LEARNING WHAT TO PERCEIVE: COUNTERFACTUAL MEDIA CREDIT AND SELECTIVE CONSOLIDATION FOR OMNI-MODAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal active perception shifts multimodal understanding from static prediction to multi-turn acquisition of task-relevant evidence. Each perception action changes subsequent states and memory, making interaction part of task solving. Yet outcome-only reinforcement learning supervises acquisition through final answers: successful trajectories may contain redundant or misleading acquisitions, while failed ones may already contain decisive evidence. Outcome-only supervision therefore obscures whether an executed acquisition returned useful content and whether its temporal interval was more informative than matched alternatives. We introduce Media-Attributed Turn-Level Credit Assignment (**MATCA**), a hierarchical credit framework that uses explicit media counterfactuals to score returned content through the executed pathway and the temporal specificity of its selected interval. **MATCA** converts these comparisons into signed content and temporal-specificity credits and routes them only to the responsible action-type and timestamp value fields, allowing locally positive acquisitions in failed trajectories to be reinforced and locally negative ones in successful trajectories to receive negative credit. To consolidate acquisition behaviors discovered by the current policy, we introduce Selective Intervention Consolidation (**SIC**). **SIC** distills teacher actions into a reset student only when their student-completed branches outperform the corresponding student-action branches, providing an initialization for renewed **MATCA** optimization. In a separate 180-action paired audit under a fixed deterministic continuation policy, the positive, abstaining, and negative content-credit groups exhibit the predicted positive, zero, and negative mean terminal changes; among the 40 gated pairs with an observable answer flip, 32 agree with the credit sign. Under matched 7B settings, **SIC-initialized MATCA** outperforms OmniAgent by 6.6 points on average across seven benchmarks. It also scores above the reported Qwen3-Omni-30B-A3B results on four benchmarks: OmniVideoBench, LVOmniBench, WorldSense, and Video-MME.

1 INTRODUCTION

Answering questions over long-form videos and omni-modal information streams requires identifying useful evidence within temporally extended, redundant, and cross-modal content. Recent agentic approaches acquire media on demand, allowing a model to choose a perception pathway and temporal interval as its information needs evolve (Wang et al., 2025a;b; Lin et al., 2026; Xing et al., 2026; Xu et al., 2026). Training these decisions from final-answer rewards alone creates an acquisition–outcome mismatch: a successful trajectory may contain misleading or redundant acquisitions, while a failed trajectory may contain useful evidence followed by incorrect reasoning. Broadcasting one outcome-derived advantage across the trajectory does not distinguish these cases.

Local credit assignment already addresses this broader problem. Prior work rewards intermediate segment selection, tracks changes in target-answer belief, and uses counterfactual multimodal perturbations to refine policy updates (Xie et al., 2025; Auzina et al., 2026; Tao et al., 2026b; Yu et al., 2026; Jiang & Zou, 2026). We build on these ideas to study the structure of an omni-modal acquisition action: its returned content and its temporal placement require different comparisons. A real-versus-null comparison evaluates the content returned through the executed pathway, whereas

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

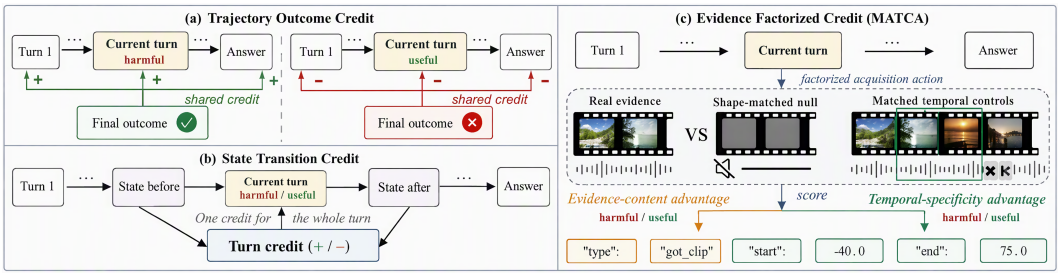

Figure 1: **Acquisition-specific credit through returned-media interventions.** Outcome-only supervision assigns a trajectory-level advantage to every acquisition. MATCA evaluates the content returned by an executed action against a structure-matched semantic null, then compares its content gain with gains from matched temporal placements. The resulting content and temporal-specificity credits supervise the action-type and timestamp value fields, respectively.

a comparison with matched intervals evaluates the specificity of its temporal placement. Our focus is to make these comparisons at the returned-media boundary, before the next policy response, and route their credit to the corresponding structured action fields (Figure 1). This defines an acquisition-specific contribution within existing local and counterfactual credit assignment.

We implement this approach in a three-stage training framework that evaluates, consolidates, and refines evidence acquisition. In the first stage, Media-Attributed Turn-Level Credit Assignment (MATCA) holds the interaction history and executed action metadata fixed and compares a frozen model's target-answer support under real media and a structure-matched semantic null. This content contrast evaluates the executed pathway's return; it does not compare alternative modality choices. A second contrast subtracts the median content gain of sampled intervals matched in pathway and acquisition budget, measuring temporal specificity relative to a typical sampled placement rather than identifying the best possible interval. Null-calibrated thresholds, abstention, and a content gate determine which signed credits are assigned to action-type and timestamp value tokens; other response tokens retain outcome supervision. These interventions are training-only and require no human turn-level annotations. We treat the frozen readout as a directional proxy and separately audit its thresholded content-credit direction against terminal behavior under a fixed deterministic continuation policy.

Because MATCA can only reinforce acquisitions sampled by the current policy, we use a reset student to consolidate its discovered acquisition behaviors before renewed exploration. In the second stage, Selective Intervention Consolidation (SIC) consolidates acquisition behaviors discovered by MATCA into a reset student while restricting supervision to the action fields tested by the paired intervention. Starting from the original agentic SFT checkpoint, SIC treats teacher actions as candidate interventions rather than unconditional imitation targets Agarwal et al. (2024); Li et al. (2026). Given the same interaction state and student-generated prefix, teacher and student propose alternative actions, and the current student completes both branches. SIC uses their realized return ordering as a selective data gate and distills only the action-value tokens of accepted teacher actions. The third stage resumes MATCA optimization from this consolidated student, forming a MATCA–SIC–MATCA loop of acquisition discovery, selective consolidation, and renewed refinement without changing the inference-time interaction loop.

Under matched 7B settings, MATCA outperforms OmniAgent by 5.9 points on average across four joint audio-visual benchmarks, with the largest gain of 9.2 points on JointAVBench. Against the matched TAURA policy on LVOmniBench, it improves accuracy by 5.2 points while using 6.5% fewer perception actions, 7.4% fewer visual frames, and 9.1% less audio duration. SIC-initialized MATCA further widens the average margin over OmniAgent to 6.6 points across all seven benchmarks and scores above the reported Qwen3-Omni-30B-A3B results on four benchmarks. SIC also yields a stronger reset-student initialization than ungated teacher distillation on all four ablation benchmarks.

Our contributions are threefold:

1. We formulate acquisition-specific credit through two paired comparisons: real versus structure-matched null returns for executed-pathway content, and selected versus matched alternative placements for temporal specificity. This separates the supervision questions associated with returned content and temporal placement while making the scope of each comparison explicit.

2. We develop a three-stage training framework: MATCA assigns signed executed-pathway content and temporal-specificity credit through explicit media counterfactuals, SIC selectively consolidates actions chosen by student-completed branch comparisons, and renewed MATCA optimization refines the consolidated policy.

3. Under matched 7B settings, SIC-initialized MATCA exceeds OmniAgent by 6.6 points on average across seven benchmarks. A paired continuation audit supports the directional use of content credit, nested ablations support the two credit components and selective consolidation, and long-video analysis shows improved accuracy with reduced perception cost.

2 Related Work

2.1 Long-Form Omni-Modal Active Perception

Agentic methods acquire visual and audio evidence on demand through temporal, spatial, and modality-specific tools (Wang et al., 2025a;b; Lin et al., 2026; Rasheed et al., 2025; Shen et al., 2025; Zeng et al., 2026; Xu et al., 2026). OmniAgent formulates native audio-visual exploration as a unified POMDP policy (Xing et al., 2026). We build on this setting to study credit for structured acquisition actions.

2.2 Credit Assignment for Agentic Active Perception

Intermediate acquisition rewards are established: Video-MTR supervises segment selection using annotated temporal relevance (Xie et al., 2025); GiGPO groups actions at repeated states (Feng et al., 2025); TAPO transfers credit between similar acquisition calls across successful and failed trajectories (Dong et al., 2026).

Belief-based methods instead derive local progress from predictions. ΔBelief-RL and IGPO use changes in target-concept log belief and target-answer probability, respectively (Auzina et al., 2026; Wang et al., 2026). TRACE evaluates a frozen answer-based potential at tool boundaries (Tao et al., 2026b), while SIOP tracks beliefs over semantic outcome clusters (Hu et al., 2026). Whereas these methods compare interaction prefixes, MATCA holds the executed action fixed and contrasts real and structure-matched null returns at the same boundary.

2.3 Counterfactual and Token-Level Credit

COMA marginalizes one agent's action alternatives while fixing other agents' actions (Foerster et al., 2017). For multimodal reasoning, CFPO perturbs visual representations (Yu et al., 2026), SRPO contrasts original and corrupted inputs for role-aware token weighting (Jiang & Zou, 2026), and PGPO measures token-level visual dependence (Ye et al., 2026). POAD uses action-decomposed Bellman backups for intra-action credit (Wen et al., 2024); APPO rewards perception-oriented tokens through attention to crucial frames (Du et al., 2026).

MATCA combines returned-media interventions with a separate comparison over matched temporal placements, routing their credits to action-type and timestamp fields. Its content credit concerns the executed pathway, not an advantage over unexecuted modalities. SIC subsequently consolidates actions through student-completed branch comparisons (Section 3.5).

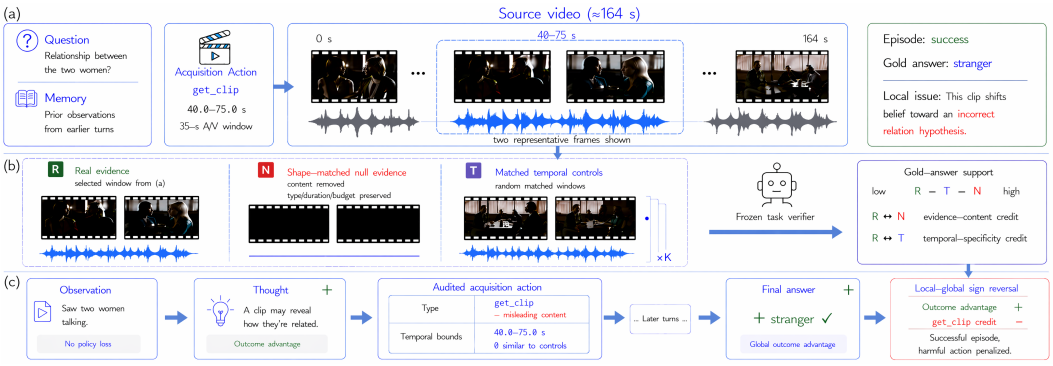

Figure 2: **Negative content credit in a successful trajectory.** Despite a correct final answer, the clip lowers frozen target-answer support relative to a structure-matched null (N). Real evidence (R) and matched temporal controls (T) yield content and temporal-specificity credits for action fields; other tokens retain outcome credit.

3 METHOD

3.1 OMNI-MODAL ACTIVE PERCEPTION

Each example is $x = (q, V, \mathcal{Y}, y^\star)$, where q is a question, V is a temporally indexed audio-visual input, $\mathcal{Y}$ is the candidate set, and $y^\star$ is the gold answer. At turn t, the policy conditions on $H_t = (q, M_t)$, where M_t contains the preceding structured responses and returned media. It produces $r_t = (o_t, z_t, c_t, a_t)$ with observation o_t, reasoning trace z_t, confidence c_t, and action $a_t = (m_t, I_t)$, where $m_t \in \{\texttt{frames}, \texttt{audio}, \texttt{clip}\}$ selects the perception pathway and I_t specifies a temporal interval or sampling location. The environment returns $e_t = \mathrm{Env}(V, a_t)$ and updates the history as $H_{t+1} = \mathrm{Update}(H_t, r_t, e_t)$.

For each question, we sample N trajectories $\{\tau_i\}_{i=1}^N$. Trajectory i terminates with prediction $\hat{y}_i$ and reward $R_i = R(\hat{y}_i, y^\star)$. Following group-relative policy optimization (Shao et al., 2024), its outcome advantage is

$$A_i^{\mathrm{out}} = \frac{R_i - \mu_R}{\sigma_R + \epsilon}. \tag{1}$$

MATCA replaces this outcome advantage on acquisition fields when local scoring passes the calibrated thresholds; remaining tokens retain outcome supervision.

