026; Wang et al., 2025b). LoCoMo, LongMemEval, and MMRC study conversational memory (Maharana et al., 2024; Wu et al., 2025; Xue et al., 2025). Laban et al. (2026) compare single- and multi-turn performance with matched task content. ReTurn examines history as both necessary information and potential interference: matched pairs preserve the dependency and target while varying historical agreement or current-media competition. Single-turn references assess target answerability (Table 1).

3 ReTurn: Benchmarking Selective History Use

ReTurn evaluates selective history use through matched conversations built from verified media–question pairs. Each construction preserves the information needed to answer the final request while varying contextual agreement or competition (Figure 2).

3.1 Task Design

The final request depends on history in two ways: it can reuse an earlier question with current media, or ask a current question about historical media. These dependencies define task-carrying and evidence-carrying history, respectively.

Task-carrying history. Reconfirm and Reground ask the model to answer the first question about current target media. The final request and current target remain fixed, while the initial media support the target answer in Reconfirm and a different answer in Reground. The question must persist while the answer may change; Reconfirm alone does not establish current-media inspection, since its historical answer remains correct.

Evidence-carrying history. Retrieve and Rebind provide a current question about historical target media alongside current non-target media. They preserve the target and preceding dialogue while replacing a weaker competitor (Retrieve) with media supporting a plausible different answer (Rebind). Both require the historical source; Retrieve reduces competition without guaranteeing its absence.

Each pair fixes the target question and media, answer, options, historical dependency, and horizon. Agreement and competition concern media-supported answers, not model replies. These are local matched contrasts rather than a full factorial design (Appendix A).

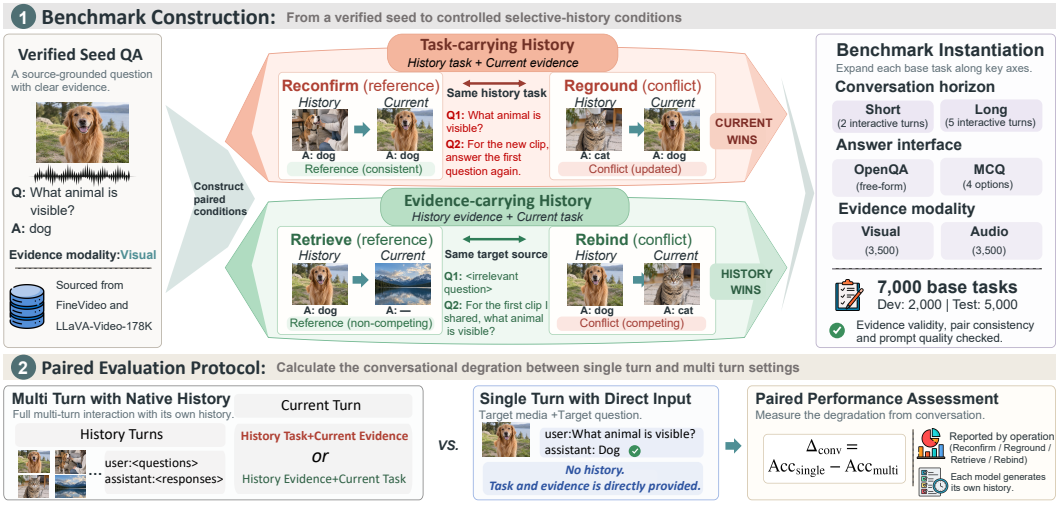


Figure 2: **ReTurn construction and evaluation.** Verified media–question pairs yield matched task-carrying and evidence-carrying conditions. Media annotations indicate supported answers; evaluation uses model-generated histories and matched single-turn references.

3.2 Dataset Construction

Construction and verification. We localize clips from FineVideo (Farré et al., 2024) and LLaVA-Video-178K (Zhang et al., 2024); source annotations guide retrieval rather than supply answer evidence. Doubao-Seed-2.0-Lite (Seed, 2026) verifies questions and answers against actual visual or acoustic content. We pair media by their answer relations; GPT-5.6 Sol (OpenAI, 2026) assists clarity, consistency, and leakage checks. Evidence-carrying histories use questions distinct from the final request to reduce answer copying. Provenance and automated audits appear in Appendices A and D.

Coverage and task views. ReTurn contains 7,000 tasks in 3,500 matched pairs, with 1,750 tasks per operation and balanced visual/audio evidence and Short/Long horizons. Evidence modality identifies the required channel, not input restriction. Short/Long contain two/five media-bearing user turns, including the final turn, and are not universally item-matched. OpenQA and four-option MCQ provide two views per task (14,000 total). Task-carrying MCQ inherits the original options and letter mapping.

Development and partitioning. Development, primarily with Qwen3-Omni, prioritized direct answerability and measurable history effects to emphasize history use over question difficulty; universal correctness and large degradation were not required. The split comprises Dev (2,000: Dev-Train 1,500; Dev-Val 500) and Test (5,000), using pair and fact-group metadata rather than final model scores. Shared media and development exposure limit unseen-source generalization (Appendix A).

3.3 Evaluation Protocol

Single-turn and multi-turn inputs. Single-turn supplies the complete question and target media directly, including historical target media when needed. Multi-turn evaluates the final request after model-generated historical replies, separately for each interface. Earlier errors remain; changed historical media can also change assistant replies. Models consume supported modalities (Appendix B).

Scoring and comparisons. We score final-answer content accuracy separately for OpenQA and MCQ. Frozen rules provide initial scores; rejected responses receive GPT-5.6 Sol review blinded to model and condition. Uncertain judgments count as incorrect. Scoring rules and alternative-judge audits appear in Appendix D. The *conversational degradation*, $\Delta_{\text{conv}} = \text{Acc}_{\text{single}} - \text{Acc}_{\text{multi}}$, is reported in percentage points (pp). Within-pair comparisons instead assess sensitivity to historical agreement or current-media competition. Scores use valid single-turn/multi-turn response pairs, with per-model aggregation detailed in Appendix C.

Table 2: **ReTurn results across 13 models.** Cells report Multi-turn accuracy (%) and the drop (↓) or gain (↑) from Single-turn in pp. Darker red indicates larger drops; blue indicates gains. Models are evaluated on supported modalities. Full Single-turn scores and valid-pair denominators appear in Appendix C.