3.2 COUNTERFACTUAL RETURNED-MEDIA SCORING

At the start of a MATCA phase, we freeze the current initialization as $\pi_{\mathrm{ref}} = \pi_{\theta_0}$. Appending a fixed read-only answer probe to context C yields the *candidate-normalized gold belief* $S_{\mathrm{ref}}(C; y^\star)$ in one forward pass. Probe serialization and calibration details are provided in the supplementary material.

We score returns before the next policy response, holding history and executed action metadata fixed:

$$C_t^{\mathrm{real}} = H_t \oplus a_t \oplus e_t^{\mathrm{real}}, \tag{2}$$

$$C_t^{\mathrm{mask}} = H_t \oplus a_t \oplus e_t^{\mathrm{mask}}, \tag{3}$$

where e_t^{mask} preserves the format, size, and token structure of the tool return while removing semantic content. Holding the action metadata and timestamps fixed, the *content gain* is

$$G_t^{\mathrm{content}} = S_{\mathrm{ref}}(C_t^{\mathrm{real}}; y^\star) - S_{\mathrm{ref}}(C_t^{\mathrm{mask}}; y^\star). \tag{4}$$

This is a directional proxy, not a terminal-value estimate; Section 4.3.1 audits its thresholded sign under a fixed deterministic continuation policy. Content gain does not determine whether the selected interval is preferable to equally costly alternatives. Let μ_t be a distribution over intervals

reachable under the same pathway m_t and matched query configuration. We sample K matched controls $I_t^{\mathrm{ctrl},j} \sim \mu_t$, preferring non-overlapping placements, and compute

$$G_t^{\mathrm{ctrl},j} = S_{\mathrm{ref}}(C_t^{\mathrm{ctrl},j,\mathrm{real}}; y^\star) - S_{\mathrm{ref}}(C_t^{\mathrm{ctrl},j,\mathrm{mask}}; y^\star). \tag{5}$$

The *temporal-specificity gain* is

$$G_t^{\mathrm{time}} = G_t^{\mathrm{content}} - \operatorname*{median}_{j\in\{1,\dots,K\}} G_t^{\mathrm{ctrl},j}. \tag{6}$$

This contrast measures gain relative to a typical sampled placement, not the best interval. The median reduces sensitivity to individual extreme controls.

3.3 Executed-Pathway Content and Temporal-Specificity Credit

On held-out data, we calibrate fixed one-sided thresholds $\tau_{m,d}^{\pm}$ and robust scales $s_{m,d}$, and define $\rho_{m,d}(g) = \mathrm{clip}\big(g/(s_{m,d}+\epsilon), -1, 1\big)$. When $G_t^{\mathrm{content}} \notin [-\tau^-_{m_t,\mathrm{content}}, \tau^+_{m_t,\mathrm{content}}]$, the executed-pathway content credit is

$$C_t^{\mathrm{content}} = \rho_{m_t,\mathrm{content}}(G_t^{\mathrm{content}}). \tag{7}$$

Otherwise $C_t^{\mathrm{content}} = \perp$ denotes abstention. This is executed-pathway content credit relative to the null, not an advantage over unexecuted pathways.

Let $g_t = G_t^{\mathrm{time}}$, $\rho_t = \rho_{m_t,\mathrm{time}}$, and $\tau_t^{\pm} = \tau^{\pm}_{m_t,\mathrm{time}}$. For a content-reliable turn,

$$\widetilde{C}_t^{\mathrm{time}} = \begin{cases} \rho_t(g_t), & g_t \notin [-\tau_t^-, \tau_t^+], \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

We then apply a content gate:

$$C_t^{\mathrm{time}} = \begin{cases} \min\{\widetilde{C}_t^{\mathrm{time}}, 0\}, & C_t^{\mathrm{content}} < 0, \\ \widetilde{C}_t^{\mathrm{time}}, & C_t^{\mathrm{content}} \geq 0, \\ \perp, & C_t^{\mathrm{content}} = \perp. \end{cases} \tag{9}$$

Negative content credit blocks positive temporal credit. When content passes the gate but temporal gain lies in the null region, timestamp credit is zero; only abstention falls back to outcome supervision.

3.4 Decoupled Turn-Level Policy Optimization

We parse each action and locate the token spans encoding its JSON values. Let $\mathcal{D}_{i,t}^{\mathrm{type}}$ and $\mathcal{D}_{i,t}^{\mathrm{time}}$ denote the action-type and timestamp value spans. Gated local credits are routed as

$$A_{i,u}^{\mathrm{local}} = \begin{cases} C_{i,t}^{\mathrm{content}}, & u \in \mathcal{D}_{i,t}^{\mathrm{type}}, \\ C_{i,t}^{\mathrm{time}}, & u \in \mathcal{D}_{i,t}^{\mathrm{time}}. \end{cases} \tag{10}$$

If a field is unreliable or unparsable, MATCA abstains and leaves its tokens on the outcome branch. With $M_{i,u}^{\mathrm{local}}$ indicating a locally credited action-value token, the complementary mask $M_{i,u}^{\mathrm{out}} = M_{i,u}^{\mathrm{response}}(1 - M_{i,u}^{\mathrm{local}})$ ensures that no token receives both local and outcome advantages. Let $r_{i,u}(\theta) = \pi_\theta(w_{i,u} \mid w_{i,<u}) / \pi_{\mathrm{old}}(w_{i,u} \mid w_{i,<u})$ and, following PPO (Schulman et al., 2017), let

$$\mathrm{PG}(r, A) = \min(rA, \mathrm{clip}(r, 1-\epsilon_l, 1+\epsilon_h)A)\,.$$

Using Avg_M for the token mean over active mask entries,

$$\begin{aligned} \mathcal{L}_{\mathrm{out}} &= -\,\mathrm{Avg}_{M^{\mathrm{out}}}\left[\mathrm{PG}(r_{i,u}, A_i^{\mathrm{out}})\right], \\ \mathcal{L}_{\mathrm{local}} &= -\,\mathrm{Avg}_{M^{\mathrm{local}}}\left[\mathrm{PG}(r_{i,u}, A_{i,u}^{\mathrm{local}})\right], \\ \mathcal{L}_{\mathrm{MATCA}} &= \mathcal{L}_{\mathrm{out}} + \lambda_{\mathrm{local}}\mathcal{L}_{\mathrm{local}} + \beta_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} - \beta_{\mathrm{H}}\mathcal{H}(\pi_\theta). \end{aligned} \tag{11}$$

Separate normalization preserves the weight of sparse action-value tokens. Their local credit can oppose the trajectory outcome, while other tokens retain outcome supervision. The masks partition token advantages; both losses still update shared parameters.

3.5 Selective Intervention Consolidation

SIC freezes the Phase-I teacher π_{T} and resets student π_ϕ to Agentic SFT. A fixed pool contains approximately 5,000 training-only SFT states that are nonterminal, restorable, history-valid, media-available, and action-executable. Replay restores histories and residual budgets; prefixes, actions, continuations, returns, and gates are regenerated online.

At reconstructed state $\bar{H}_i$, the student generates prefix $p_i^{\mathrm{S}} = (o_i^{\mathrm{S}}, z_i^{\mathrm{S}}, c_i^{\mathrm{S}})$. Given this same context, student and teacher sample actions a_i^{S} and a_i^{T}. The current student independently completes both branches from copies of the post-prefix state, yielding $\tau_i^{\mathrm{S}\to\mathrm{S}}$ and $\tau_i^{\mathrm{T}\to\mathrm{S}}$. Only the candidate action is systematically changed; continuation sampling introduces Monte Carlo variation.

Let v_i^{S} and v_i^{T} indicate that the student and teacher actions are valid and executable, and let n_i^b indicate that branch b terminates normally. We accept the teacher intervention iff

$$\gamma_i = \mathbf{1}\begin{bmatrix} v_i^{\mathrm{S}} \wedge v_i^{\mathrm{T}} \\ \wedge\, n_i^{\mathrm{S}\to\mathrm{S}} \wedge n_i^{\mathrm{T}\to\mathrm{S}} \\ \wedge\, R(\tau_i^{\mathrm{T}\to\mathrm{S}}) > R(\tau_i^{\mathrm{S}\to\mathrm{S}}) \end{bmatrix}. \tag{12}$$

This single-sample gate selects by realized return ordering, not expected action advantage; return differences do not weight accepted tokens.

For accepted interventions, we teacher-force the student along the accepted teacher action and distill only executable action-value positions $\mathcal{D}_i^{\mathrm{act}}$. Let $w_{i,u}^{\mathrm{T}}$ denote the teacher-generated token at action position u. We optimize

$$\ell_{i,u}^{\mathrm{SIC}} \triangleq -\log \pi_\phi\big(w_{i,u}^{\mathrm{T}} \mid \bar{H}_i, p_i^{\mathrm{S}}, w_{i,<u}^{\mathrm{T}}\big)\,. \tag{13}$$

$$\mathcal{L}_{\mathrm{SIC}} = \frac{\sum_i \gamma_i \sum_{u\in\mathcal{D}_i^{\mathrm{act}}} \ell_{i,u}^{\mathrm{SIC}}}{\sum_i \gamma_i \,|\mathcal{D}_i^{\mathrm{act}}|}. \tag{14}$$

Batches without accepted interventions are skipped. Because the branches share a student prefix and systematically vary the candidate action, SIC applies loss only to accepted executable action values; JSON keys, scaffolding, reasoning, confidence, and continuations remain context (Kim & Rush, 2016; Dong et al., 2023).

3.6 Three-Phase Training Procedure

Starting from agentic SFT, Phase I applies MATCA to obtain π_{T}. Instead of directly continuing π_{T}, Phase II freezes it, resets the student to SFT, and uses Eq. equation 14 to distill only teacher action fields accepted by Eq. equation 12, yielding the consolidated student π_{C}. Phase III freezes $\pi_{\mathrm{ref}} = \pi_{\mathrm{C}}$, recalibrates $\{\tau_{m,d}^{\pm}, s_{m,d}\}$, and resumes MATCA on new on-policy trajectories. SIC uses two student-completed rollouts per replay state during training. At deployment, the consolidated policy follows the original interaction loop without an auxiliary scorer or additional tool calls.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We report multiple-choice accuracy on seven benchmarks. Daily-Omni (Zhou et al., 2025), OmniVideoBench (Li et al., 2025b), JointAVBench (Chao et al., 2025), and LVOmniBench (Tao et al., 2026a) require joint audio-visual understanding; LVOmniBench further extends the evaluation to videos lasting 10–90 minutes. WorldSense (Hong et al., 2025), Video-MME without subtitles (Fu et al., 2025a), and VSI-Bench (Yang et al., 2024) evaluate real-world omni-modal understanding, general video understanding, and visual-spatial reasoning, respectively.

Baselines. Table 1 compares proprietary and open-source omni-modal models, single-stream specialists, and active-perception agents. For the controlled post-training comparison, MATCA uses the same tools, media budget, turn limit, answer parser, and decoding settings as the Qwen2.5-Omni-7B backbone and OmniAgent entries.

Table 1: Accuracy on seven benchmarks; Video-MME is evaluated without subtitles. *Size* denotes language-model parameters and *Mod.* the evaluation input. Boldface marks the best results among 7B open-source models. SIC-Initialized MATCA is obtained through Phase-I MATCA → Phase-II SIC → Phase-III MATCA.

			Joint audio-visual				General and visual		
Model	Size	Mod.	Daily-Omni	Omni-Video	JointAV	LV OmniBench	World-Sense	Video-MME	VSI-Bench
Proprietary models									
GPT-4o	—	V	56.5	—	45.0	—	42.6	71.9	34.0
Gemini-2.5-Flash	—	A+V	73.1	52.4	58.0	—	52.3	—	—
Gemini-2.5-Pro	—	A+V	—	58.9	65.3	—	65.1	—	51.5
Open-source omni-modal models									
VideoLLaMA2 (2024)	7B	A+V	35.2	29.2	46.8	27.2	25.4	47.9	—
VITA-1.5 (2025b)	7B	A+V	—	30.5	—	—	36.9	—	—
Ola (2025)	7B	A+V	50.7	—	—	—	—	—	—
video-SALMONN 2+ (2025)	7B	A+V	—	—	46.7	32.7	50.9	—	—
OmniVinci (2025)	9B	A+V	66.5	32.1	—	—	48.2	68.2	—
Qwen3-Omni (2025b)	30B-A3B	A+V	71.9	38.4	63.6	35.8	54.0	70.5	—
Qwen2.5-Omni (backbone) (2025a)	7B	A+V	60.1	29.3	52.3	32.0	45.4	64.8	35.5
Open-source video models (visual stream only)									
LLaVA-OneVision (2025a)	7B	V	—	—	—	—	37.7	58.2	32.4
InternVL2.5 (2024)	8B	V	—	—	51.7	—	39.1	—	—
Qwen2.5-VL (2025)	7B	V	40.7	29.8	47.7	—	—	65.1	33.5
Open-source audio models (audio stream only)									
Qwen2-Audio (2024)	7B	A	31.5	—	39.5	24.7	—	—	—
Kimi-Audio (2025)	7B	A	—	—	45.6	—	—	—	—
Audio Flamingo 3 (2025)	7B	A	49.9	—	—	—	—	—	—
Agentic active-perception models									
LongVT (2025)	7B	V	—	—	—	—	—	55.9	34.4
VITAL (2025)	7B	V	—	—	—	—	—	64.1	41.8
Zoom-Zero (2025)	7B	V	—	—	—	—	—	66.0	—
OmniAgent (2026)	7B	A+V	64.8	37.1	52.6	33.3	47.2	67.8	48.4
Ours									
MATCA	7B	A+V	69.9	41.2	61.8	38.5	52.6	69.3	48.6
Δ over backbone			+9.8	+11.9	+9.5	+6.5	+7.2	+4.5	+13.1
Δ over OmniAgent			+5.1	+4.1	+9.2	+5.2	+5.4	+1.5	+0.2
SIC-Initialized MATCA	7B	A+V	**71.8**	**43.1**	**62.7**	**39.8**	**55.7**	**72.2**	**52.0**

Implementation. We fine-tune Qwen2.5-Omni-7B on synthesized Observation–Thought–Action trajectories covering `get_frames`, `get_audio`, and `get_clip`, obtaining the common Agentic SFT initialization. RL prompts combine VSI-Train-10k (Yang et al., 2024), Video-Holmes (Cheng et al., 2025), and LongVideo-Reason (Chen et al., 2025) in multiple-choice format. We use $N = 8$ trajectories per prompt, batches of 32 prompts, at most 22 turns, and 6×8 GPUs. SIC-Initialized MATCA denotes the full MATCA–SIC–MATCA schedule; SIC replays a fixed state pool while regenerating branches under the current student. Scoring, calibration, optimization, and compute details are in the supplementary material.

4.2 Main Results

Accuracy across benchmarks. MATCA improves over OmniAgent on all seven benchmarks (Table 1), with a 5.9-point mean gain across the four joint audio-visual benchmarks and a 9.2-point gain on JointAVBench. Gains on Video-MME and VSI-Bench are smaller (1.5 and 0.2 points). SIC-Initialized MATCA improves all seven scores further, reaching 55.7 on WorldSense, 72.2 on Video-MME, and 52.0 on VSI-Bench, with a 6.6-point mean margin over OmniAgent. The ablations below distinguish local credit, consolidation, and final training-path performance.

Table 2: MATCA ablation (accuracy, %).

Method	Daily-Omni	Omni-Video	JointAV	Video-MME
Qwen2.5-Omni-7B	60.1	29.3	52.3	64.8
+ Agentic SFT	63.3	35.2	55.2	67.2
+ GRPO	62.8	35.6	57.1	68.9
+ Content credit only	67.8	40.9	59.4	68.0
+ MATCA (full)	**69.9**	**41.2**	**61.8**	**69.3**

Table 3: SIC ablation (accuracy, %).

Training path	Daily-Omni	Omni-Video	JointAV	Video-MME
Agentic SFT	63.3	35.2	55.2	67.2
Ungated distillation	61.2	34.7	55.2	63.9
Random-gate distillation	61.4	34.6	54.8	64.3
SIC initialization (Phase II)	65.1	37.1	59.7	66.9
Phase-I MATCA teacher	69.9	41.2	61.8	69.3
Continued MATCA (compute-matched)	69.8	41.3	61.4	69.6
SIC-Initialized MATCA (Phase III)	**71.8**	**43.1**	**62.7**	**72.2**

Notes. Columns report Daily-Omni, OmniVideoBench, JointAVBench, and Video-MME. Left: GRPO uses outcome reward only; "Content credit only" removes temporal-specificity credit. All RL settings are matched. Right: random gate is update-matched in selected complete actions and nonempty steps (action-value-token difference: 0.8%). Continued MATCA matches the full SIC path in total training compute.

4.3 Ablation Studies

4.3.1 MATCA Ablation

Table 2 traces a matched progression from Agentic SFT and outcome-only GRPO to action-local content credit and full MATCA. Temporal-specificity credit is evaluated only after action-local content scoring, and all RL variants share the same training and inference budgets. We separately compare explicit media-counterfactual credit with a state-difference signal in the supplementary material.

Content and temporal credit. Content credit improves GRPO by 5.0, 5.3, and 2.3 points on Daily-Omni, OmniVideoBench, and JointAVBench, but reduces Video-MME by 0.9 points. Adding temporal-specificity credit raises all four scores by 2.1, 0.3, 2.4, and 1.3 points. This nested comparison supports its incremental benefit given content credit; it does not isolate a standalone temporal effect.

Behavioral audit of content-credit direction. We audit content-credit sign on 180 acquisitions by replaying each history and executed action with real or structure-matched null returns, then measuring terminal correctness changes under the same frozen deterministic continuation policy. The positive, abstaining, and negative credit groups have mean changes of $+0.200$, 0.000, and -0.175, respectively. All 180 pairs are retained, including 130 (72.2%) with no answer change. Among the 40 gated pairs with an observable terminal flip, $32/40 = 80.0\%$ agree with the predicted sign (Wilson 95% CI $[65.2\%, 89.5\%]$); two-sided exact McNemar tests give $p = 0.0043$ and $p = 0.0309$ for the positive and negative groups. This audit supports the directional use of gated content credit under the fixed continuation policy; the complete protocol and a separate same-state, same-window cross-pathway diagnostic are provided in the supplementary material.

4.3.2 SIC Ablation

We compare SIC with ungated and matched random-gate distillation at initialization, and with compute-matched continued MATCA after Phase III. All distillation paths reset to the Agentic SFT checkpoint: ungated uses all valid teacher actions, random gate matches SIC per update in selected complete actions and nonempty optimization steps, and SIC retains only actions accepted by Eq. equation 12. Continued MATCA starts from π_{T} and receives additional online optimization to match the full SIC path in total training compute.

Phase-II SIC exceeds matched random gating by 3.4 points on average and outperforms both distillation controls on all four benchmarks, although its Video-MME score remains 0.3 points below SFT. Phase-II results compare reset-student initialization quality. Phase III compares the complete SIC-initialized path with compute-matched continued MATCA, yielding a 1.9-point average gain.

4.4 LONG-VIDEO COMPUTE ANALYSIS

Table 4 compares MATCA with OmniAgent's TAURA objective (Xing et al., 2026), matching Agentic SFT initialization, action space, turn limit, and media budget. This comparison evaluates the full objectives.

Table 4: MATCA versus a matched OmniAgent baseline trained with TAURA on LVOmniBench. Bold marks the better value. Duration bins are in minutes and contain 500, 350, 120, and 44 examples; All is sample-weighted. Frames and audio seconds count the returned streams, so a synchronized clip contributes to both.

Metric	Method	10–30	30–50	50–70	70–90	All
Effectiveness (↑)						
Accuracy	OmniAgent	34.7	32.7	30.7	28.7	33.3
	MATCA	**39.4**	**38.1**	**36.7**	**35.4**	**38.5**
Temporal acc.	OmniAgent	30.5	28.2	25.8	23.5	28.8
	MATCA	**35.7**	**34.1**	**32.3**	**30.7**	**34.5**
Perception cost (↓)						
Actions/Q	OmniAgent	8.5	9.5	10.5	11.0	9.2
	MATCA	**8.1**	**8.8**	**9.4**	**9.8**	**8.6**
Visual frames/Q	OmniAgent	110	125	145	165	121.7
	MATCA	**104**	**115**	**130**	**145**	**112.7**
Audio sec./Q	OmniAgent	300	430	620	800	404.4
	MATCA	**280**	**390**	**550**	**690**	**367.7**

Accuracy and perception cost. MATCA improves accuracy in every duration bin: the margin grows from 4.7 points at 10–30 minutes to 6.7 at 70–90 minutes (5.2 overall). Temporal-question gains range from 5.2 to 7.2 points (5.7 overall). Overall perception actions, visual frames, and audio duration fall by 6.5%, 7.4%, and 9.1%, respectively; the longest bin shows reductions of 10.9%, 12.1%, and 13.8%. These results show higher accuracy with lower measured perception cost, but do not establish lower end-to-end latency or isolate the mechanism behind the duration trend.

5 CONCLUSION

We presented MATCA, which intervenes on returned media to derive signed executed-pathway content and temporal-specificity credits and routes them to the corresponding action fields. SIC then consolidates branch-compared acquisition actions into a reset student, enabling renewed MATCA optimization from a stronger initialization. A 180-action paired continuation audit supports the directional relation between gated content credit and terminal behavior under a fixed continuation policy. Across seven benchmarks, the combined training pipeline improves accuracy over OmniAgent; on long videos, MATCA also raises accuracy while lowering perception cost. Nested MATCA ablations and Phase-II consolidation controls support the roles of local acquisition credit and selective transfer, respectively. Together, MATCA and SIC improve omni-modal active perception without changing the deployment-time interaction loop.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, assist with the retrieval and discovery of related work, draft portions of the manuscript, and generate candidate multi-turn Observation–Thought–Action trajectories for Agentic SFT through model rollouts. As described in the supplementary material, these trajectories were generated from training-split seed questions and filtered through outcome verification and a rationality audit. We did not use generative AI tools to prove mathematical claims. The authors reviewed and verified all AI-assisted text, literature suggestions, and generated training data, and take responsibility for the final content of this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Mohammad Sadegh Akhondzadeh, Vijay Lingam, Atula Tejaswi, Chanakya Ekbote, Sujay Sanghavi, and Aleksandar Bojchevski. Reward-gated on-policy distillation, 2026. URL https://arxiv.org/abs/2607.04037.

Ilze Amanda Auzina, Joschka Strüber, Sergio Hernández-Gutiérrez, Shashwat Goel, Ameya Prabhu, and Matthias Bethge. Intrinsic credit assignment for long horizon interaction, 2026. URL https://arxiv.org/abs/2602.12342.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-vl technical report, 2025.

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. Jointavbench: A benchmark for joint audio-visual reasoning evaluation, 2025. URL https://arxiv.org/abs/2512.12772.

Yukang Chen, Wei Huang, Baifeng Shi, et al. Scaling RL to long videos. 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/fc2865a2fdd5fd401050d37ee4946850-Abstract-Conference.html.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2024.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can MLLM think like holmes for complex video reasoning?, 2025.

Zesen Cheng, Sicong Leng, Hang Zhang, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024. URL https://arxiv.org/abs/2406.07476.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, et al. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Chengqi Dong, Chuhuai Yue, Hang He, Yandong Liu, Fenghe Tang, S. Kevin Zhou, Xiaohan Wang, Jiajun Chai, and Guojun Yin. TAPO: Tool-aware policy optimization via credit transfer for multimodal search agents, 2026. URL https://arxiv.org/abs/2606.05784.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: reward ranked finetuning for generative foundation model alignment. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=m7p5O7zblY.

Henghui Du, Chang Zhou, Xi Chen, and Di Hu. APPO: Attention-guided perception policy optimization for video reasoning, 2026. URL https://arxiv.org/abs/2602.23823.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training, 2025. URL https://arxiv.org/abs/2505.10978.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients, 2017.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24108–24118. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Chaoyou Fu, Haojia Lin, Xiong Wang, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 75300–75320. Curran Associates, Inc., 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6ce9d51dded7dac82b3b4d3dbb1d73bc-Paper-Conference.pdf.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models, 2025.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms, 2025.

Senkang Hu, Yong Dai, Xudong Han, Zhengru Fang, Yuzhi Zhao, et al. Self-induced outcome potential: Turn-level credit assignment for agents without verifiers, 2026.

Bingqing Jiang and Difan Zou. Structured role-aware policy optimization for multimodal reasoning, 2026. URL https://arxiv.org/abs/2605.07274.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1317–1327. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1139. URL https://doi.org/10.18653/v1/d16-1139.

KimiTeam, Ding Ding, Zeqian Ju, et al. Kimi-audio technical report, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, et al. 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, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms, 2025b.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, and Jinqiao Wang. On-policy distillation with curriculum turn-level guidance for multi-turn agents, 2026.

Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Videoseek: Long-horizon video agent with tool-guided seeking, 2026.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment, 2025.

Hanoona Rasheed, Mohammed Zumri, Muhammad Maaz, Ming-Hsuan Yang, Fahad Shahbaz Khan, and Salman H. Khan. Video-com: Interactive video reasoning via chain of manipulations, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Xiaoqian Shen, Min-Hung Chen, Yu-Chiang Frank Wang, Mohamed Elhoseiny, and Ryo Hachiuma. Zoom-zero: Reinforced coarse-to-fine video understanding via temporal zoom-in, 2025. URL https://arxiv.org/abs/2512.14273.