Model Name	OpenQA					MCQ				
	Overall	Reconfirm	Reground	Retrieve	Rebind	Overall	Reconfirm	Reground	Retrieve	Rebind
(a) Short conversations · 2 native interactive turns										
OMNI · Visual + audio evidence										
Qwen3.8-Omni-Flash	91.1 (↓3.4)	94.4 (↓1.8)	95.0 (↓1.1)	94.4 (↑1.4)	80.6 (↓12.3)	93.4 (↓5.4)	98.6 (↓0.0)	93.9 (↓4.6)	99.0 (↓0.0)	81.9 (↓17.1)
Qwen3-Omni	91.6 (↓3.0)	95.4 (↑0.3)	91.0 (↓4.0)	94.2 (↓0.0)	85.9 (↓8.3)	89.7 (↓9.4)	97.0 (↓1.6)	67.4 (↓31.2)	99.2 (↓0.3)	95.2 (↓4.3)
Doubao-Seed-2.0-Lite	89.8 (↓6.6)	95.4 (↓2.1)	93.1 (↓4.3)	94.7 (↓0.6)	75.8 (↓19.5)	89.0 (↓10.4)	98.7 (↓0.6)	71.8 (↓27.5)	98.9 (↓0.6)	86.7 (↓12.8)
MiniCPM-o 4.5	78.7 (↓14.4)	88.5 (↓5.4)	80.8 (↓13.1)	91.0 (↓1.3)	54.4 (↓37.9)	68.4 (↓29.2)	96.0 (↓1.3)	32.3 (↓65.0)	96.5 (↓1.4)	49.0 (↓49.0)
MiMo-V2.5	80.2 (↓11.2)	90.1 (↓2.6)	79.5 (↓13.1)	87.7 (↓2.6)	63.5 (↓26.7)	75.9 (↓21.7)	89.0 (↓8.8)	46.9 (↓50.9)	96.3 (↓1.1)	71.4 (↓26.1)
Nemotron-3-Nano-Omni	85.9 (↓5.2)	92.5 (↑2.4)	75.4 (↓14.7)	93.4 (↑1.4)	82.2 (↓9.8)	78.7 (↓19.0)	99.0 (↑1.1)	34.4 (↓63.5)	99.4 (↑1.9)	82.1 (↓15.4)
LVLM · Visual evidence										
Qwen3.7-Plus	92.6 (↓0.9)	94.6 (↓0.3)	92.0 (↓2.9)	93.3 (↑1.3)	90.4 (↓1.6)	93.5 (↓4.7)	98.4 (↓0.3)	80.1 (↓18.6)	98.1 (↑0.3)	97.4 (↓0.3)
Qwen3-VL	87.9 (↓3.1)	90.7 (↓1.3)	87.8 (↓4.2)	90.7 (↑0.6)	82.4 (↓7.7)	80.3 (↓17.6)	98.7 (↑1.6)	32.1 (↓65.1)	98.4 (↓0.3)	92.0 (↓6.7)
MiniCPM-V 4.5	82.3 (↓6.8)	88.8 (↑1.3)	60.9 (↓26.6)	91.7 (↑1.0)	87.9 (↓2.9)	76.4 (↓20.7)	94.2 (↓2.2)	20.8 (↓75.6)	97.8 (↓0.0)	92.7 (↓5.1)
InternVL3.5	56.4 (↓28.7)	42.9 (↓37.8)	29.8 (↓51.0)	87.5 (↓1.9)	65.1 (↓24.4)	70.5 (↓24.6)	93.3 (↓1.0)	11.9 (↓82.4)	97.8 (↑1.9)	78.9 (↓16.9)
LALM · Audio evidence										
StepAudio 3	80.0 (↓18.7)	99.0 (↑0.6)	99.4 (↑1.0)	87.5 (↓11.6)	34.1 (↓65.0)	84.8 (↓15.2)	100.0 (↓0.0)	100.0 (↓0.0)	99.7 (↓0.3)	39.5 (↓60.5)
Audio Flamingo 3	72.6 (↓23.2)	95.5 (↓2.2)	47.3 (↓50.5)	93.3 (↓0.6)	54.5 (↓39.4)	65.8 (↓34.0)	100.0 (↑0.3)	6.4 (↓93.3)	99.4 (↓0.6)	57.7 (↓42.3)
MiMo-Audio	62.7 (↓32.2)	86.3 (↓9.9)	72.5 (↓23.6)	73.1 (↓20.5)	18.9 (↓74.7)	49.6 (↓48.2)	97.4 (↓1.3)	13.1 (↓85.6)	66.0 (↓30.8)	21.8 (↓75.0)
(b) Long conversations · 5 native interactive turns										
OMNI · Visual + audio evidence										
Qwen3.8-Omni-Flash	83.2 (↓11.8)	91.3 (↓5.6)	73.1 (↓23.9)	93.8 (↑0.8)	74.6 (↓18.4)	90.2 (↓8.6)	95.3 (↓3.2)	89.9 (↓8.7)	98.9 (↓0.2)	76.8 (↓22.2)
Qwen3-Omni	62.9 (↓31.7)	84.8 (↓10.1)	72.5 (↓22.4)	74.7 (↓19.5)	19.5 (↓74.7)	64.9 (↓34.3)	95.8 (↓3.0)	48.2 (↓50.7)	91.5 (↓8.0)	24.0 (↓75.5)
Doubao-Seed-2.0-Lite	63.0 (↓33.6)	82.9 (↓14.9)	77.9 (↓19.8)	76.5 (↓18.9)	14.7 (↓80.6)	66.3 (↓33.1)	82.2 (↓17.0)	62.6 (↓36.6)	91.2 (↓8.3)	29.1 (↓70.4)
MiniCPM-o 4.5	42.6 (↓50.5)	40.0 (↓53.8)	35.8 (↓57.9)	78.9 (↓13.4)	15.5 (↓76.8)	47.8 (↓49.8)	64.6 (↓32.6)	23.7 (↓73.6)	89.0 (↓9.0)	13.8 (↓84.2)
MiMo-V2.5	47.3 (↓44.7)	59.7 (↓34.1)	48.5 (↓45.3)	65.6 (↓24.6)	15.4 (↓74.9)	53.2 (↓44.2)	71.7 (↓25.6)	31.8 (↓65.4)	84.8 (↓12.6)	24.3 (↓73.1)
Nemotron-3-Nano-Omni	40.5 (↓50.8)	34.7 (↓55.8)	25.8 (↓64.8)	81.8 (↓10.2)	19.7 (↓72.3)	58.6 (↓39.0)	97.9 (↓0.0)	17.9 (↓80.0)	93.8 (↓3.7)	25.0 (↓72.5)
LVLM · Visual evidence										
Qwen3.7-Plus	81.4 (↓12.6)	89.4 (↓6.4)	64.1 (↓31.7)	92.7 (↑0.6)	79.2 (↓12.8)	83.7 (↓14.4)	95.2 (↓3.2)	54.8 (↓43.6)	97.4 (↓0.3)	87.2 (↓10.5)
Qwen3-VL	56.7 (↓34.5)	80.1 (↓12.2)	63.8 (↓28.5)	60.1 (↓30.0)	23.0 (↓67.1)	59.5 (↓38.1)	93.9 (↓2.6)	32.1 (↓64.4)	84.3 (↓14.4)	27.8 (↓70.9)
MiniCPM-V 4.5	64.8 (↓24.3)	79.5 (↓8.0)	28.8 (↓58.7)	82.4 (↓8.3)	68.4 (↓22.4)	64.6 (↓32.1)	93.3 (↓2.2)	8.0 (↓87.5)	91.1 (↓6.7)	65.8 (↓31.9)
InternVL3.5	31.8 (↓53.4)	8.7 (↓72.4)	7.7 (↓73.4)	73.8 (↓15.7)	37.1 (↓52.4)	58.3 (↓36.9)	92.9 (↓1.6)	12.2 (↓82.4)	88.5 (↓7.3)	39.6 (↓56.2)
LALM · Audio evidence										
StepAudio 3	85.2 (↓13.7)	98.7 (↓0.0)	99.0 (↑0.3)	91.0 (↓8.0)	51.9 (↓47.1)	81.4 (↓18.6)	100.0 (↓0.0)	100.0 (↓0.0)	97.1 (↓2.9)	28.5 (↓71.5)
Audio Flamingo 3	54.2 (↓41.7)	64.2 (↓33.5)	45.7 (↓52.1)	86.9 (↓7.1)	19.9 (↓74.0)	55.3 (↓44.6)	98.7 (↓1.0)	4.8 (↓94.9)	94.2 (↓5.8)	23.4 (↓76.6)
MiMo-Audio	37.4 (↓58.2)	50.8 (↓46.6)	42.2 (↓55.3)	52.2 (↓41.3)	4.2 (↓89.4)	34.6 (↓63.6)	58.8 (↓40.9)	18.8 (↓80.8)	53.8 (↓42.9)	7.1 (↓89.7)

4 Evaluating Selective History Use

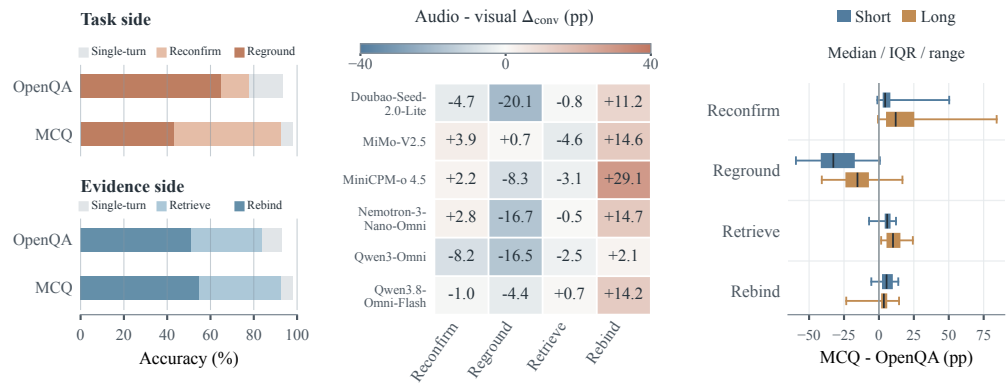
We examine whether strong single-turn performance translates into selective history use across omni-modal models, LVLMs, and LALMs. Table 2 reports all 13 models across conversation horizons and reference/conflict conditions; Figure 3 summarizes paired conditions, evidence modality, and answer interface.

4.1 Overall Results and the Access–Binding Hypothesis

Models answer the underlying questions reliably when the question and target media are supplied directly, yet struggle when the same targets must be resolved through history. Median model-level OpenQA accuracy falls from 93.7% in Single-turn to 72.3% in Multi-turn, with median $\Delta_{\text{conv}} = 20.1$ pp. This contrast exposes history-use difficulties beyond direct answerability.

Extended histories expose larger gaps. Overall Multi-turn accuracy is lower in Long than Short for 12 of 13 models on OpenQA and all 13 on MCQ. Median model-level Δ_{conv} rises from 6.8 to 34.5 pp on OpenQA and from 19.0 to 36.9 pp on MCQ. Short and Long differ in both history composition and extent, so this is a configuration comparison rather than an isolated length effect.

Conflict conditions expose weaknesses that reference conditions can conceal. In Long OpenQA, median accuracy falls from 79.5% in Reconfirm to 48.5% in Reground, and from 78.9% in Retrieve to 19.9% in Rebind. Across both sides, success with relevant history does not guarantee robustness when other contextual information conflicts with the target.



(a) Paired conditions (b) Evidence modality (c) Answer interface

Figure 3: **Paired conditions, modality, and answer interface.** (a) Mean accuracy across 13 models on complete operation pairs, pooling horizons; bars overlay rather than stack accuracies. (b) Audio minus visual Δ_{conv} for six omni models, averaging interfaces and pooling horizons; task subsets differ by modality. (c) MCQ minus OpenQA accuracy across 13 models by horizon; boxes, lines, and whiskers indicate the interquartile range, median, and minimum–maximum. Aggregation details appear in Appendix C.

We hypothesize that these failures reflect two complementary demands: accessing the required history and binding it to the current request.

(1) History access. The ability to recover the historical question or media content needed to answer the current request.

(2) Cross-turn binding. The ability to pair the question with its intended evidence across turns while resisting competing within-turn associations.

Figure 3a illustrates increasing demands along two parallel ladders: Single-turn–Reconfirm–Reground and Single-turn–Retrieve–Rebind. Single-turn establishes direct answerability; reference conditions introduce historical dependencies; conflict conditions add interference while preserving the target. The resulting drops motivate a distinction between accessing historical information and using it correctly: Reground must combine an earlier question with current evidence, whereas Rebind must combine the current question with earlier evidence. The ladders do not isolate the two abilities: reference conditions also require binding, and Reconfirm permits answer-reuse shortcuts. Section 5 therefore examines these demands with recall probes and targeted interventions.

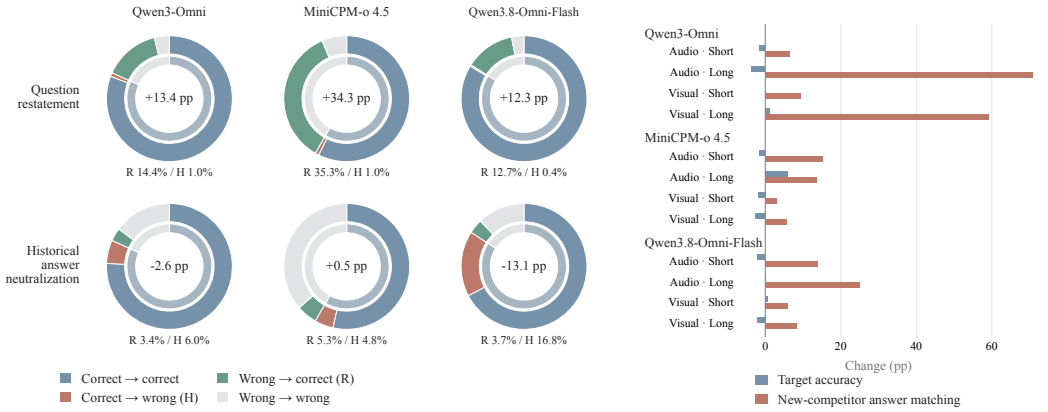
4.2 Modality and Interface Expose Distinct Weaknesses

Evidence modality. Modality sensitivity depends on the operation (Figure 3b). All six omni-modal models have a larger Rebind Δ_{conv} on audio than visual tasks, whereas five have a larger Reground Δ_{conv} on visual tasks. Thus an overall claim that one modality is harder would hide opposing patterns across historical dependencies. Preserving a historical evidence target and reapplying a historical question need not share the same modality sensitivity. This motivates examining the two history roles separately when diagnosing failures.