Changli Tang, Yixuan Li, Yudong Yang, et al. video-salmonn 2: Caption-enhanced audio-visual large language models, 2025. URL https://arxiv.org/abs/2506.15220.

Keda Tao, Yuhua Zheng, Xu Jia, et al. Lvomnibench: Pioneering long audio-video understanding evaluation for omnimodal llms, 2026a.

Leitian Tao, Baolin Peng, Wenlin Yao, Tao Ge, Hao Cheng, Mike Hang Wang, Jianfeng Gao, and Sharon Li. TRACE: Turn-level reward assignment via credit estimation for long-horizon agents, 2026b. URL https://arxiv.org/abs/2607.13988.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *International Conference on Learning Representations*, 2026.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 58–76, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-72989-8.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, et al. Active video perception: Iterative evidence seeking for agentic long video understanding, 2025b.

Muning Wen, Ziyu Wan, Weinan Zhang, Jun Wang, and Ying Wen. Reinforcing language agents via policy optimization with action decomposition, 2024. URL https://arxiv.org/abs/2405.15821.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-MTR: Reinforced multi-turn reasoning for long video understanding, 2025. URL https://arxiv.org/abs/2508.20478.

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, et al. Native active perception as reasoning for omni-modal understanding, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, et al. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-omni technical report, 2025b.

Ke Xu, Yuhao Wang, Ziyang Cheng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. Agentic active omni-modal perception for multi-hop audio-visual reasoning, 2026.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces, 2024.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, et al. Longvt: Incentivizing "thinking with long videos" via native tool calling, 2025.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding LLM, 2025.

Zekai Ye, Qiming Li, Xiaocheng Feng, Ruihan Chen, Ziming Li, Haoyu Ren, Kun Chen, Dandan Tu, and Bing Qin. Not all tokens see equally: Perception-grounded policy optimization for large vision-language models, 2026. URL https://arxiv.org/abs/2604.01840.

Zhangyuan Yu, Wanran Sun, Guangjing Yang, Xiaohu Wu, and Qicheng Lao. CFPO: counterfactual policy optimization for multimodal reasoning, 2026.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning, 2026. URL https://arxiv.org/abs/2601.23224.

Haoji Zhang, Xin Gu, Jiawen Li, et al. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning, 2025.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. SAGE-OPD: Selective agent-guided intervention for multi-turn on-policy distillation, 2026. URL https://arxiv.org/abs/2606.19659.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities, 2025.

A Exact Counterfactual Scoring Protocol

A.1 Eligible acquisition turns and scoring time

We score only successfully executed multiple-choice acquisition turns whose structured action can be parsed, whose returned media can be decoded, and whose decision-value token spans can be located exactly. An eligible action must be one of `get_frames`, `get_audio`, or `get_clip`; its timestamps must lie inside the source media; and all matched controls must be executable under the same tool configuration. A failed tool call, missing media file, invalid candidate set, non-single-token option label, truncated action value, or failed span alignment causes abstention rather than an imputed score.

The scoring context ends immediately after the environment return and before the next policy response. Consequently, the current acquisition is represented by the returned media itself and cannot receive credit for a summary or claim that the policy writes on the following turn. Persistent evidence from earlier turns is serialized identically in every branch of the current comparison.

A.2 Real, masked, and matched-control branches

For an executed action $a_t = (m_t, I_t)$, the real and masked contexts preserve the question, candidate set, pre-action history, action JSON, timestamps, and probe. Only the current media return changes. The mask is deterministic and action-specific:

- `get_frames`: every returned image is replaced by a black RGB image with the same resolution; the number and ordering of images are kept;
- `get_audio`: the waveform is replaced by silence with the same duration, sampling rate, and channel layout;
- `get_clip`: both streams are retained, with black video frames at the same sampling rate and a silent audio track of the same duration.

The masks remove semantic content without inserting a textual description of what was removed. Media-decoding failures invalidate the complete comparison.

For temporal controls, we retain m_t and every non-temporal parameter, including the requested number of frames and the interval duration. Candidate placements are generated on a deterministic grid over the same source media. We reject a candidate whose overlap with the executed support exceeds 5%, shuffle the remaining candidates with a hash of a fixed seed and the logical action identity, and take the first K sufficiently separated placements. Point-valued frame queries use one frame period as their support. If fewer than K valid controls exist, the action is not temporally scored. Every control is paired with its own shape-matched mask, yielding $2(K + 1)$ branches for a scored action.

A.3 Answer probe and option canonicalization

Every branch ends with the same read-only instruction:

> Based only on the evidence available so far, answer the original multiple-choice question with exactly one option letter. Do not explain.

The instruction, decoding temperature, candidate set, and attended context length are identical across branches. The probe is used only to obtain the frozen model's next-token logits, is never sampled into a policy trajectory, and contains neither the gold answer nor future interaction. The gold label $y^\star$ is applied only after the forward pass to index the resulting logit vector; neither $y^\star$, the indexed score, nor the MATCA credit is appended to a branch context or exposed to the rollout policy.

Before scoring, option content is canonicalized and relabeled by a deterministic hash of the option text. The same permutation is used for every branch of the example. This removes a fixed answer-position prior while preserving the gold option content. Examples for which an option letter is not a single token under the scorer tokenizer are excluded.

Writing $z_y(C)$ for the frozen reference model's next-token logit for option y, the reported runs use $T = 1$ and compute

$$S_{\mathrm{ref}}(C; y^\star) = \log \frac{\exp z_{y^\star}(C)}{\sum_{y \in \mathcal{Y}} \exp z_y(C)}. \tag{15}$$

Candidate normalization removes any branch-wise shift common to all option logits, yielding a comparable gold-option support readout across the paired contexts.

A.4 Controlled interpretation and behavioral audit of content gain

MATCA's paired construction defines a controlled contrast for the content returned by an executed acquisition. We now connect its score direction to downstream task value and assess that relation with terminal agent behavior. Fix the pre-acquisition history $H_t = h_t$, the metadata of the executed action $A_t = a_t$, and a continuation policy π. Let E_t denote the media content returned by that action, let e_t be its realized content, and let $e_\varnothing$ be the corresponding structure-matched null. For the binary terminal correctness reward $R \in \{0, 1\}$, define

$$V_t^\pi(e) = \mathbb{E}_\pi[R \mid H_t = h_t, A_t = a_t, \mathrm{do}(E_t = e)] \tag{16}$$

and the additive controlled content effect

$$\Delta_t^{\mathrm{content}} = V_t^\pi(e_t) - V_t^\pi(e_\varnothing). \tag{17}$$

This estimand holds the chosen pathway, interval, budget, and prior interaction fixed and intervenes only on the semantic content of the returned media.

The paired intervention design fixes all continuation inputs outside E_t, preserves the action structure in $e_\varnothing$, and holds the continuation policy and remaining interaction budget fixed. This design defines the returned-media contrast in Eq. equation 17. Connecting the frozen readout to the direction of downstream task value uses the following sufficient calibration condition: for the paired branches, there is a branch-invariant strictly increasing map ϕ_t such that

$$S_{\mathrm{ref}}(C_t(e); y^\star) = \phi_t(V_t^\pi(e)), \qquad e \in \{e_t, e_\varnothing\}. \tag{18}$$

Directional consistency. Under the controlled paired design and the sufficient calibration condition,

$$\mathrm{sign}\big(G_t^{\mathrm{content}}\big) = \mathrm{sign}\big(\Delta_t^{\mathrm{content}}\big). \tag{19}$$

Thus, the score contrast and the controlled continuation-value contrast have the same direction under the stated calibration condition.

Proof. The real and masked branches instantiate $C_t(e_t)$ and $C_t(e_\varnothing)$, respectively. Substituting Eq. equation 18 into the definition of G_t^{content} gives

$$G_t^{\mathrm{content}} = \phi_t(V_t^\pi(e_t)) - \phi_t(V_t^\pi(e_\varnothing)). \tag{20}$$

Because ϕ_t is strictly increasing, this difference and $V_t^\pi(e_t) - V_t^\pi(e_\varnothing)$ have the same sign. □

The scale of G_t^{content} has a sharper interpretation under log value calibration. If $V_t^\pi(e) > 0$ in both branches and, for a branch-invariant offset b_t,

$$S_{\mathrm{ref}}(C_t(e); y^\star) = \log V_t^\pi(e) + b_t, \tag{21}$$

then

$$\begin{aligned} G_t^{\mathrm{content}} &= \log \frac{V_t^\pi(e_t)}{V_t^\pi(e_\varnothing)}, \\ \Delta_t^{\mathrm{content}} &= V_t^\pi(e_\varnothing)\left(\exp G_t^{\mathrm{content}} - 1\right). \end{aligned} \tag{22}$$

Under this stronger condition, the raw gain is the log ratio of the two controlled continuation values, rather than their additive difference. More generally, if $S_{\mathrm{ref}}(C_t(e); y^\star) = \log V_t^\pi(e) + b_t + \varepsilon_t(e)$ with $|\varepsilon_t(e)| \le \epsilon_t$ in both branches, then

$$\left| G_t^{\mathrm{content}} - \log \frac{V_t^\pi(e_t)}{V_t^\pi(e_\varnothing)} \right| \le 2\epsilon_t. \tag{23}$$

MATCA gate	N	$+1$	0	-1	Mean δR
Positive	70	18	48	4	$+0.200$
Abstain	53	5	43	5	0.000
Negative	57	4	39	14	-0.175
All	180	27	130	23	$+0.022$

Table 5: Terminal changes in the real-versus-null paired continuation audit, stratified by the fixed MATCA content-credit gate. All pairs, including the 130 zero-change outcomes, are retained; gate-stratified sampling directly tests the directional separation used by MATCA.

This bound motivates MATCA's calibrated gate: gains that clear the null threshold supply the directional credit used by the objective, while Eq. equation 22 characterizes their magnitude under log-value calibration. Thresholding, normalization, and clipping preserve the sign of every accepted nonzero credit. The paired continuation audit below evaluates this direction against terminal behavior, and Section B controls scorer variation across matched branches.

Operationally, MATCA assigns this executed-pathway content credit to the responsible action field rather than to unrelated tokens in the trajectory. Section F.1 complements this local estimand with same-state, same-window interventions over visual, acoustic, and joint pathways, directly testing whether differences between independently computed pathway-specific content credits predict terminal differences across executed and alternative pathways.

Outcome-level paired intervention audit. We evaluate the directional relation in Eq. equation 19 using 180 acquisition actions stratified by the frozen MATCA content gate. For each action, we replay the same pre-action history and executed action in two branches: the *real* branch retains the returned media, whereas the *null* branch replaces only its semantic content with the structure-matched mask used by MATCA. From that point onward, both branches use the same frozen deterministic continuation policy π_{det}, checkpoint, decoding configuration, remaining turn and tool budgets, and environment. Deterministic decoding yields one reproducible continuation per branch. If $R_i^{\mathrm{real}}, R_i^{\mathrm{null}} \in \{0, 1\}$ denote terminal correctness, the observed paired outcome is

$$\delta R_i = R_i^{\mathrm{real}} - R_i^{\mathrm{null}} \in \{-1, 0, +1\}. \tag{24}$$

Crucially, the answer probe, gold option $y^\star$, and MATCA score are used only by the frozen scorer and are never inserted into either continuation context. Each continuation generates its terminal answer without access to these values; only after generation is that answer compared with $y^\star$ to compute R_i. Thus, δR_i is an end-to-end behavioral endpoint separate from S_{ref}, and each pair is a fully specified downstream contrast for the returned content.

Table 5 exhibits the predicted ordering: the mean paired change is positive for positive credit, zero for abstention, and negative for negative credit. In the positive group, the confirmed and contradictory directions occur in $18/70 = 25.7\%$ (Wilson 95% CI $[16.9\%, 37.0\%]$) and $4/70 = 5.7\%$ ($[2.2\%, 13.8\%]$) of all pairs, respectively. In the negative group, $14/57 = 24.6\%$ ($[15.2\%, 37.1\%]$) are confirmed harmful changes and $4/57 = 7.0\%$ ($[2.8\%, 16.7\%]$) are contradictions. The corresponding two-sided exact McNemar tests give $p = 0.0043$ for positive credit (18 versus 4 discordant pairs) and $p = 0.0309$ for negative credit (14 versus 4). Among the 40 gated actions that produce a nonzero terminal change, $32/40 = 80.0\%$ agree with the predicted sign (Wilson 95% CI $[65.2\%, 89.5\%]$). This statistic is the directional agreement over the 40 gated pairs with an observable terminal flip; Table 5 reports all 180 pairs.

Table 6 shows action-dependent coverage. In particular, $23/35 = 65.7\%$ of audio acquisitions abstain, consistent with the calibrated gate suppressing weak acoustic evidence. The pooled audit provides the primary directional estimate, and the action-type rows make its coverage explicit. Credit also crosses the original terminal outcome: 36 positive-credit actions occur in originally failed trajectories, while 29 negative-credit actions occur in originally successful trajectories. This cross-outcome allocation demonstrates that MATCA credit is action-specific rather than inherited from trajectory reward. Together with the paired terminal contrasts, this distribution supports reinforcing useful acquisitions within failed trajectories while suppressing harmful or misleading acquisitions within otherwise successful trajectories.

Stratum	Pos.	Abst.	Neg.	Total
Action type				
Frames	30 (46.2%)	10 (15.4%)	25 (38.5%)	65
Audio	10 (28.6%)	23 (65.7%)	2 (5.7%)	35
Clip	30 (37.5%)	20 (25.0%)	30 (37.5%)	80
Original trajectory outcome				
Successful	34 (40.0%)	22 (25.9%)	29 (34.1%)	85
Failed	36 (37.9%)	31 (32.6%)	28 (29.5%)	95
All	70	53	57	180

Table 6: Composition of the 180-action audit by action type and original trajectory outcome. Parentheses give percentages normalized within each row; the final row gives marginal counts.

Delayed effects and diagnostic slices. The nonzero gated directional agreement is $14/18 = 77.8\%$ for short-horizon actions and $18/22 = 81.8\%$ when multiple interaction turns remain. Thus the directional relationship persists when multiple subsequent agent decisions mediate the terminal outcome. Agreement is $14/18$ for frame retrieval and $16/20$ for clip retrieval. The two observed nonzero audio cases both agree with the predicted direction, and the larger pooled audit provides the primary cross-modal estimate. Agreement is also $17/20$ for gated terminal flips drawn from originally unsuccessful trajectories and $15/20$ for those from originally successful trajectories, consistent with the motivation that useful acquisitions can occur in failed trajectories and harmful acquisitions in successful ones.

Outcome-level assessment of the training signal. Terminal correctness changes only when the final answer flips; the table therefore retains all $130/180 = 72.2\%$ unchanged endpoints, while the primary directional comparisons use confirmed-versus-contradictory flips within each gate. The opposing asymmetries of the positive and negative gates, the balanced abstention group, and the delayed-horizon agreement together provide direct outcome-level evidence that the gated sign of G_t^{content} tracks whether the returned content helps or harms terminal task completion under π_{det}. This is exactly the property consumed by MATCA, whose objective is defined on thresholded, sign-preserving credit. Together, these results support G_t^{content} as the directional controlled-content credit used by MATCA, including for acquisitions followed by multiple interaction turns; Eq. equation 22 gives its magnitude relation under log-value calibration.

B Null Calibration and Deterministic Score Selection

B.1 Null construction, thresholds, and scales

Calibration uses 128 held-out records that are disjoint from RL prompts and all reported evaluation sets. Thresholds are determined solely from these null records, independently of development-set accuracy. For each action type, the content null is built from differences between the fully masked executed placement and fully masked control placements. The temporal null uses pairwise differences among fully masked controls and therefore requires at least two controls per calibration record. Each null difference is inserted in both directions so that arbitrary ordering cannot create asymmetric noise.

For action type m and component $d \in \{\text{content}, \text{time}\}$, the positive threshold is the 95th percentile of positive null values and the negative threshold is the 99th percentile of negative-null magnitudes. The 99th-percentile gate provides conservative control of negative credit. A component is declared ready only when its null contains at least 64 samples. The scale is the 75th percentile of the absolute non-null effects accepted by the corresponding gate. All constants are frozen before the MATCA phase begins. Phase III repeats calibration after resetting the reference model to the consolidated student.

The normalized credit is

$$\rho_{m,d}(g) = \text{clip}\left(\frac{g}{s_{m,d} + \epsilon}, -1, 1\right). \tag{25}$$

Action	Comp.	n_0	τ^+	τ^-	s
Frames	Content	300	0.425	0.691	1.446
	Time	300	0.365	0.475	0.995
Audio	Content	156	0.613	1.875	2.568
	Time	156	0.315	1.023	2.125
Clip	Content	312	0.534	0.610	1.197
	Time	312	0.340	0.717	0.682

Table 7: Frozen one-sided null thresholds and robust scales. Positive and negative thresholds use the 0.95 and 0.99 null quantiles, respectively.

Table 7 reports the frozen Phase-I constants used by the submitted configuration. "Null n" counts the symmetric null samples, not the number of videos.

B.2 Deterministic selection under a scoring budget

We first group eligible actions by (data source, action type, trajectory outcome). Within a stratum, actions are ordered by a SHA-256 hash of the fixed seed and the logical action identity, rather than arrival order. We then take one action per stratum in round-robin order until the update budget B is exhausted. When a per-trajectory cap is enabled, it is enforced during this global traversal; enforcing it before stratification would systematically favor the first acquisition in a trajectory, which is often a frame query.

In the streaming implementation, the global budget is divided deterministically over prompt groups using the original prompt index. Unused capacity is assigned only after all groups are known, using the same hash order, which preserves selection independence from arrival time. The prompt identifier, trajectory identifier, turn index, action JSON, control placements, option permutation, and selected branches are written to an audit record so the decision can be reconstructed from the seed.

C Action-Field Routing and Loss Normalization

The response is parsed as JSON and each trainable action value is aligned to a half-open token span using tokenizer offset mappings. If offsets are unavailable, we use exact token-subsequence matching; ambiguous or missing matches are not guessed. The routing is:

- content credit C_t^{content} to the value of `type`;
- temporal credit C_t^{time} to the values of `start`, `end`, or `timestamps`;
- no MATCA credit to JSON keys, punctuation, fixed scaffolding, observations, reasoning, confidence, or the final answer.

When content is reliable but temporal gain is inside the calibrated null band, the timestamp fields receive a zero local advantage. This is distinct from abstention: a zero prevents outcome credit from declaring the same timestamp good or bad, whereas an abstained field remains on the outcome branch. If content credit is negative, the content gate clips temporal credit above at zero.

For the submitted objective, locally credited tokens and outcome tokens form the complementary masks defined in the main text, and the two branches are normalized independently. Consistent with Eq. (13) in the main paper, Avg_M is the token mean over the active entries of mask M:

$$\mathrm{Avg}_M[x] = \frac{\sum_{i,u} M_{i,u} x_{i,u}}{\sum_{i,u} M_{i,u}}. \tag{26}$$

Accordingly, $\mathcal{L}_{\text{local}}$ averages all active M^{local} tokens and $\mathcal{L}_{\text{out}}$ averages all active M^{out} tokens; no additional field-, action-, turn-, or trajectory-level averaging is applied. Separate normalization keeps the sparse action-value objective from being diluted by the larger outcome-token set while matching the loss definition used in the main paper.

D Training Data, Streaming Verification, and Cost

Agentic SFT data construction. We construct the Agentic SFT data from seed questions sampled exclusively from the training splits of LongVideo-Reason, Video-Holmes, VSI-Train-10k, LongVALE, and MultiHop-EgoQA. For each seed question, best-of-N exploration with self-correction produces multi-turn Observation–Thought–Action trajectories. Outcome verification then retains task-successful episodes, and a rationality audit removes unsupported interaction traces. The resulting curated mixture contains approximately 19.8K trajectories and is enriched with joint audio-visual trajectories containing `get_clip`, while retaining visual-only `get_frames` and audio-only `get_audio` exploration. Only training splits are used, and the final mixture is the common Agentic SFT initialization for all controlled post-training variants.

MATCA execution. The submitted training uses 48 NVIDIA H100 GPUs arranged as six eight-GPU nodes. Four nodes hold the policy and perform rollout generation, old-log-probability computation, and policy updates; two nodes hold a frozen copy of the reference weights and execute counterfactual scoring. A completed prompt group contains all $N = 8$ trajectories for one prompt. It is assigned its deterministic action budget and placed on a bounded scorer queue while actors continue generating the remaining groups. Small groups are coalesced for at most one second and media preprocessing is bounded to avoid an unbounded CPU-memory backlog. Decoded media, masks, and scorer-only clip proxies are cached by logical action and branch.

The scorer may overlap with rollout and old-log-probability computation, but the policy update begins only after all 32 prompt groups and their selected scores have been merged into the final batch. A scorer result is accepted only when its prompt, trajectory, and turn identifiers match a row generated under the current π_{old}. No score or trajectory crosses a policy-version boundary, so the optimization remains synchronous and on-policy.

With B scored actions and K controls per action, an update contains at most

$$N_{\text{ctx}} = 2B(K + 1) \tag{27}$$

multimodal scorer contexts. The final launcher uses $K = 3$, $B \leq 1024$, a 32,768-token scorer-context limit, 2-fps scorer clip proxies, at most eight actions per scorer call, and a scorer micro-batch of two contexts per GPU. Thus the configured upper bound is 8,192 contexts per update, although invalid or ineligible actions reduce the realized count. Each context performs one frozen forward pass and reads all answer candidates at the final position.

The training schedule contains approximately 100 policy updates in Phase I and 50 in Phase III. With 32 prompt groups and eight trajectories per prompt, this corresponds to approximately 25,600 and 12,800 generated trajectories, respectively. The observed end-to-end throughput is about four updates per four hours (approximately one update per hour), including rollout, counterfactual scoring, and optimization. We therefore estimate Phase I at approximately 100 wall-clock hours on 48 H100 GPUs (4,800 H100-hours) and Phase III at approximately 50 hours (2,400 H100-hours). Excluding the intervening SIC stage, the two online MATCA phases total approximately 150 stage-hours and 7,200 H100-hours. These rounded stage-level values are estimated from the observed end-to-end throughput. Phase-II SIC compute is reported separately in Section E.5.

For device-aware cost accounting, we additionally log scorer forward time, preprocessing time, queue wait, tail wait after rollout completion, selected-action count, reliable-action count, and peak memory. These measurements complement context counts because frame, audio, and clip inputs contain different numbers of multimodal tokens and benefit differently from caching and prefix reuse. MATCA adds no trainable parameters and no inference-time scorer, mask, control branch, or tool call.

E Selective Intervention Consolidation

Selective Intervention Consolidation (SIC) is Phase II of the complete training procedure. We freeze the final Phase-I MATCA policy as the teacher π_T, reset the student π_ϕ to the original Agentic SFT checkpoint, and optimize the student on a fixed replay pool of intermediate interaction states. At each state, the teacher and student propose alternative actions under the same student-generated prefix; the current student completes both resulting branches, and the teacher action is distilled

Training setting	Value
Backbone	Qwen2.5-Omni-7B
Prompt groups per update	32
Trajectories per prompt (N)	8
Maximum interaction turns	22
Actor/scorer nodes	$4/2$
GPUs per node	8
GPU type	NVIDIA H100
Actor learning rate	1×10^{-6}
Local-loss coefficient	0.2
Content/time multipliers	$1/1$ after normalization
KL/entropy coefficients	$0/0$
Efficiency credit	Disabled
Matched controls in RL (K)	3
Control/selection seed	20260721

Table 8: Core settings shared by the submitted MATCA runs.

only when its branch obtains a strictly higher terminal return. The consolidated checkpoint π_C then initializes Phase-III MATCA. The frozen MATCA belief readout π_{ref} is not used in the SIC acceptance decision.

SIC is offline only with respect to state acquisition: the replay pool is fixed before optimization. Student prefixes and actions, teacher candidate actions, student-completed continuations, terminal returns, and acceptance decisions are generated online by their respective policies rather than precomputed.

E.1 Replay-State Pool Construction

Sources and composition. The replay pool is drawn from the complete training-only Agentic SFT corpus described in Section D and contains no examples from any validation or evaluation split. After applying the eligibility criteria below, the fixed pool contains approximately 5,000 replay states. We sample these states uniformly throughout Phase II without source-level reweighting.

State reconstruction. Each replay entry stores the task and media metadata needed to initialize the standard OmniAgent environment, including the question, answer format and options when applicable, video identifier, duration, frame rate, and audio availability. It also stores the fork turn and the exact ordered sequence of preceding structured Observation–Thought–Action responses. To reconstruct $\bar{H}_i$, we reset the environment from the task metadata and replay these prefix responses verbatim. The environment thereby regenerates the associated tool outputs, restores the conversation history and residual turn budget, and then applies the standard OmniAgent chat template.

Following OmniAgent's memory-consolidation protocol, the most recent retrieved frames, audio segment, or video clip remains available as multimodal input, whereas older raw media are replaced by a textual omission marker after their content has been summarized in the subsequent agent response. The previous textual Observation–Thought–Action trace is retained. Ground-truth answers and replay identifiers are kept on the environment side for reward computation and bookkeeping and are not exposed as policy input.