Answer interface. The median absolute OpenQA–MCQ multi-turn difference is only 2.5 pp overall, but 24.4 pp on Reground. Pooling horizons, eleven of 13 configurations score lower on Reground MCQ than OpenQA; Figure 3c separates the model-level distributions by horizon. Qwen3-Omni illustrates the cancellation: overall accuracies are 77.26% and 77.28%, while Reground scores are 81.76% and 57.76%. Providing answer candidates therefore does not uniformly simplify the conversational task. MCQ also differs in single-turn answerability, so an interface contrast is not itself a pure history effect. In task-carrying MCQ, however, the final request inherits both the question and the original option mapping. Thus even a small answer space may impose an additional history-use burden, motivating separate question and option restatement in Section 5.1.

Table 3: **Separating question and option inheritance.** Format-matched MCQ accuracy (%), $n = 128$ per row and condition. Gains (pp) are relative to Neither; omitted components remain in history.

Model	Operation	Accuracy (%)				Gain over Neither (pp)	
		Neither	Question	Options	Both	Question	Options
Qwen3-Omni	Reconfirm	97.7	94.5	99.2	100.0	-3.1	+1.6
	Reground	57.8	72.7	99.2	98.4	+14.8	**+41.4**
MiniCPM-o 4.5	Reconfirm	85.9	90.6	98.4	97.7	+4.7	+12.5
	Reground	29.7	42.2	92.2	96.1	+12.5	**+62.5**



(a) Task-side interventions

(b) Evidence-side interventions

Figure 4: **Task- and evidence-side interventions (OpenQA).** (a) Reground: inner/outer rings show correctness before/after intervention; centers show accuracy changes (pp), R/H repaired/harmed percentages. (b) Rebind: media-replacement changes in target accuracy and new-competitor answer matching. Protocols and sample counts: Appendix E.

5 Diagnosing Selective History Failures

We examine whether conversational failures reflect difficulty accessing required history or applying it to the intended task and evidence. We probe historical information and intervene separately on task inheritance and evidence selection, following the hypotheses in Section 4.1.

5.1 Task-Side: Access and Application

Recalling a question versus applying it. Qwen3-Omni and Qwen3.8-Omni-Flash recall the inherited question on 96.8% and 99.8% of Reground OpenQA probes, yet achieve only 81.8% and 84.1% accuracy on the conversational task. Question restatement raises these scores to 95.2% and 96.4%. MiniCPM-o 4.5 has weaker recall (60.5%) and a larger restatement gain (Figure 4a, top). Across these separate probes, high explicit recall coexists with weaker task application, motivating interventions that make the inherited request explicit.

Repair versus disruption. Question restatement repairs 12.7–35.3% of valid Reground OpenQA pairs while harming at most 1.0%. Historical-answer neutralization instead yields inconsistent gains, reducing accuracy for two of the three models (Figure 4a, bottom). Neutralizing an earlier answer therefore does not reproduce the benefit of making the inherited task explicit, limiting a stale-answer-only explanation.

Separating question and option inheritance. MCQ additionally requires inheriting the answer interface. Options-only restatement improves Reground by 41.4/62.5 pp for Qwen3-Omni/MiniCPM-o 4.5, versus 14.8/12.5 pp for questions alone (Table 3); Reconfirm gains are smaller. Restoring both components adds no gain over options alone for Qwen3-Omni, but further improves MiniCPM-o

Table 4: **Two routes to mitigation.** Accuracy (%). (a) Dev-Val with cached histories and matched execution ($n = 500$ per model/interface). (b) Qwen3-Omni on Test with checkpoint-native histories ($n = 5{,}000$ per interface); Single/Multi denote single-/multi-turn evaluation. Compare methods within each panel.

Method	Qwen3-Omni		MiniCPM-o 4.5	
	OpenQA	MCQ	OpenQA	MCQ
Neutral	71.4	78.8	59.8	59.6
Restatement	76.0	**86.4**	74.2	**78.2**
Reweighting	73.2	77.6	61.4	60.2
Routed policy	**78.2**	86.0	**77.8**	**78.2**

(a) Inference-time assistance

Method	OpenQA		MCQ	
	Single	Multi	Single	Multi
Base	94.60	77.26	99.12	77.28
SFT	94.84	80.68	99.12	79.64
SFT+Inv	94.92	80.70	99.12	79.84

(b) Conversational adaptation

4.5. Task inheritance therefore extends beyond question recall to the answer interface, whose options jointly supply candidate answers and their letter mapping.

5.2 Evidence-Side: Access and Selection

Accessing historical media. A distinct source-fact question probes information about historical media. OpenQA accuracy reaches 66.9–85.7%, versus 90.4–92.1% with direct input (Appendix Table 13). These separate source-fact probes indicate retained historical information, potentially including textual cues, and complement the original-task evidence-selection analysis below.

Changing the competing evidence. Rebind supplies the complete question. Removing current media improves OpenQA by 39.5/52.9 pp for Qwen3-Omni/MiniCPM-o 4.5, but also reduces input load. We instead replace the current competitor while fixing the historical target, measuring agreement with the *same replacement-media answer* before and after the swap. Qwen3-Omni's target accuracy barely changes (52.6% to 51.7%), while this matching rate rises from 0.1% to 36.5%.

Stable accuracy can conceal answer shifts. Across 12 configurations, target accuracy changes by -3.6 to $+6.0$ pp, yet replacement-answer matching increases throughout (Figure 4b). For Qwen3-Omni, this increase is larger in Long (59.1–71.0 pp) than Short (6.5–9.5 pp). An audit restricted to cases where all three media endpoints are answerable alone retains this behavior (Appendix Table 17). This consistent shift under whole-media replacement supports sensitivity to competing content. Together with the task-side results, it motivates assistance that restores inherited requests and strengthens evidence selection.

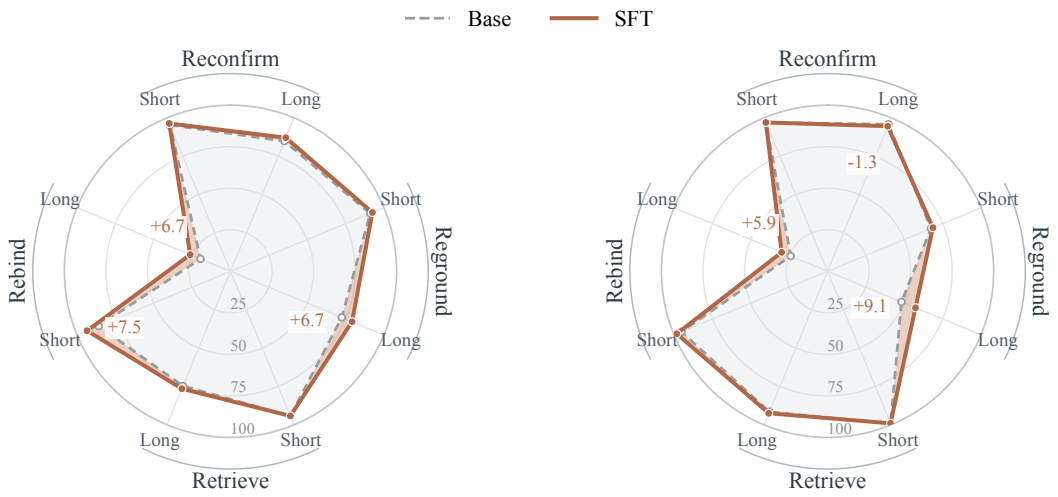
6 Mitigating Selective History Failures

We test whether the diagnosed failures can be reduced through inference-time assistance or conversational adaptation (Table 4). We evaluate each approach against matched controls within its respective protocol.

6.1 Assisting Selection at Inference Time

Restatement supplies the inherited question and, for MCQ, its original options. *Reweighting* resolves the user-visible media reference and scales attention to non-target media and historical assistant-text keys by 0.5, with renormalization at final-answer positions. A *routed policy* uses restatement for historical-question requests and reweighting for requests containing a complete question. It uses no gold answer or operation label, but relies on the benchmark's explicit reference structure (Appendix G).

Reweighting yields small, mixed Dev-Val gains: $+1.8/-1.2$ pp on OpenQA/MCQ for Qwen3-Omni and $+1.6/+0.6$ pp for MiniCPM-o 4.5. Routing improves over neutral execution, but its advantage over restatement is mixed across models and interfaces (Table 4(a)). Assistance therefore remains partial.



(a) OpenQA (b) MCQ

Figure 5: **Capability profiles before and after SFT.** Qwen3-Omni Test multi-turn accuracy (%), pooled across modalities; all axes span 0–100. Shading highlights changes between contours; selected gains and losses are labeled in pp. Training uses Short OpenQA only.

6.2 Learning from Native Conversational Histories

We adapt Qwen3-Omni on 1,018 Short OpenQA views (509 Dev-Train pairs). SFT+Inv adds a Jensen–Shannon penalty on answer-preserving pairs. Initialization, data order, 204 updates, and auxiliary-forward budgets are matched; each checkpoint generates its own test history (Appendix F).

SFT improves multi-turn accuracy by 3.42 pp on OpenQA and 2.36 pp on MCQ while retaining single-turn accuracy (Table 4(b)). Long OpenQA improves by 4.36 pp despite Short-only training. Figure 5 shows improvements on Rebind across both horizons and interfaces, and larger Reground gains on Long conversations. Excluding all input-parent overlap with gradient-training media retains a 3.14 pp OpenQA gain ($n = 3,125$; Appendix F). The improvement thus extends beyond tasks sharing these training sources.

Residual errors concentrate in difficult conditions: Long Rebind reaches 26.2% OpenQA and 29.9% MCQ accuracy after SFT, while Long MCQ Reconfirm declines by 1.3 pp. The OpenQA net gain combines 261 repaired answers with 90 newly incorrect answers (Appendix Table 22). These uneven changes identify evidence selection under extended histories as a priority for further improvement.