Sampling and eligibility. We retain only nonterminal states for which the interaction environment can be restored, the historical action sequence is valid, the required media remain available, and at least one subsequent executable decision can be evaluated. The pool size reported above is measured after applying these eligibility criteria.

Pool reuse. The resulting replay pool remains fixed throughout SIC. Sampling a replay state restores its interaction history, residual turn budget, and per-call media limits, but all subsequent student prefixes and actions, teacher candidate actions, student continuations, rewards, and gate decisions are recomputed online by their respective policies.

E.2 Paired Intervention and Acceptance Gate

Paired branch construction. At replay state $\bar{H}_i$, the current student first generates a pre-action prefix

$$p_i^S = (o_i^S, z_i^S, c_i^S). \tag{28}$$

Conditioned on the same pair $(\bar{H}_i, p_i^S)$, the current student and frozen teacher sample actions a_i^S and a_i^T, respectively. Conditioning the teacher on the student prefix prioritizes actions compatible with states the student can produce, although it can miss actions available only under teacher-native reasoning.

We create independent copies of the post-prefix decision state, insert a_i^S into one copy and a_i^T into the other, and let the current student complete both trajectories:

$$\tau_i^{S\to S} \quad \text{and} \quad \tau_i^{T\to S}. \tag{29}$$

The candidate action is the only systematic difference between the branches. Their continuations are independent samples from the current student under the same decoding configuration.

Action agreement and answer actions. Both candidate actions are parsed before branch rollout. A pair is discarded if either action is invalid or unexecutable, or if the parsed actions have the same type and values; differences only in whitespace or JSON formatting do not create an intervention. The `answer` action remains in the candidate space whenever it is valid at the reconstructed turn, allowing SIC to transfer both evidence-acquisition decisions and decisions to stop and answer.

Environment and continuation budgets. For a video of duration D seconds, the environment sets the trajectory budget to

$$N_{\max}(D) = \min(5 + \lfloor D/60 \rfloor\,,\ 22)\,. \tag{30}$$

The replayed history has already consumed part of the turn budget, so both branches inherit the same residual turn budget and per-call media limits at the fork. A `get_frames` call may return at most 60 frames, a `get_audio` call may cover at most 300 seconds, and a `get_clip` call may cover at most 60 seconds. Candidate-action and continuation decoding settings are reported in Table 10.

Validity and normal termination. An action is valid only if exactly one supported structured action can be parsed and executed. Its type and required fields must match the environment schema: `get_frames`, `get_audio`, `get_clip`, or `answer`. Media intervals must be numeric, ordered, within the video duration and per-call limits, and compatible with the available modalities and residual budgets. An `answer` must contain a parseable answer value. Schema errors, invalid timestamps, unavailable modalities, exhausted budgets, and execution errors invalidate the action.

A branch terminates normally when it executes a valid `answer` action within its inherited turn and token budgets; the answer may be correct or incorrect. Timeout, maximum-turn or token-limit termination without a valid answer, environment exceptions, media-tool failures, decode failures, and unparseable final answers are abnormal terminations. A pair is discarded if either branch terminates abnormally.

Terminal return and gate. The gate uses the binary terminal task-success reward $R(\tau) \in \{0, 1\}$: $R(\tau) = 1$ iff the environment marks the final answer as correct, and $R(\tau) = 0$ for a normally terminated but incorrect answer. Let v_i^S and v_i^T indicate valid, executable student and teacher actions, and let $n_i^{S\to S}$ and $n_i^{T\to S}$ indicate normal termination. A teacher intervention is accepted iff

$$\begin{aligned}\gamma_i = \mathbb{1}\big[&v_i^S \wedge v_i^T \wedge n_i^{S\to S} \wedge n_i^{T\to S} \\ &\wedge R(\tau_i^{T\to S}) > R(\tau_i^{S\to S})\big].\end{aligned} \tag{31}$$

Under the binary reward, this strict comparison is equivalent to the teacher-action branch being correct and the student-action branch being incorrect. The return magnitude is not used as a token-level weight.

Randomness. The run uses base seed 42 for replay sampling, environment stochasticity, and rollout generation. The two branch continuations are drawn from distinct random substreams derived from that base seed, so they are independent samples of the current student and the run remains reproducible.

Action	Supervised value fields
`get_frames`	type, start, end, num
`get_audio`	type, start, end
`get_clip`	type, start, end
`answer`	type, content

Table 9: Executable action values used to construct D_i^{act}. The corresponding JSON keys and syntactic delimiters are not supervised.

E.3 Action-Only Distillation

For an accepted intervention, we teacher-force the student along the accepted teacher action and supervise only its executable action-value positions D_i^{act}. For teacher-generated token $w_{i,u}^T$,

$$\ell_{i,u}^{\text{SIC}} = -\log \pi_\phi \left(w_{i,u}^T \mid \bar{H}_i, p_i^S, w_{i,<u}^T \right). \tag{32}$$

The batch loss is

$$\mathcal{L}_{\text{SIC}} = \frac{\sum_i \gamma_i \sum_{u \in D_i^{\text{act}}} \ell_{i,u}^{\text{SIC}}}{\sum_i \gamma_i |D_i^{\text{act}}|}. \tag{33}$$

Mini-batches with no accepted intervention are skipped. Keys, delimiters, response scaffolding, observations, reasoning, confidence, and continuation tokens do not receive the SIC loss.

This is hard-target negative log-likelihood, as specified in the main paper. SIC does not match the teacher's full or truncated soft token distribution and does not use a top-k forward-KL objective: supervision targets the accepted executable action tested by the paired branches.

Token mask and empty batches. We obtain the mask by parsing the teacher turn, locating the character spans of values inside its top-level `action` object, and mapping fully contained response-token offsets to D_i^{act}. Prompt tokens, right padding, JSON keys, and boundary tokens shared with syntax are excluded. The teacher action is used at its own token length and need not align with the rejected student action. The loss is normalized once over all active action-value tokens in the distributed batch rather than first by action or trajectory.

When the global active-token count is zero, we skip the backward pass, the optimizer step, and the learning-rate-scheduler step on every worker.

E.4 Training Configuration

The checkpoint identifiers are defined by training provenance rather than machine-local paths: the teacher is the final Phase-I MATCA policy, and the student is reset to the same Agentic SFT checkpoint used to initialize Phase I. Table 10 summarizes the remaining Phase-II configuration.

E.5 Control Baselines and Compute Matching

Ungated distillation. Ungated distillation resets the student from the same Agentic SFT checkpoint and uses SIC's replay pool, frozen teacher, student-generated prefix, action-value mask, NLL objective, optimizer, and one-epoch schedule. The teacher proposes one action at each state, and every parseable and executable teacher action is distilled, including one identical to the student's choice; no agreement filter is applied. This baseline generates neither a competing student action nor the two terminal branches, so it is cheaper than SIC and tests unconditional rather than outcome-filtered teacher-action transfer.

Random-gate distillation. Random-gate distillation uses the same initialization, replay pool, teacher, student-prefix conditioning, candidate-action decoding, eligibility rules, action-value objective, and optimizer as SIC. It retains the student–teacher action disagreement filter but replaces the branch-return comparison with a reward-independent random selection over complete eligible teacher actions; it does not generate paired continuations or observe branch returns.

Algorithm 1 Selective Intervention Consolidation

Require: Fixed replay pool $\mathcal{B}$, frozen teacher π_T, student π_ϕ reset from Agentic SFT
for each SIC optimization batch **do**
  Sample reconstructed states $\{\bar{H}_i\}$ from $\mathcal{B}$
  for each $\bar{H}_i$ **do**
    Generate student prefix p_i^S
    Sample $a_i^S \sim \pi_\phi(\cdot \mid \bar{H}_i, p_i^S)$
    Sample $a_i^T \sim \pi_T(\cdot \mid \bar{H}_i, p_i^S)$
    Parse and validate both candidate actions
    if either action is invalid, unexecutable, or identical after parsing **then**
      Reject the pair and continue
    end if
    Execute the actions in independent state copies
    Let the current student complete both branches
    if either branch terminates abnormally **then**
      Reject the pair and continue
    end if
    Compute γ_i using Eq. equation 31
    if $\gamma_i = 1$ **then**
      Add the teacher action-value tokens to the training batch
    end if
  end for
  if the global batch contains accepted action-value tokens **then**
    Update ϕ using Eq. equation 33
  else
    Skip the optimization step
  end if
end for
Ensure: Consolidated student π_C

Matching is performed separately at every Phase-II update. Let n_u^{SIC} be the number of complete teacher actions accepted by SIC at update u. We record only this count, not the accepted action identities or branch returns. The control follows the same replay-state order as SIC. From the eligible student–teacher disagreements in the corresponding random-gate update, we select $\min(n_u^{\mathrm{SIC}}, |\mathcal{E}_u|)$ actions uniformly without replacement, where $\mathcal{E}_u$ denotes the eligible-action set. Selection is implemented by sorting actions with a deterministic hash of the common base seed 42, the update index, and the replay-state identifier, which defines an order-independent random stream separate from replay sampling and decoding. If $n_u^{\mathrm{SIC}} = 0$, the corresponding optimization update is skipped.

Selection operates on whole actions, so no action is truncated or fractionally weighted. This construction matches SIC per update in the number of selected actions and in the pattern of nonempty optimization updates whenever the eligible set has sufficient capacity. In the matched run, both discrepancies are zero. Because complete actions can contain different numbers of value tokens, the aggregate number of supervised action-value tokens differs by 0.8% between random-gate distillation and SIC.

Compute-matched continued MATCA. The continued-MATCA baseline starts from the Phase-I MATCA policy and receives additional online optimization to match the complete post-Phase-I SIC path, defined as Phase-II SIC plus Phase-III MATCA. We measure this budget in accelerator-hours:

$$C_{\mathrm{full}} = G_{\mathrm{II}} T_{\mathrm{II}} + G_{\mathrm{III}} T_{\mathrm{III}}, \tag{34}$$

where G and T denote the GPU count and wall-clock training time of each phase. Continued MATCA runs from the same Phase-I policy until its accelerator-hours match C_{full}. Phase-II SIC uses 16 H100 GPUs for approximately 30 wall-clock hours, or 480 H100-hours. Phase-III MATCA uses 48 H100 GPUs for approximately 50 hours, or 2,400 H100-hours. Thus, the complete post-Phase-I SIC path consumes approximately 2,880 H100-hours. The compute-matched continued-MATCA baseline uses 48 H100 GPUs for approximately 60 hours, also totaling 2,880 H100-hours; the discrepancy is 0.0% under this rounded stage-level accounting.

Item	Final setting
Replay pool	Approximately 5,000 states from the complete training-only Agentic SFT corpus; uniform state-level sampling without source reweighting
Candidate actions	Student: $T = 1.0$, top-$p = 1.0$, $k = -1$, one sample, at most 4096 response tokens; teacher: the same sampling settings, one sample, at most 512 new tokens
Branch continuations	Current student with $T = 1.0$, top-$p = 1.0$, $k = -1$, one sample per turn, and at most 4096 response tokens
Objective and duration	Action-value-token NLL, normalized over accepted tokens; one epoch over the fixed replay pool
Optimizer and schedule	AdamW, learning rate 1×10^{-7}, $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01, no warmup, and constant learning rate
Batching and clipping	Global batch 14, mini-batch 14, dynamic token batching with at most 69,632 tokens per GPU, and global-norm clipping at 0.05
Sequence and precision	65,536 prompt, 4,096 response, and 81,920 model tokens; BF16 execution with FP32 reductions; remove-padding disabled
Trainable components	Student language-model parameters are updated; visual and audio encoders and the teacher are frozen
Parallelism and inference	FSDP student optimization with parameter and optimizer offload; vLLM student rollouts and Hugging Face teacher inference
Hardware	Phase-II SIC: 2×8 NVIDIA H100 GPUs (14 student/rollout ranks and two teacher replicas)
Software	VERL 0.9.0.dev (local fork), PyTorch 2.10.0, Transformers 5.5.4, vLLM 0.19.0, Ray 2.55.0, and CUDA 12.8
Randomness	Seed 42 for replay sampling, rollouts, and training

Table 10: Phase-II SIC training configuration.

E.6 Scope of Consolidation

SIC evaluates action transfer at replayed states using student-prefix conditioning and student-completed branches. The fixed replay pool focuses consolidation on executable interventions compatible with the reset student, which then undergoes renewed on-policy refinement in Phase-III MATCA.

F Capability Labels and Additional Diagnostic Protocols

We define a *modality-sensitive* item as one whose official task annotation requires audio or synchronized audio-visual evidence, and a *temporal-sensitive* item as one whose annotation requires event location, ordering, duration, or cross-time comparison. An item may belong to both subsets; items belonging to neither are excluded from both rather than assigned by default. The mapping is fixed from official benchmark taxonomies before comparing methods. All aggregate subset scores are sample-level accuracies, not unweighted means of category accuracies.