SFT+Inv adds only 0.02 pp on OpenQA and 0.20 pp on MCQ over matched SFT, with derivation-root sensitivity intervals spanning zero. In this single-model, single-seed study, the observed gains are attributable to SFT, with no clear additional benefit from the consistency term.

Scope of the evidence. The experiments establish controlled within-benchmark comparisons. Appendix H discusses explicit references, source reuse, and automated verification, and outlines extensions to broader histories and replicated adaptation.

7 Conclusion

ReTurn evaluates selective history use through controlled contrasts that preserve historical dependencies while varying agreement or competition. Its paired design controls target question difficulty when comparing historical agreement and competing evidence. Across omni-modal models, LVLMs, and LALMs, strong single-turn answerability does not ensure correct task inheritance for current evidence or evidence selection from history. Diagnostics and partial mitigation highlight the need to preserve useful history while resisting contextual interference.

AI USE STATEMENT

Generative AI tools assisted task construction, automated verification, and response scoring, as documented in the main text and appendix. They also supported literature retrieval, feedback on experimental designs and analyses, coding, and manuscript drafting and editing. The source audiovisual media were drawn from existing datasets, not generated by AI. The authors take responsibility for all research decisions and the final content.

ETHICS STATEMENT

ReTurn uses audiovisual media from existing research datasets to evaluate model behavior in controlled conversations. Any release of source media will be subject to the applicable licenses and access restrictions. Automated task verification and scoring may retain annotation errors or biases inherited from the source data and models. Benchmark results concern model behavior and should not be interpreted as assessments of individuals depicted in the media.

REPRODUCIBILITY STATEMENT

Section 3.2 and Appendix A describe task construction, data provenance, and partitioning. Appendices B–G document model configurations, scoring, diagnostic interventions, and adaptation protocols, including the distinction between native-history and cached-history evaluation. We plan to release the ReTurn benchmark after the review process concludes, subject to the licenses and access restrictions of the source datasets.

REFERENCES

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *CoRR*, abs/2604.27393, 2026. doi: 10.48550/ARXIV.2604.27393. URL https://doi.org/10.48550/arXiv.2604.27393.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. FineVideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/3babb6b453cb59d87cb58a1219ef914b-Abstract-Conference.html.

Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and Yunzhong He. Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 35740–35770. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1654. URL https://doi.org/10.18653/v1/2026.acl-long.1654.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026.

Young-Jun Lee, Byung-Kwan Lee, Jianshu Zhang, Yechan Hwang, Byungsoo Ko, Han-Gyu Kim, Dongyu Yao, Xuankun Rong, Eojin Joo, Seung-Ho Han, Bowon Ko, and Ho-Jin Choi. Multiverse: A multi-turn conversation benchmark for evaluating large vision and language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 708–719. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00074. URL https://doi.org/10.1109/ICCV51701.2025.00074.

Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, and Kaipeng Zhang. Convbench: A multi-turn conversation evaluation benchmark with hierarchical ablation capability for large vision-language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/b69396afc07a9ca3428d194f4db84c02-Abstract-Datasets_and_Benchmarks_Track.html.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. MMDU: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/1057053100de064a44286239724f7865-Abstract-Datasets_and_Benchmarks_Track.html.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.747. URL https://doi.org/10.18653/v1/2024.acl-long.747.

OpenAI. GPT-5.6 System Card. OpenAI Deployment Safety Hub, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6. Published July 9, 2026.

Yaning Pan, Qianqian Xie, Guohui Zhang, Zekun Moore Wang, Yongqian Wen, Yuanxing Zhang, Haoxuan Hu, Zhiyu Pan, Yibing Huang, Zhidong Gan, Yonghong Lin, An Ping, Shihao Li, Yanghai Wang, Tianhao Peng, and Jiaheng Liu. Mt-video-bench: A holistic video understanding benchmark for evaluating multimodal llms in multi-turn dialogues. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 8105–8126. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.397. URL https://doi.org/10.18653/v1/2026.findings-acl.397.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Qwen Team. Qwen3-omni technical report. *CoRR*, abs/2509.17765, 2025a. doi: 10.48550/ARXIV.2509.17765. URL https://doi.org/10.48550/arXiv.2509.17765.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025b. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, JingJing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025a. doi: 10.48550/ARXIV.2508.18265. URL https://doi.org/10.48550/arXiv.2508.18265.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 18925–18935. Computer Vision Foundation / IEEE, 2025b. doi: 10.1109/CVPR52734.2025.01763. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_OmniMMI_A_Comprehensive_Multi-modal_Interaction_Benchmark_in_Streaming_Video_Contexts_CVPR_2025_paper.html.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

LLM-Core Xiaomi. Mimo-audio: Audio language models are few-shot learners. *CoRR*, abs/2512.23808, 2025. doi: 10.48550/ARXIV.2512.23808. URL https://doi.org/10.48550/arXiv.2512.23808.

Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, Yutong Xie, Imran Razzak, Zongyuan Ge, Jionglong Su, Junjun He, and Yu Qiao. MMRC: A large-scale benchmark for understanding multimodal large language model in real-world conversation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *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*, pp. 22477–22503. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1096. URL https://doi.org/10.18653/v1/2025.acl-long.1096.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. URL https://arxiv.org/abs/2410.02713.

A CONSTRUCTION, PROVENANCE, AND EXPOSURE

Evidence before dialogue construction. ReTurn uses traceable clips from FineVideo and LLaVA-Video-178K, including previously audited assets with retained provenance. Source annotations and captions support localization; they are not themselves the answer evidence. Newly introduced clips or changed evidence requirements are checked against actual audio/video with Doubao-Seed-2.0-Lite (Seed, 2026). GPT-5.6 Sol (`gpt-5.6-sol`) (OpenAI, 2026) assists question analysis, task-consistency checks, and ambiguous cases, with visual-frame inspection where needed. Text-only review does not substitute for audio verification. Each base question records its parent video, temporal window, required modality, answer granularity, and supporting evidence. Verification outcomes are reused only for identical inputs and evidence requirements. Verification is automated; human annotation agreement was not measured.

Two target-preserving constructions. Reconfirm and Reground share the current target media, initial complete question, final follow-up, gold answer, options and order, and intervening user media and questions. The historical media supports either the same or a different answer to that question. The model produces its own historical replies, so changing historical media can also change subsequent assistant text. Retrieve and Rebind instead share the historical target, its position and preceding prefix, final complete question, gold answer, and options; their current non-target media provides weaker or stronger answer competition. Both conditions contain current media. The historical question is constructed separately from the final question to reduce direct answer copying. These are two local matched constructions, not a complete factorial crossing of independent task and evidence factors. Correctness on the agreeing Reconfirm control alone does not establish that the model re-inspected current media: the historical answer can already match the target.

Instantiation and partitioning. The benchmark contains 7,000 base tasks in 3,500 matched pairs: 1,750 tasks per operation, 3,500 per required modality, and 3,500 per horizon. Short and Long contain two and five media-bearing user turns, respectively, including the final turn. Long task-carrying histories insert three intervening media question/answer turns; evidence-carrying targets can occupy different historical positions. Short and Long are not universally item-matched. Each task has OpenQA and four-option MCQ views, yielding 14,000 task/interface views rather than two independent datasets. Modality labels identify necessary answer evidence, not necessarily the modalities presented to an omni-modal model.

Development calibration uses Qwen3-Omni single-turn and native multi-turn runs, supplemented by selective checks on other models, to guide refinement and expansion toward high single-turn answerability and measurable history effects. Small development batches establish questions, gold answers, pairing constraints, and scoring granularity; low single-turn performance prompts checks of evidence, wording, input consumption, and truncation. The approximate 80% single-turn target was a development-set observation goal, not an item-level acceptance threshold. Neither universal single-turn correctness nor large degradation in every condition was required. The final split uses metadata grouping for matched pairs, recorded fact/derivation roots, and shared requests, without reading final model accuracies: 2,000 Dev tasks, comprising 1,500 Dev-Train and 500 Dev-Val tasks, plus 5,000 Test tasks. Different facts can still share parent media. Of the 5,000 Test tasks, 1,028 previously belonged to TRAIN/DEV partitions; 2,268 carry broader component-level development tags, with 536 tasks in both groups. These indicators describe different exposure histories, not gradient-training counts. No exact adaptation task ID occurs in Test, but repartitioning does not remove prior research exposure.

Source exposure. Table 5 audits media actually sent to the model, excluding unused partner/donor metadata. Gradient exposure refers to the 750 unique task IDs used for adaptation (Appendix F), rather than the entire Dev-Train partition. Indicators overlap and count test tasks rather than unique parents. Target-parent disjointness concerns target media; other inputs may still share development sources. Shared media and prior development exposure also preclude interpreting task-level sample size as the number of independent source observations. Requiring every input parent to be disjoint from Dev retains 1,928 tasks but removes all Long Retrieve/Rebind tasks. Source-filtered aggregates describe performance under different exposure and task-composition profiles.

Table 5: **Actual-input source and development exposure.** Counts are test tasks satisfying each indicator, not numbers of distinct videos; indicators overlap.

Exposure indicator	Tasks / 5,000
Any input parent occurs in development	3,072
Any input parent occurs in gradient-used data	1,875
An exact input window occurs in development	2,947
An exact input window occurs in gradient-used data	1,571
Task has an earlier training/development tag	2,268
Target parent occurs in development	918
Target parent occurs in gradient-used data	486

B Models and Evaluation Protocol

Table 6: **Evaluated model configurations.** Display names are used throughout the paper; configuration identifiers retain size and variant suffixes. A/V denote the supported evidence subsets. Limits are configured output-token limits; arrows indicate a length-retry cap where recorded. Sampling and media parameters below are configuration-specific.

Display name	Scope; limit	Execution and recorded settings
Qwen3.8-Omni-Flash	A/V; 1,024 → 4,096	Hosted `qwen3.8-omni-flash`; temperature 0, reasoning effort none, 2 fps; companion WAV and silent video.
Qwen3-Omni	A/V; 128	Checkpoint: `Qwen3-Omni-30B-A3B-Instruct`. Local native runner; frozen formal decoding configuration.
Doubao-Seed-2.0-Lite	A/V; 1,024 → 4,096	Hosted `doubao-seed-2-0-lite-260428`; temperature 0, 2 fps; audio, silent video, and text in order.
MiniCPM-o 4.5	A/V; 256	Recorded runner key: `minicpmo_45_instruct`. Local native runner; frozen formal decoding configuration.
MiMo-V2.5	A/V; 4,096	Recorded runner key: `mimo_v2_5`. Local SGLang, tensor parallelism 4; configured adaptive cap 8,192.
Nemotron-3-Nano-Omni	A/V; 1,024	Model: `Nemotron-3-Nano-Omni-30B`. Local SGLang single-turn and vLLM multi-turn execution.
Qwen3.7-Plus	V; 1,024	Hosted `qwen3.7-plus`; temperature 0, 2 fps, video only, without added clip labels.
Qwen3-VL	V; 256 → 1,024	Checkpoint: `Qwen/Qwen3-VL-30B-A3B-Instruct`. Official local Hugging Face implementation; bfloat16, SDPA.
MiniCPM-V 4.5	V; 256 → 1,024	Checkpoint: `openbmb/MiniCPM-V-4_5`. Official local implementation; bfloat16, SDPA.
InternVL3.5	V; 256 → 1,024	Checkpoint: `OpenGVLab/InternVL3_5-30B-A3B`. Official remote-code implementation; bfloat16, FlashAttention 2 when available.
StepAudio 3	A; 1,024	Hosted `stepaudio-3-chat-preview`; temperature 0; original companion WAV.
Audio Flamingo 3	A; 256 → 1,024	Recorded runner key: `audio_flamingo3_hf`. Official local implementation; SDPA and bfloat16 autocast.
MiMo-Audio	A; 256	Model: `MiMo-Audio-7B-Instruct`. Local implementation; original companion WAV and frozen turn boundaries.

Native histories and media scope. Each model generates its own assistant replies recursively from the chronological prefix, separately for OpenQA and MCQ. Earlier errors remain in the history; gold replies are not substituted. Exact successful requests may be reused, but responses to a different wording or prefix are not interchangeable. LVLMs consume visual-only media and LALMs consume the original companion audio without an ASR substitution. The two hosted omni-modal con-

figurations explicitly use full synchronized audio/video for both visual-evidence and audio-evidence tasks. This input policy does not make every task a joint audio-visual reasoning task.

Runtime details. Table 6 records the evaluated configurations. For the visual-only local models, the shared silent-video cache samples at 2 fps from frame zero with maximum dimension 768; native temporal padding is permitted. Qwen3-VL, MiniCPM-V 4.5, InternVL3.5, and Audio Flamingo 3 use greedy generation with a 256-token limit and a 1,024-token length-retry limit. Their batch sizes are 2, 4, 1, and 8. Qwen3-VL and Audio Flamingo 3 use Transformers 5.12.1; MiniCPM-V 4.5 and InternVL3.5 use 4.51.2, all with PyTorch 2.12.1. MiniCPM-V 4.5 disables image IDs and thinking and uses one slice with temporal packing 1. Audio Flamingo 3 retains the official FP32 positional embeddings under bfloat16 autocast. MiMo-Audio uses Python 3.10.20, PyTorch 2.6.0, Transformers 4.49.0, and FlashAttention 2.7.4.post1. Local omni-modal models retain their recorded decoding configurations; hosted-model temperature settings do not apply to these runs. Nemotron-3-Nano-Omni uses SGLang for single-turn and vLLM for multi-turn execution; its measured gap may therefore include backend effects. Cross-model comparisons also reflect model-specific input processing and decoding settings.

Single-turn reference and valid responses. The single-turn input supplies the target media and complete question: current media for task-carrying history, historical target media for evidence-carrying history. It is an empirical answerability reference, not a mathematical ceiling or simply the multi-turn prompt with history deleted. Scores use task views with valid responses in both conditions. The 95,000 expected model/task/interface views yield 94,984 valid pairs: Qwen3.8-Omni-Flash excludes 9, InternVL3.5 excludes 1, and StepAudio 3 excludes 6. Other model views are complete. The precise affected operation/horizon cells are listed beneath the full result tables.

C Full Interface-Specific Results and Aggregation

Tables 7–10 report single-turn (ST) and multi-turn (MT) accuracy on n common valid response pairs per cell. History harm counts ST-correct/MT-incorrect pairs; rescue counts the reverse, giving $\Delta_{\text{conv}} = 100(\text{harm} - \text{rescue})/n$. Intervention rescue/harm instead compares two multi-turn conditions.

Table 7: **OpenQA, Short: full Test results.** Single-turn (ST) and multi-turn (MT) accuracy (%); $\Delta_{\text{conv}} = \text{ST} - \text{MT}$ (pp).

Model	Task-carrying history						Evidence-carrying history					
	Reconfirm			Reground			Retrieve			Rebind		
	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}
Omni-modal												
Qwen3.8-Omni-Flash	96.2	94.4	1.8	96.2	95.0	1.1	93.0	94.4	-1.4	92.9	80.6	12.3
Qwen3-Omni	95.0	95.4	-0.3	95.0	91.0	4.0	94.2	94.2	0.0	94.2	85.9	8.3
Doubao-Seed-2.0-Lite	97.4	95.4	2.1	97.4	93.1	4.3	95.4	94.7	0.6	95.4	75.8	19.5
MiniCPM-o 4.5	93.9	88.5	5.4	93.9	80.8	13.1	92.3	91.0	1.3	92.3	54.4	37.9
MiMo-V2.5	92.6	90.1	2.6	92.6	79.5	13.1	90.2	87.7	2.6	90.2	63.5	26.7
Nemotron-3-Nano-Omni	90.1	92.5	-2.4	90.1	75.4	14.7	92.0	93.4	-1.4	92.0	82.2	9.8
LVLM												
Qwen3.7-Plus	94.9	94.6	0.3	94.9	92.0	2.9	92.0	93.3	-1.3	92.0	90.4	1.6
Qwen3-VL	92.0	90.7	1.3	92.0	87.8	4.2	90.1	90.7	-0.6	90.1	82.4	7.7
MiniCPM-V 4.5	87.5	88.8	-1.3	87.5	60.9	26.6	90.7	91.7	-1.0	90.7	87.9	2.9
InternVL3.5	80.8	42.9	37.8	80.8	29.8	51.0	89.5	87.5	1.9	89.4	65.1	24.4
LALM												
StepAudio 3	98.4	99.0	-0.6	98.4	99.4	-1.0	99.0	87.5	11.6	99.0	34.1	65.0
Audio Flamingo 3	97.8	95.5	2.2	97.8	47.3	50.5	93.9	93.3	0.6	93.9	54.5	39.4
MiMo-Audio	96.2	86.3	9.9	96.2	72.5	23.6	93.6	73.1	20.5	93.6	18.9	74.7

Valid n per operation: omni-modal 625; LVLM 312/313 and LALM 313/312 for task-/evidence-carrying history. Exceptions: InternVL3.5 Rebind 312; Qwen3.8-Omni-Flash Rebind 624; StepAudio 3 Rebind 311; StepAudio 3 Retrieve 311.

Table 8: **OpenQA, Long: full Test results.** Single-turn (ST) and multi-turn (MT) accuracy (%); $\Delta_{\text{conv}} = \text{ST} - \text{MT}$ (pp).

Model	Task-carrying history						Evidence-carrying history					
	Reconfirm			Reground			Retrieve			Rebind		
	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}
Omni-modal												
Qwen3.8-Omni-Flash	96.9	91.3	5.6	97.0	73.1	23.9	93.0	93.8	-0.8	93.0	74.6	18.4
Qwen3-Omni	94.9	84.8	10.1	94.9	72.5	22.4	94.2	74.7	19.5	94.2	19.5	74.7
Doubao-Seed-2.0-Lite	97.8	82.9	14.9	97.8	77.9	19.8	95.4	76.5	18.9	95.4	14.7	80.6
MiniCPM-o 4.5	93.8	40.0	53.8	93.8	35.8	57.9	92.3	78.9	13.4	92.3	15.5	76.8
MiMo-V2.5	93.8	59.7	34.1	93.8	48.5	45.3	90.2	65.6	24.6	90.2	15.4	74.9
Nemotron-3-Nano-Omni	90.6	34.7	55.8	90.6	25.8	64.8	92.0	81.8	10.2	92.0	19.7	72.3
LVLM												
Qwen3.7-Plus	95.8	89.4	6.4	95.8	64.1	31.7	92.0	92.7	-0.6	92.0	79.2	12.8
Qwen3-VL	92.3	80.1	12.2	92.3	63.8	28.5	90.1	60.1	30.0	90.1	23.0	67.1
MiniCPM-V 4.5	87.5	79.5	8.0	87.5	28.8	58.7	90.7	82.4	8.3	90.7	68.4	22.4
InternVL3.5	81.1	8.7	72.4	81.1	7.7	73.4	89.5	73.8	15.7	89.5	37.1	52.4
LALM												
StepAudio 3	98.7	98.7	0.0	98.7	99.0	-0.3	99.0	91.0	8.0	99.0	51.9	47.1
Audio Flamingo 3	97.8	64.2	33.5	97.8	45.7	52.1	93.9	86.9	7.1	93.9	19.9	74.0
MiMo-Audio	97.4	50.8	46.6	97.4	42.2	55.3	93.6	52.2	41.3	93.6	4.2	89.4

Valid n per operation: omni-modal 625; LVLM 312/313 and LALM 313/312 for task-/evidence-carrying history. Exceptions: Qwen3.8-Omni-Flash Reconfirm 621; Qwen3.8-Omni-Flash Reground 624.

Table 9: **MCQ, Short: full Test results.** Single-turn (ST) and multi-turn (MT) accuracy (%); $\Delta_{\text{conv}} = \text{ST} - \text{MT}$ (pp).