The full audit export retains, for every benchmark item, its official identifier, official category, the two binary labels, prediction, and correctness. This permits recomputation of every aggregate and avoids hiding low-frequency categories. For LVOmniBench, duration bins are 10–30, 30–50, 50–70, and 70–90 minutes and contain 500, 350, 120, and 44 items, respectively; the "All" column in the main paper is sample-weighted.

F.1 Input-pathway ablation and cross-pathway interventions

Structure-matched clip decomposition. We first decompose the gated content credit within executed `get_clip` actions. From 84 deterministically selected actions, two were excluded for unavailable media and two for branch-structure mismatch, leaving 80 valid clips. For every clip, the four conditions preserve duration, frame timestamps, resolution, audio length, sample rate, and channel count: S_0 uses matched null frames and silence, S_V restores only the real visual stream, S_A restores only the real audio stream, and S_{AV} restores both. All 80 silent-audio tracks and visual

Quantity	N	Mean	Med.	95% CI	> 0	< 0
Φ_V	80	$+0.065$	$+0.052$	$[-0.070, +0.195]$	53	27
Φ_A	80	-0.002	$+0.002$	$[-0.100, +0.080]$	46	34
I_{AV}	80	-0.170	-0.015	$[-0.410, +0.005]$	30	50
D_{AV}	80	-0.220	-0.006	$[-0.445, -0.080]$	29	51

Table 11: Stream-level decomposition on 80 structure-matched `get_clip` interventions. No quantity is exactly zero.

Path	Corr.	Acc. (95% CI)	Media	Total	N-dom.
V	29/60	48.3% [36.2, 60.7]	1920	8450	38/60
A	26/60	43.3% [31.6, 55.9]	1280	7730	36/60
AV	34/60	56.7% [44.1, 68.4]	3100	9610	42/60

Table 12: Terminal outcomes and realized mean costs for matched pathway interventions. "Non-dom." denotes sample-wise Pareto non-domination in terminal correctness and media-token cost.

tensors passed the shape audit. Define

$$\Phi_V = \tfrac{1}{2}[(S_V - S_0) + (S_{AV} - S_A)], \tag{35}$$

$$\Phi_A = \tfrac{1}{2}[(S_A - S_0) + (S_{AV} - S_V)], \tag{36}$$

with interaction $I_{AV} = S_{AV} - S_V - S_A + S_0$ and complementarity $D_{AV} = S_{AV} - \max(S_V, S_A)$. The Shapley efficiency identity $\Phi_V + \Phi_A = S_{AV} - S_0$ holds for every valid clip, with maximum absolute error below 10^{-12}.

Table 11 shows heterogeneous stream credit: visual evidence is dominant in $44/80 = 55.0\%$ of clips and audio evidence in $36/80 = 45.0\%$. The complete clip exceeds its better single-stream condition in $29/80 = 36.3\%$ of cases. The decomposition also follows MATCA's gate. For positive, abstaining, and negative gates (40, 27, and 13 clips), the mean (Φ_V, Φ_A) is $(+0.290, +0.140)$, $(+0.130, -0.070)$, and $(-0.760, -0.300)$, respectively. The corresponding mean interaction is $-0.130, -0.310, -0.005$, and mean D_{AV} is $-0.010, -0.280, -0.740$. Thus the attribution exposes which stream supplies the gated clip credit, while retaining sample-specific audio–visual complementarity rather than imposing a globally preferred stream.

Same-state, same-window pathway interventions. We next test cross-pathway predictive validity on 60 frozen pre-action states, with 20 originally selected from each of `get_frames`, `get_audio`, and `get_clip`. Each state is replayed through visual (V), acoustic (A), and joint (AV) branches. The branches share the identical pre-action history and temporal window; the AV visual timestamps and waveform are exactly those used in the corresponding single-stream branches. Each performs one acquisition and then uses the same frozen deterministic continuation policy, decoding configuration, and remaining budget. Gold labels and all scorer outputs remain outside the continuation context. Let $R_i(m) \in \{0, 1\}$ be terminal correctness for pathway m and let $\widetilde{G}_i^m$ be its robustly normalized real-versus-null content credit.

The paired terminal tables (both correct, first only, second only, both wrong) are $(13, 16, 13, 18)$ for V/A, $(18, 11, 16, 15)$ for V/AV, and $(17, 9, 17, 17)$ for A/AV; the corresponding two-sided exact McNemar p-values are 0.711, 0.442, and 0.169. Consequently, no single pathway explains the sample-level differences by a uniform marginal advantage.

As Table 13 shows, all three pathway comparisons agree with terminal flips above the 50% chance direction. Across the three pair views, the pooled directional agreement is $48/72 = 66.7\%$; a state-clustered bootstrap gives a 95% interval of approximately $[54\%, 77\%]$, accounting for the dependence among the three comparisons from the same state. This directly extends the executed-pathway result in Section A.4: the relative credit also predicts which of two pathway interventions will improve the terminal answer under the fixed continuation.

Pathway selection and cost. The pathway with maximum normalized credit is the unique successful pathway in $13/20 = 65.0\%$ of unique-best states and is weakly optimal (allowing terminal

Pair	$\Delta R \neq 0$	Ties	Agree	Agree. (95% CI)	p
V/A	29	4	16	64.0% $[44.5, 79.8]$	.230
V/AV	27	3	17	70.8% $[50.8, 85.1]$	.064
A/AV	26	3	15	65.2% $[44.9, 81.2]$	.210

Table 13: Agreement between the sign of the normalized pathway-credit difference and the observed terminal contrast. Agreement excludes terminal and credit ties; all 60 states enter each pairwise audit. Confidence intervals are Wilson intervals and p-values are two-sided exact sign tests.

ties) in $49/60 = 81.7\%$ of all states. Its mean binary terminal regret is $11/60 = 0.183$. The originally executed pathway is weakly optimal in $43/60 = 71.7\%$, with regret $17/60 = 0.283$; strict and tie-aware agreement between the original policy and credit ranking are $30/60 = 50.0\%$ and $35/60 = 58.3\%$. Using realized media-token costs, the original pathway is strictly Pareto dominated in $17/60 = 28.3\%$ of states, compared with $11/60 = 18.3\%$ for the credit-ranked pathway; the latter remains non-dominated in $49/60 = 81.7\%$. An equal-or-better terminal result is available at lower cost in $23/60 = 38.3\%$ of states. Together, these paired results support the relative predictive validity of independently computed executed-pathway content credits while separately measuring cost efficiency rather than folding an arbitrary cost weight into the credit.

F.2 Long-form audio-visual evaluation

LVOmniBench is evaluated with identical maximum turns, tool-call budget, requested frame count, audio duration, and media-token budget for the matched OmniAgent and MATCA policies. Actions per question counts executed perception calls. Visual frames and audio seconds count returned streams, so `get_clip` contributes to both. Failed tool calls are reported separately and do not contribute decoded media. The duration-bin accuracy and cost values in the main paper are computed directly from item-level logs with no macro-averaging across bins.

G Positioning and Scope of Acquisition Credit

G.1 Relation to concurrent credit-assignment methods

Returned-media scoring. MATCA builds on local acquisition rewards, belief-based progress signals, and counterfactual credit assignment. Video-MTR rewards intermediate segment selection using temporal relevance annotations (Xie et al., 2025); GiGPO and TAPO construct local comparisons from repeated states or similar information-acquisition calls (Feng et al., 2025; Dong et al., 2026). ΔBelief-RL, IGPO, and TRACE evaluate changes in target-answer support across interaction prefixes (Auzina et al., 2026; Wang et al., 2026; Tao et al., 2026b), while SIOP uses beliefs over semantic outcome clusters (Hu et al., 2026). TRACE evaluates this state transition at tool-interaction boundaries. MATCA's distinguishing comparison holds the history and executed action fixed and substitutes the returned media, scoring real and structure-matched null returns before the next policy response.

Multimodal interventions and token-level credit are likewise established: CFPO perturbs visual representations, SRPO uses visual corruption for role-aware weighting, and PGPO measures token-level visual dependence (Yu et al., 2026; Jiang & Zou, 2026; Ye et al., 2026); POAD decomposes action credit over tokens, while APPO uses frame attention to guide perception-token rewards (Wen et al., 2024; Du et al., 2026). MATCA specializes these broader ideas to structured acquisition actions. It separates a real–null content contrast from a selected–control temporal contrast and routes their signed, gated credits to `type` and timestamp value fields. Content credit concerns the executed pathway's return, not its advantage over alternative pathways; temporal credit is relative to the median of sampled matched placements, not the globally best interval. The frozen readout supplies a null-calibrated directional proxy. The paired continuation audit evaluates its thresholded content-credit direction under a fixed deterministic continuation policy (Section A.4).

Branch-compared consolidation. The main paper's second innovation is SIC: candidate teacher actions are consolidated only when they achieve higher realized return on the reset student's own

continuation distribution. SAGE-OPD selects multi-turn teacher interventions using environment feedback, teacher judgment, and teacher-confidence weighting Zhou et al. (2026), while RG-OPD uses verifier reward to gate teacher-logit supervision on student trajectories Akhondzadeh et al. (2026). SIC performs a paired action-level transfer test at the same student-generated prefix. Teacher and student propose alternative executable acquisition actions, the current student completes both branches under matched budgets, and the teacher action is supervised only when its branch achieves a strictly higher realized terminal return. Supervision is restricted to the executable action-value tokens before renewed MATCA optimization. The defining SIC contribution is therefore branch-compared transfer of a concrete acquisition action into a reset student. The returned-media and cross-pathway audits in Sections A.4 and F.1, together with the ungated and random-gate controls in Section E.5, respectively support MATCA's acquisition-local credit and SIC's selective transfer.

G.2 Counterfactual readout scope

MATCA uses the frozen reference as a consistent task-conditioned readout. Candidate normalization makes branch scores comparable, while one-sided null gates and abstention restrict local credit to effects that exceed matched-branch variation. The paired continuation audit in Section A.4 then evaluates the selected credit direction against terminal agent behavior.

The structure-matched null fixes the action pathway, media shape, and duration, so the resulting estimand isolates the controlled contribution of the returned content under a reproducible intervention. Applying the same masking rule to the executed placement and its matched controls further cancels common offsets in the temporal difference. The deterministic null also prevents uncontrolled answer evidence from entering the reference branch and supports exact matching of pathway, shape, and duration, providing a consistent and auditable comparison throughout training.

Temporal credit ranks the executed interval against K deterministically sampled feasible placements. Conservative thresholds, strict validity checks, and explicit coverage reporting ensure that a nonzero credit is assigned only when the evaluated comparison is supported. SIC analogously uses a matched paired continuation under fixed budgets, keeping its consolidation criterion aligned with the same controlled-intervention principle.

H Case Study

We present three qualitative case studies illustrating how MATCA assigns factorized credit to acquisition type and temporal placement. The first case shows a useful acquisition embedded in a failed trajectory, the second shows a misleading acquisition embedded in a successful trajectory, and the third shows an acquisition for which both the returned content and the selected temporal interval are harmful.

Coverage of credit configurations. The examples cover three outputs of the same hierarchical credit rule. Figure 3 has $(A_{\text{out}}, C_{\text{content}}, C_{\text{time}}) = (-0.725, +0.522, 0)$: useful content reverses the trajectory sign, while temporal credit abstains. Figure 4 has $(+1.620, -1.000, 0)$: a misleading acquisition is suppressed inside a successful trajectory without a timestamp penalty. Figure 5 has $(-0.935, -0.540, -1.000)$: both returned content and temporal placement are harmful. The first two visualize the cross-outcome allocation in Table 6 (36 positive actions in failed trajectories and 29 negative actions in successful ones); the third makes the independent temporal decision explicit.

How to read the diagrams. Each figure uses the same five-panel audit layout: Panel 1 gives the question and terminal outcome; Panel 2 locates the acquisition on the media timeline; Panel 3 preserves the audited action, returned media, and subsequent trajectory; Panel 4 shows the real–null content contrast and matched temporal controls; and Panel 5 maps calibrated credit to the executable JSON fields while identifying tokens that retain outcome supervision. The fixed layout makes the intervention, score change, gate, and optimization target directly traceable.

From score contrasts to field-level supervision. Panels 4 and 5 express the main paper's field-level claim. The real–null contrast supervises the `type` value; conditional on reliable content credit, the selected–control contrast separately supervises `start` and `end`. A comparison inside the calibrated null region leaves timestamps unchanged even when content receives a strong signed update.

All other tokens retain trajectory supervision. MATCA can therefore reverse local acquisition credit relative to terminal success while independently reinforcing, suppressing, or abstaining on pathway and temporal fields.