Model	Task-carrying history						Evidence-carrying history					
	Reconfirm			Reground			Retrieve			Rebind		
	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}
Omni-modal												
Qwen3.8-Omni-Flash	98.6	98.6	0.0	98.6	93.9	4.6	99.0	99.0	0.0	99.0	81.9	17.1
Qwen3-Omni	98.6	97.0	1.6	98.6	67.4	31.2	99.5	99.2	0.3	99.5	95.2	4.3
Doubao-Seed-2.0-Lite	99.4	98.7	0.6	99.4	71.8	27.5	99.5	98.9	0.6	99.5	86.7	12.8
MiniCPM-o 4.5	97.3	96.0	1.3	97.3	32.3	65.0	97.9	96.5	1.4	97.9	49.0	49.0
MiMo-V2.5	97.8	89.0	8.8	97.8	46.9	50.9	97.4	96.3	1.1	97.4	71.4	26.1
Nemotron-3-Nano-Omni	97.9	99.0	-1.1	97.9	34.4	63.5	97.4	99.4	-1.9	97.4	82.1	15.4
LVLM												
Qwen3.7-Plus	98.7	98.4	0.3	98.7	80.1	18.6	97.8	98.1	-0.3	97.8	97.4	0.3
Qwen3-VL	97.1	98.7	-1.6	97.1	32.1	65.1	98.7	98.4	0.3	98.7	92.0	6.7
MiniCPM-V 4.5	96.5	94.2	2.2	96.5	20.8	75.6	97.8	97.8	0.0	97.8	92.7	5.1
InternVL3.5	94.2	93.3	1.0	94.2	11.9	82.4	95.8	97.8	-1.9	95.8	78.9	16.9
LALM												
StepAudio 3	100.0	100.0	0.0	100.0	100.0	0.0	100.0	99.7	0.3	100.0	39.5	60.5
Audio Flamingo 3	99.7	100.0	-0.3	99.7	6.4	93.3	100.0	99.4	0.6	100.0	57.7	42.3
MiMo-Audio	98.7	97.4	1.3	98.7	13.1	85.6	96.8	66.0	30.8	96.8	21.8	75.0

Valid n per operation: omni-modal 625; LVLM 312/313 and LALM 313/312 for task-/evidence-carrying history. Exceptions: StepAudio 3 Rebind 311; StepAudio 3 Reconfirm 312; StepAudio 3 Reground 312; StepAudio 3 Retrieve 311.

Table 10: **MCQ, Long: full Test results.** Single-turn (ST) and multi-turn (MT) accuracy (%); $\Delta_{\text{conv}} = \text{ST} - \text{MT}$ (pp).

Model	Task-carrying history						Evidence-carrying history					
	Reconfirm			Reground			Retrieve			Rebind		
	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}
Omni-modal												
Qwen3.8-Omni-Flash	98.6	95.3	3.2	98.6	89.9	8.7	99.0	98.9	0.2	99.0	76.8	22.2
Qwen3-Omni	98.9	95.8	3.0	98.9	48.2	50.7	99.5	91.5	8.0	99.5	24.0	75.5
Doubao-Seed-2.0-Lite	99.2	82.2	17.0	99.2	62.6	36.6	99.5	91.2	8.3	99.5	29.1	70.4
MiniCPM-o 4.5	97.3	64.6	32.6	97.3	23.7	73.6	97.9	89.0	9.0	97.9	13.8	84.2
MiMo-V2.5	97.3	71.7	25.6	97.3	31.8	65.4	97.4	84.8	12.6	97.4	24.3	73.1
Nemotron-3-Nano-Omni	97.9	97.9	0.0	97.9	17.9	80.0	97.4	93.8	3.7	97.4	25.0	72.5
LVLM												
Qwen3.7-Plus	98.4	95.2	3.2	98.4	54.8	43.6	97.8	97.4	0.3	97.8	87.2	10.5
Qwen3-VL	96.5	93.9	2.6	96.5	32.1	64.4	98.7	84.3	14.4	98.7	27.8	70.9
MiniCPM-V 4.5	95.5	93.3	2.2	95.5	8.0	87.5	97.8	91.1	6.7	97.8	65.8	31.9
InternVL3.5	94.6	92.9	1.6	94.6	12.2	82.4	95.8	88.5	7.3	95.8	39.6	56.2
LALM												
StepAudio 3	100.0	100.0	0.0	100.0	100.0	0.0	100.0	97.1	2.9	100.0	28.5	71.5
Audio Flamingo 3	99.7	98.7	1.0	99.7	4.8	94.9	100.0	94.2	5.8	100.0	23.4	76.6
MiMo-Audio	99.7	58.8	40.9	99.7	18.8	80.8	96.8	53.8	42.9	96.8	7.1	89.7

Valid n per operation: omni-modal 625; LVLM 312/313 and LALM 313/312 for task-/evidence-carrying history. Exceptions: Qwen3.8-Omni-Flash Reconfirm 623; Qwen3.8-Omni-Flash Reground 624.

Counts are pooled over both modalities for omni-modal models, visual tasks for LVLMs, and audio tasks for LALMs. The main table separates interfaces; analyses that average them use $(s_{\text{OpenQA}} + s_{\text{MCQ}})/2$, with each rate computed on its own valid denominator. Figure 3a pools horizons within models and weights the 13 models equally. Its operation-pair contrasts use the common intersection of valid ST/MT responses for both members, which cannot be recovered from marginal accuracies alone.

D Scoring Rules and Automated Sensitivity Audits

Frozen rule and content scores. The initial scorer applies frozen answer and interface rules, including approved normalization/aliases where applicable. Rule-rejected answers receive blind GPT-5.6 Sol review using the complete question, gold answer, options, and requested answer granularity; model identity, condition, and aggregate performance are hidden. Content scores incorporate confirmed semantic correctness, while uncertain judgments receive no confirmed-correct credit. Rule scores, content scores, format compliance, uncertainty, and review reasons are retained separately. The gold answer and aliases are not modified in response to final test errors. Table 11 reports Δ_{conv} under both scoring levels. The production rubric accepts semantic equivalence at the granularity requested by the question; unrequested modifiers, such as finer color shades, are not mandatory solely because they appear in the reference. A text judge assesses answer equivalence, not the validity of the original audiovisual evidence.

Cross-judge and rubric sensitivity. We sample 80 logical OpenQA outputs in each of six model/condition groups, stratified by rule-correct, review-recovered, still-incorrect, and uncertain outcomes. Deduplication yields 367 distinct Doubao-Seed-2.0-Lite judge inputs for 480 logical outputs. The judge does not see the evaluated model, condition, or original score. Table 12 reports sampling-weighted accuracy estimates and unweighted disagreements. The original/alternative weighted Δ_{conv} estimates are 17.34/10.82 pp for Qwen3-Omni, 32.46/29.23 pp for MiniCPM-o 4.5, and 7.62/5.61 pp for Qwen3.8-Omni-Flash.

The audit changes both judge and rubric: its simplified prompt did not fully reproduce the production rules for semantic equivalence and requested answer granularity. All 56 logical disagreements (50 distinct inputs) were originally correct and rejected by the alternative judge. Some involve synonyms, basic color categories, or more specific answers; text-only disagreements do not establish

Table 11: **Frozen rule and reviewed content scoring.** Both columns use the same valid single-turn/multi-turn response pairs. Conversational degradation Δ_{conv} is in pp; a rule score is not necessarily an exact-string score.

Model	OpenQA			MCQ		
	n	Rule Δ_{conv}	Content Δ_{conv}	n	Rule Δ_{conv}	Content Δ_{conv}
Qwen3.8-Omni-Flash	4994	9.5	7.6	4997	7.7	7.0
Qwen3-Omni	5000	13.9	17.3	5000	21.9	21.8
Doubao-Seed-2.0-Lite	5000	19.2	20.1	5000	10.2	21.7
MiniCPM-o 4.5	5000	-0.9	32.5	5000	42.5	39.5
MiMo-V2.5	5000	3.2	28.0	5000	29.3	33.0
Nemotron-3-Nano-Omni	5000	22.8	28.0	5000	25.0	29.0
Qwen3.7-Plus	2500	-2.1	6.7	2500	42.0	9.6
Qwen3-VL	2500	16.2	18.8	2500	0.0	27.8
MiniCPM-V 4.5	2500	-8.1	15.6	2500	44.2	26.4
InternVL3.5	2499	16.5	41.1	2500	30.9	30.7
StepAudio 3	2498	17.1	16.2	2496	17.1	16.9
Audio Flamingo 3	2500	20.3	32.4	2500	41.6	39.3
MiMo-Audio	2500	29.0	45.2	2500	55.9	55.9

Table 12: **Cross-judge/rubric OpenQA sensitivity audit.** Accuracy estimates (%) use stratified sampling weights. There are 480 logical outputs and 367 distinct judge inputs. Both judge and rubric differ; these estimates do not replace full-Test scores.

Model	Condition	n	Original	Alternative	Disagreements
Qwen3-Omni	ST	80	94.60	82.79	17
Qwen3-Omni	MT	80	77.26	71.96	9
MiniCPM-o 4.5	ST	80	93.08	83.50	2
MiniCPM-o 4.5	MT	80	60.62	54.28	3
Qwen3.8-Omni-Flash	ST	80	94.76	81.86	14
Qwen3.8-Omni-Flash	MT	80	87.14	76.25	11

media-grounded error. For MiniCPM-o single-turn, two sampled disagreements account for the 9.58 pp estimated accuracy reduction because each carries 4.79 pp of sampling weight. The single-group design standard error is 6.56 pp, excluding rubric uncertainty. These estimates quantify joint judge/rubric sensitivity rather than an adjudicated scoring-error rate. Shared judge inputs are not independent replications; no cross-condition confidence interval is inferred.

Media-grounded checks. An additional 32-pair/64-task sample contains 31 distinct necessary-evidence questions. Qwen3.8-Omni-Flash answers these from the complete media without seeing the gold; 30/31 agree with it. The remaining question asks for the first two spoken words: the gold is "what is," Qwen3.8-Omni-Flash responds "Most important," and an additional blind Doubao-Seed-2.0-Lite media check responds "What is." We retain this disagreement. The paired single-turn input/gold equality and the 64 tasks' option uniqueness and valid answer indices pass programmatic checks. This sample checks necessary evidence, not every competing endpoint, and does not estimate the gold-answer error rate of the full dataset. No human review is included in this audit; Doubao-Seed-2.0-Lite's construction role and Qwen3.8-Omni-Flash's evaluation role also mean that these model checks are not independent ground truth.

E Additional Diagnostic Protocols and Results

E.1 Task Recall, Restatement, and Historical Answer Text

The recall probe explicitly asks for the historical task, whereas restatement supplies that task at the final turn. Table 13 uses only Reground recall outcomes, preserving its operation-specific denominator. Qwen3-Omni and Qwen3.8-Omni-Flash attain high explicit recall despite lower original Reground accuracy, while MiniCPM-o 4.5 has substantial recall difficulty. The probes characterize

Table 13: **Task-recall and historical source-fact probes.** Entries are correct/valid responses. The recall column is restricted to Reground. Source-fact probes ask a distinct historical-evidence question; their 712-task cohort is not the recall or full evaluation cohort.