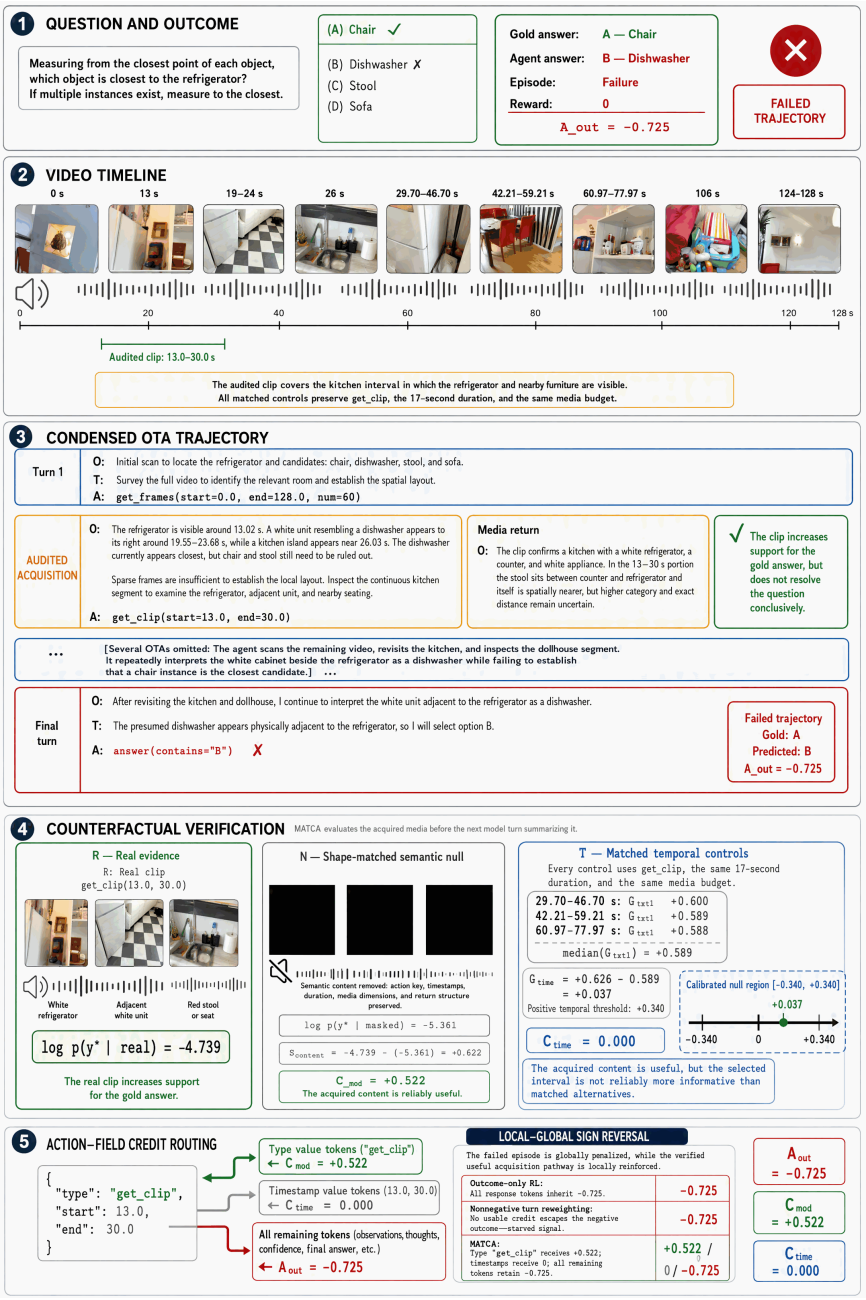

Figure 3: **MATCA reinforces a useful acquisition pathway in a failed trajectory without over-crediting its temporal placement.** The agent ultimately predicts *dishwasher* instead of the gold answer *chair*, resulting in a negative outcome advantage of -0.725. Nevertheless, the executed `get_clip(13.0, 30.0)` action increases gold-answer support relative to its structure-preserving masked return, yielding a positive content gain of 0.626 and a positive executed-pathway content credit of 0.522. Its gain is only marginally higher than the median gain of matched same-pathway, same-duration controls, placing the temporal-specificity signal in the calibrated null region. MATCA therefore positively supervises the `type` value while assigning zero credit to `start` and `end`. Outcome-only RL and TAURA cannot produce this local sign reversal because both retain the negative trajectory-level sign.

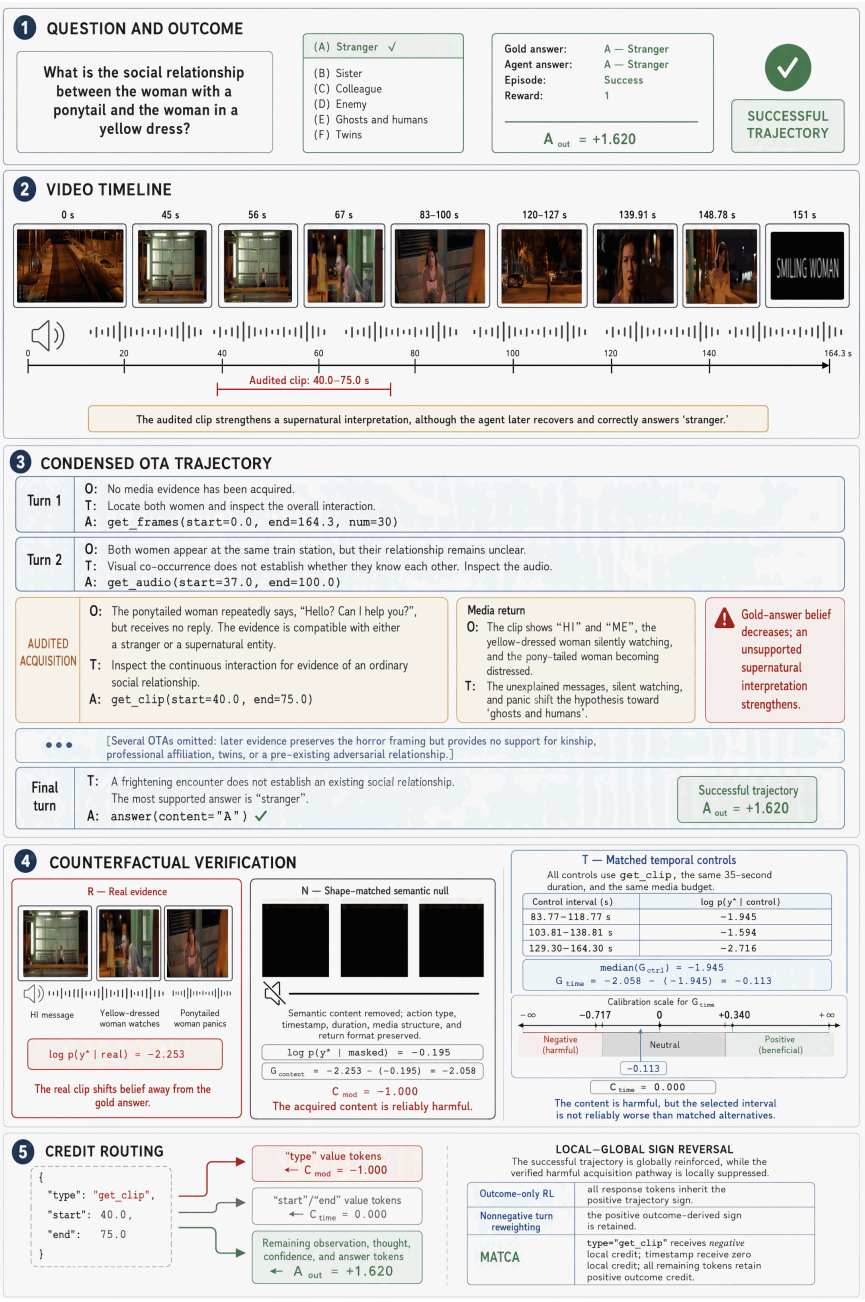

Figure 4: **MATCA suppresses a misleading acquisition inside an otherwise successful trajectory.** The trajectory eventually selects the correct relationship, *stranger*, and therefore receives a positive outcome advantage of 1.620. However, the acquired clip from 40.0–75.0 s shifts the frozen readout away from the gold answer by emphasizing the characters' fearful and apparently supernatural interaction. Relative to a structure-matched masked return, the clip produces a content gain of −2.058 and receives an executed-pathway content credit of −1.000. Its temporal-specificity difference lies within the calibrated null region, so the timestamp values receive zero credit. MATCA thus penalizes only the responsible acquisition-type field while preserving positive outcome supervision for the remaining reasoning and answer tokens.

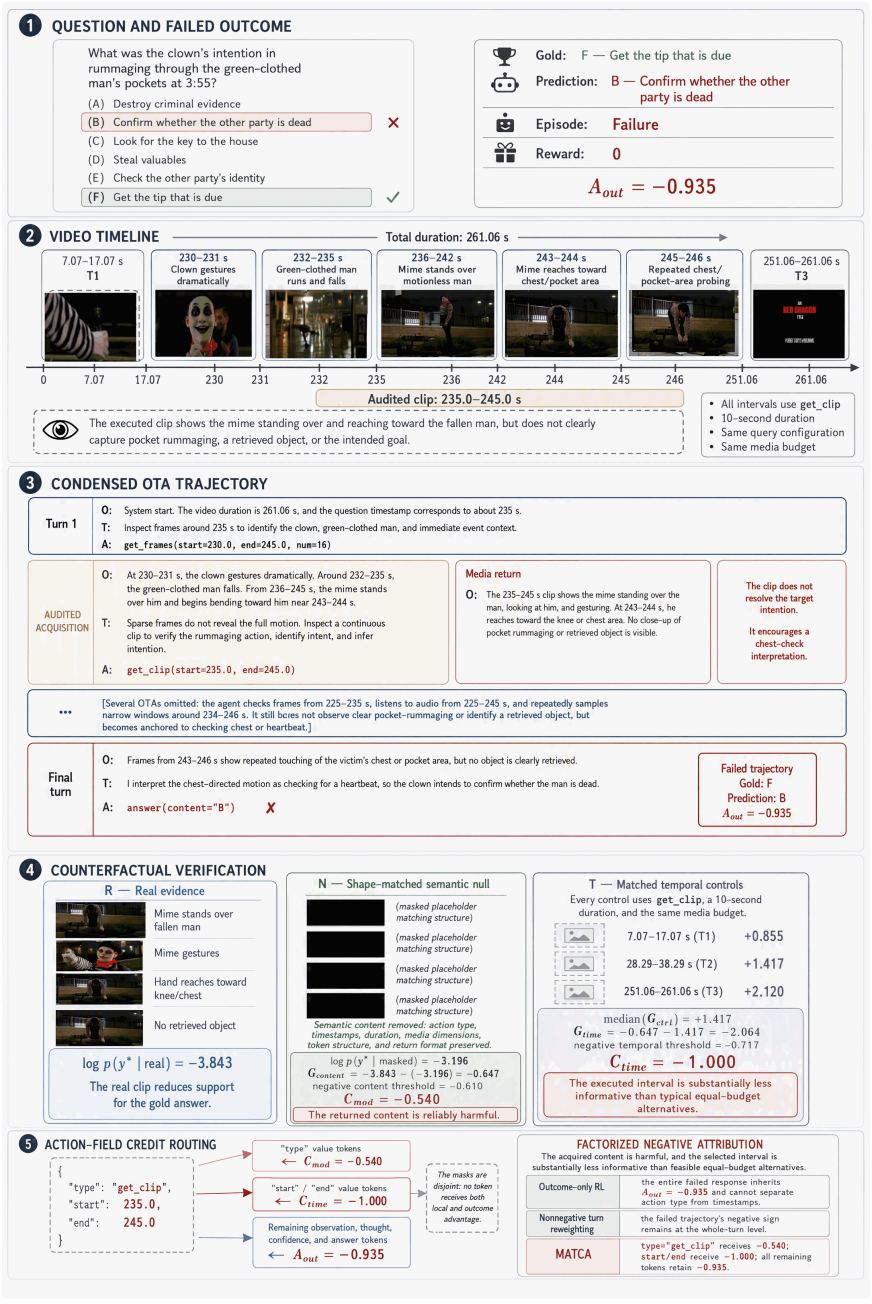

Figure 5: **MATCA separately penalizes harmful acquired content and an uninformative temporal interval.** To determine why the clown searches the green-clothed man's pockets, the agent requests a clip from 235.0–245.0 s. The returned clip shows the clown standing over the motionless man but does not capture the pocket-searching action or reveal its intention. The real media reduces gold-answer support relative to its masked counterpart, producing a content gain of -0.647. Moreover, matched clips of equal duration provide substantially stronger evidence, resulting in a temporal-specificity gain of -2.064. MATCA assigns negative credit to both the acquisition type ($C_{\text{content}} = -0.540$) and its temporal bounds ($C_{\text{time}} = -1.000$), explicitly distinguishing an unhelpful pathway execution from a poorly selected interval.