Model	Interface	Reground recall	Source-fact ST	Source-fact MT
Qwen3-Omni	OpenQA	1210/1250	656/712	532/712
Qwen3-Omni	MCQ	1226/1250	708/712	640/712
MiniCPM-o 4.5	OpenQA	756/1250	644/712	476/712
MiniCPM-o 4.5	MCQ	940/1250	692/712	582/712
Qwen3.8-Omni-Flash	OpenQA	1246/1248	652/712	610/712
Qwen3.8-Omni-Flash	MCQ	1239/1250	696/712	676/712

Table 14: **Component interventions on common valid test responses.** Native and changed accuracies are percentages. Rescue/harm counts compare the changed response with the native multi-turn response, not single-turn with multi-turn.

Reground: explicit question restatement						
Model	Interface	n	Native	Changed	Δ (pp)	Rescue/harm
Qwen3-Omni	OpenQA	1250	81.76	95.20	+13.44	180/12
Qwen3-Omni	MCQ	1250	57.76	98.80	+41.04	514/1
MiniCPM-o 4.5	OpenQA	1250	58.32	92.64	+34.32	441/12
MiniCPM-o 4.5	MCQ	1250	28.00	97.04	+69.04	867/4
Qwen3.8-Omni-Flash	OpenQA	1249	84.07	96.40	+12.33	159/5
Qwen3.8-Omni-Flash	MCQ	1249	91.91	98.72	+6.81	86/1
Reground: neutralize historical answer						
Model	Interface	n	Native	Changed	Δ (pp)	Rescue/harm
Qwen3-Omni	OpenQA	1250	81.76	79.12	-2.64	42/75
Qwen3-Omni	MCQ	1250	57.76	44.56	-13.20	21/186
MiniCPM-o 4.5	OpenQA	1250	58.32	58.80	+0.48	66/60
MiniCPM-o 4.5	MCQ	1250	28.00	27.84	-0.16	60/62
Qwen3.8-Omni-Flash	OpenQA	1245	84.10	71.00	-13.09	46/209
Qwen3.8-Omni-Flash	MCQ	1248	91.91	89.34	-2.56	21/53
Rebind: remove current media						
Model	Interface	n	Native	Changed	Δ (pp)	Rescue/harm
Qwen3-Omni	OpenQA	1250	52.72	92.24	+39.52	502/8
Qwen3-Omni	MCQ	1250	59.60	98.24	+38.64	484/1
MiniCPM-o 4.5	OpenQA	1250	34.96	87.84	+52.88	676/15
MiniCPM-o 4.5	MCQ	1250	31.36	95.20	+63.84	804/6
Qwen3.8-Omni-Flash	OpenQA	1249	77.58	93.76	+16.17	219/17
Qwen3.8-Omni-Flash	MCQ	1250	79.36	99.12	+19.76	251/4

elicited recall at the aggregate level; their separate requests do not identify recall within each original error.

Table 14 reports full-question restatement, historical-answer neutralization, and current-media removal on common valid responses. For MCQ, full restatement also supplies the original answer options. It therefore changes the information supplied at the final turn. Neutralizing historical answer text is not a reliable general repair: for example, Qwen3-Omni Reground decreases by 2.64 pp on OpenQA and 13.20 pp on MCQ. The contrast motivates examining task application alongside historical-answer persistence.

E.2 Separating Question and Option Restatement

We sample 128 Reconfirm/Reground pairs (256 tasks per model) with seed 2801, allocating 32 pairs to each modality/horizon cell. All arms retain the same media, model-native prefix, SDPA implementation, and decoding. Question and options remain available in history when absent from the final turn. All reported arms contain the same two output-format instructions. The reference arm was rerun to enforce this match; the other three arms were reused without modification. Thus Q_0O_0

Table 15: **Format-matched question/option restatement.** Entries are numbers of correct MCQ answers. Q_1 repeats the historical question at the final turn; O_1 repeats its original options. All four arms have identical output-format instructions.

Model	Operation	n	Q_0O_0	Q_1O_0	Q_0O_1	Q_1O_1
Qwen3-Omni	Reconfirm	128	125	121	127	128
Qwen3-Omni	Reground	128	74	93	127	126
Qwen3-Omni	Overall	256	199	214	254	254
MiniCPM-o 4.5	Reconfirm	128	110	116	126	125
MiniCPM-o 4.5	Reground	128	38	54	118	123
MiniCPM-o 4.5	Overall	256	148	170	244	248

in Table 15 is a format-matched diagnostic baseline, not an identical copy of the main-table native prompt.

Writing a_{qo} for accuracy, the question effect is $[(a_{10}-a_{00})+(a_{11}-a_{01})]/2$ and the option effect is $[(a_{01}-a_{00})+(a_{11}-a_{10})]/2$. For Reground, these are 7.03/33.59 pp for Qwen3-Omni and 8.20/58.20 pp for MiniCPM-o 4.5. The corresponding descriptive interactions, $a_{11}-a_{10}-a_{01}+a_{00}$, are -15.62 and -8.59 pp. Repeating options can assist option inheritance, letter mapping, and interpretation of the task; this manipulation does not distinguish those mechanisms.

E.3 Competing Media and Source-Fact Access

Table 16: **Following a newly introduced current-media answer in Rebind.** Counts use strict value matching. MCQ uses only cases where the new value is expressible by the original options; its denominator therefore differs from swap-accuracy evaluation.

Model	Interface	n	Before swap	After swap	Adopted/abandoned
Qwen3-Omni	OpenQA	1024	1	374	373/0
Qwen3-Omni	MCQ	980	1	400	399/0
MiniCPM-o 4.5	OpenQA	1024	2	97	96/1
MiniCPM-o 4.5	MCQ	980	0	34	34/0
Qwen3.8-Omni-Flash	OpenQA	1022	2	136	134/0
Qwen3.8-Omni-Flash	MCQ	980	1	212	211/0

Table 17: **Competitor-following sensitivity to media answerability.** The first three columns evaluate each endpoint independently with the same complete question. The final columns count post-swap matches to the new competitor; before-swap counts are zero in every row. The all-three-correct subset accompanies the full 128-task estimate.

Model	Target ST	Old ST	New ST	All selected	All three correct
Qwen3-Omni	120/128	120/128	123/128	40/128	37/111
MiniCPM-o 4.5	113/128	122/128	121/128	6/128	4/107
Qwen3.8-Omni-Flash	119/128	119/128	117/128	13/128	12/102

Removing current media improves Rebind, but also reduces input load. The media-swap condition instead retains a current competitor and substitutes a different clip supporting another answer to the same question, while preserving the historical target. Table 16 measures strict matches to that newly introduced answer. Its MCQ denominator includes only cases in which the destination value is expressible using the original options; it must not be confused with the larger swap-accuracy cohort. For example, Qwen3-Omni OpenQA target accuracy changes from 539/1,024 to 529/1,024, while new-value matches change from 1 to 374. Thus accuracy alone can miss substantial changes in the content of wrong answers. A whole-media substitution changes multiple perceptual attributes, not just a single value.

The answerability sensitivity sample uses seed 2804 to select 128 Rebind OpenQA tasks, 32 per modality/horizon cell. Each target, original competitor, and replacement competitor is queried separately with the same complete question, using exact cached endpoint requests. Three tasks share

competitor endpoints from development caches. Table 17 reports both all selected tasks and the post-hoc subset where all three endpoints are answered correctly. The latter is an explanatory sensitivity analysis and does not replace the full-sample estimate.

Historical source-fact probes ask a distinct question about the referenced historical media (Table 13). Each model/interface has 712 valid probes. The subsets without literal gold strings in historical assistant text contain 624, 590, and 620 tasks for Qwen3-Omni, MiniCPM-o 4.5, and Qwen3.8-Omni-Flash. Absence of a literal answer does not rule out every textual cue. These probes therefore complement original-task performance without constituting a direct decomposition of its errors.

E.4 Historical Slots and Explicit Reference Wording

Table 18: **History-slot interventions on a donor-eligible test-derived cohort.** Values are accuracy (%). Far-N inserts a neutral slot and Far-C replaces it with a competing slot. Near removes this slot. These conditions vary content/load as well as target distance.

Model	Interface	Operation	n	ST	Near	Far-N	Far-C
Qwen3-Omni	OpenQA	Reground	400	97.0	81.8	71.8	75.5
Qwen3-Omni	MCQ	Reground	400	99.8	43.8	43.2	47.2
Qwen3-Omni	OpenQA	Rebind	531	94.9	79.7	44.4	31.1
Qwen3-Omni	MCQ	Rebind	531	99.8	92.1	74.2	63.5
Qwen3.8-Omni-Flash	OpenQA	Reground	399	98.5	82.5	75.7	78.2
Qwen3.8-Omni-Flash	MCQ	Reground	400	99.5	92.2	89.5	91.0
Qwen3.8-Omni-Flash	OpenQA	Rebind	531	94.2	71.9	71.9	66.1
Qwen3.8-Omni-Flash	MCQ	Rebind	530	99.1	72.6	72.3	71.7

Table 19: **Two fixed reference rewrites for Qwen3.8-Omni-Flash.** Entries are correct OpenQA answers; R/H denotes method rescue/harm relative to cached original responses. The 128 tasks comprise 64 matched pairs.

Operation	n	Original	Message	R/H	Explicit media	R/H
Reconfirm	32	31	26	0/5	25	0/6
Reground	32	26	24	1/3	24	3/5
Retrieve	32	30	30	0/0	30	0/0
Rebind	32	25	26	4/3	28	4/1
Overall	128	112	106	5/11	107	7/12

The history-slot study uses 932 Test-derived tasks with eligible donor media. Near removes one historical media/assistant slot; Far-Neutral inserts such a slot; Far-Competing replaces its media and reply with a competing donor. Some donors are reused. In constructing the Rebind cohort, 400 targets originally at positions two through four are moved to the first position, also changing the ordinal cue. Table 18 uses common valid denominators separately for each operation and interface. Because source content, load, reply text, and sometimes position co-vary, this is not a pure distance experiment.

The wording sensitivity study samples 64 pairs/128 OpenQA tasks with seed 2805, giving 32 tasks per operation. Two rewrites are fixed in advance: a message-based reference and an explicit reference to the media attached to that message. Target ordinals are preserved, including non-first targets. Task-carrying requests continue to inherit the first question and use current media; the complete factual question is not restated. Table 19 reports heterogeneous changes and an overall accuracy decrease. The original response is a cached earlier response from the same hosted endpoint and native prefix; internal version/time changes are not separately controlled. The study neither rules out reference ambiguity nor establishes a general causal penalty for clearer references.

F Adaptation Protocol and Additional Results

Training data and candidates. Qwen3-Omni is adapted on 1,018 Short OpenQA views in 509 pairs, derived from 750 unique Dev-Train task IDs. There are 375 answer-preserving pairs and 134

target-switching pairs. Up to three verified candidate values are used per pair; duplicate detection strips and collapses whitespace and applies Unicode case folding while retaining the frozen display value for tokenization. Candidate scores average answer-token log probabilities, excluding prompt padding and EOS. Among answer-preserving pairs, 284 have at least two candidates for the JS term; the other 91 receive answer-token cross-entropy only. Historical text literally contains the relevant gold in 74/134 first-target views of the target-switching set (59/73 audio and 15/61 visual), so these views do not universally require media re-access.

For eligible answer-preserving pairs, the consistency-augmented objective is

$$\mathcal{L}_{\mathrm{SFT+Inv}} = \mathcal{L}_{\mathrm{CE}} + \lambda \, \mathrm{JS}(p_\theta(\cdot \mid h_i), p_\theta(\cdot \mid h_j)) \,, \tag{1}$$

where the distributions are over the verified candidate set and $\lambda = 0.1$. The SFT control sets the auxiliary coefficient to zero while retaining the auxiliary-forward budget. The candidate set is used during training, not supplied to the model at test time.

Matched optimization. SFT and SFT+Inv independently start from the same base checkpoint with seed 53126, identical data order, and identical LoRA initialization. LoRA targets text-attention query/key/value/output projections in layers 12–47 with rank 8, scale 16, dropout 0, and no bias adaptation. AdamW uses learning rate 10^{-5}, betas $(0.9, 0.999)$, epsilon 10^{-8}, no weight decay, a constant schedule, no warmup, and gradient clipping at 1.0. Global batch size is 5 for 204 updates, giving 1,020 answer exposures and 3,352 candidate forwards per arm, on two GPUs in one model-parallel rank. Both arms execute the same candidate forwards/backwards; SFT sets the auxiliary coefficient to zero. Cross-entropy is accumulated per answer view and the pair auxiliary term once per left-view exposure, divided by the same global batch size rather than separately normalized by pair count.

Table 20: **Adaptation results on Dev-Val.** Accuracy (%) on 500 tasks per interface. The additional sensitivity objective was evaluated on Dev-Val only.

Training objective	OpenQA		MCQ	
	ST	MT	ST	MT
SFT	90.8	74.8	96.8	80.8
SFT+Inv	90.0	75.0	96.8	80.6
SFT+Inv+sensitivity	90.8	74.4	96.8	81.0

Table 21: **Native-history adaptation on test tasks.** Accuracies are percentages; conversational degradation Δ_{conv} is in pp. Each checkpoint generates its own history; all methods have the same valid denominators. SFT and SFT+Inv have matched examples, initialization, updates, and auxiliary-forward budget.

Metric / subset	n	OpenQA			MCQ		
		Base	SFT	SFT+Inv	Base	SFT	SFT+Inv
Overall ST	5000	94.60	94.84	94.92	99.12	99.12	99.12
Overall MT	5000	77.26	80.68	80.70	77.28	79.64	79.84
Overall Δ_{conv}	5000	17.34	14.16	14.22	21.84	19.48	19.28
Reconfirm MT	1250	90.08	91.68	91.92	96.40	95.76	95.68
Reground MT	1250	81.76	85.84	85.52	57.76	62.96	63.36
Retrieve MT	1250	84.48	85.36	85.20	95.36	95.76	96.24
Rebind MT	1250	52.72	59.84	60.16	59.60	64.08	64.08
Short MT	2500	91.64	94.12	94.36	89.68	90.72	90.76
Long MT	2500	62.88	67.24	67.04	64.88	68.56	68.92

Selection and interpretation. A Dev-Val-only sensitivity-hinge variant ($\lambda = 0.1$, margin 0.2) was not evaluated on Test (Table 20). SFT and SFT+Inv generate native test histories without candidate reranking. Tables 21 and 22 report accuracy, repair/harm counts, and conditional derivation-root intervals. SFT reduces MCQ Reconfirm accuracy by 0.64 pp; JS adds only one correct OpenQA answer and ten MCQ answers over SFT, with intervals spanning zero. The observed improvement

Table 22: **Paired changes in native multi-turn answers.** $n = 5,000$ per row. Intervals are 95% percentile estimates from 2,000 derivation-root resamples (1,875 roots), not evidence of independence across parent media.

Comparison	Interface	Δ (pp)	Rescue/harm	Root interval
SFT vs. Base	OpenQA	+3.42	261/90	$[2.60, 4.23]$
SFT vs. Base	MCQ	+2.36	163/45	$[1.80, 2.90]$
SFT+Inv vs. SFT	OpenQA	+0.02	26/25	$[-0.24, 0.30]$
SFT+Inv vs. SFT	MCQ	+0.20	41/31	$[-0.14, 0.54]$

comes from SFT, with no clear incremental gain from consistency regularization. Development includes mixed horizons and interfaces, so Long/MCQ are not wholly unseen conditions. The study uses one model and seed, without an unpaired-data control.

Source-overlap sensitivity. Metadata-only filtering excludes tasks sharing any input-media parent with gradient-training inputs, retaining 3,125 tasks per interface (Table 23). SFT gains remain 3.14 pp on OpenQA (148 repairs, 50 harms) and 2.50 pp on MCQ (106 repairs, 28 harms), so improvement is not confined to shared training parents. Filtering changes task composition and does not remove all development exposure. The all-input Dev-disjoint subset contains no Long Retrieve/Rebind tasks, so its gains describe that restricted task mixture.

Table 23: **SFT gains under source-overlap exclusions.** Training refers to actual gradient-training inputs. Each subset is fixed without reading model scores; n is common to both interfaces. Gains are paired accuracy changes (pp); R/H counts repaired/harmed answers. Source filtering changes the task mixture.

Retained subset	n	OpenQA		MCQ	
		Gain	R/H	Gain	R/H
All Test tasks	5,000	+3.42	261/90	+2.36	163/45
Target parent disjoint from Dev	4,082	+3.58	218/72	+2.13	127/40
All input parents disjoint from Dev	1,928	+2.65	78/27	+1.66	54/22
Target parent disjoint from training	4,514	+3.57	239/78	+2.10	139/44
All input parents disjoint from training	3,125	+3.14	148/50	+2.50	106/28

G Inference-Time Assistance: Test and Dev-Val Protocols

Actions and routing. Restatement supplies the inherited question and original MCQ options. Reweighting resolves the user-visible target-media reference, scales non-target media and historical assistant-text keys by 0.5 at final-answer positions, and renormalizes attention. All inputs remain available; historical representations are unchanged. Routing selects restatement for historical-question requests and reweighting for complete-question requests, falling back to neutral execution for unresolved references. Its deterministic parser operates on explicit benchmark references without gold answers or operation labels.

Table 24: **Qwen3-Omni inference actions on Test.** $n = 5,000$ per interface. All actions share SDPA execution and cached base-model histories. ST is their shared neutral single-turn reference; Δ_{conv} is ST$-$MT (pp). This table compares actions; the runtime-matched neutral multi-turn comparison is reported separately on Dev-Val.

Action	OpenQA			MCQ		
	ST	MT	Δ_{conv}	ST	MT	Δ_{conv}
Restatement	95.00	81.48	13.52	99.12	87.80	11.32
Reweighting	95.00	79.40	15.60	99.12	77.32	21.80
Routed	95.00	84.02	10.98	99.12	88.02	11.10

Table 25: **Runtime-matched neutral comparison on Dev-Val.** Accuracy (%), with $n = 500$ per row. Neutral and all three actions use identical cached histories, media, SDPA execution, and decoding. The neutral ST condition is shared within each row.

Model	Interface	ST	Neutral	Restate	Reweight	Routed
Qwen3-Omni	OpenQA	89.2	71.4	76.0	73.2	78.2
Qwen3-Omni	MCQ	96.8	78.8	86.4	77.6	86.0
MiniCPM-o 4.5	OpenQA	89.2	59.8	74.2	61.4	77.8
MiniCPM-o 4.5	MCQ	99.6	59.6	78.2	60.2	78.2

Test action comparison. On 5,000 Test tasks per interface with cached Qwen3-Omni base histories, routing rescues/harms 140/13 OpenQA answers (+2.54 pp) and 21/10 MCQ answers (+0.22 pp) relative to restatement (Table 24). Routing reuses the selected action's output without another forward pass. Actions share SDPA execution, whereas the formal native-history baseline uses its original runtime; their difference therefore cannot isolate intervention effects or directly rank assistance against adaptation.

Validation neutral comparison. Table 25 matches neutral execution and actions in cached histories, media, SDPA runtime, and decoding on 500 Dev-Val tasks per model/interface. Reweighting changes OpenQA/MCQ accuracy by +1.8/$-$1.2 pp for Qwen3-Omni and +1.6/+0.6 pp for MiniCPM-o 4.5. Routing includes restatement and trails it by 0.4 pp on Qwen3-Omni MCQ. MiniCPM-o 4.5's single-turn OpenQA reference is 446/500 here versus 456/500 under the earlier native configuration; this difference is not a routing effect. These controls separate runtime and action effects on validation data, not in a full-Test neutral ablation.

H Scope and Limitations

ReTurn makes these requirements measurable through controlled task constructions; it does not exhaust open-ended conversation. Its references are explicit, histories are short relative to lifelong interaction, and source media can recur across partitions. Construction and scoring rely on automated checks, with alternative-judge and media-grounded sensitivity audits rather than human agreement estimates. The diagnostics characterize behavioral responses to controlled interventions; broader generalization remains to be assessed. Broader histories, independently held-out media, and replicated adaptation would test how far the present findings extend.