TEXT-CENTRIC POST-TRAINING FOR OMNI-MODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Improving joint audio-visual reasoning in Omni Large Language Models typically incurs substantial data construction and training costs. Our diagnostics reveal multi-hop reasoning difficulties despite correct answers to all corresponding single-hop questions and suggest partial decoupling in the local optimization of perception and reasoning objectives. This motivates post-training with different emphases on these capabilities. Text-only reasoning training yields gains across data sources, model scales, and families. With the best-performing text-only configuration, supervised fine-tuning followed by reinforcement learning (RL) raises Qwen2.5-Omni-7B's geometric mean of nine reasoning scores by 25.83% over the base model, outperforming the complete native audio-visual route with 56.6% fewer GPU-hours. Training on data synthesized entirely by a text-only LLM raises this geometric mean by 21.01% without audio-visual data in construction or training. However, text-only training degrades perception. We therefore propose a text-centric post-training paradigm: text-only training provides the main reasoning optimization, and reduced-data native audio-visual RL then refines perception. Refinement uses about 90% fewer input tokens than full-data audio-visual RL, restores perception above the base level, and retains 93.5% of the best-performing text-only pipeline's reasoning gain.

1 INTRODUCTION

Omni Large Language Models (Omni-LLMs) have made remarkable progress in unified multimodal perception and understanding, enabling a single model to process text, audio, images, and video (Xu et al., 2025a;b; Team, 2026; Deshmukh et al., 2026; Team et al., 2026). However, strong perceptual capabilities do not necessarily translate into reliable multimodal reasoning (Yao et al., 2026; Wang et al., 2026c). Existing Omni-LLMs still struggle with complex audio-visual tasks that require combining information across modalities and over time (Chen et al., 2025; Dai et al., 2026).

To address this limitation, existing approaches typically use specially constructed multimodal reasoning data for supervised fine-tuning (SFT) or reinforcement learning (RL) (Xing et al., 2025; Yang et al., 2025b; Kulkarni & Fazli, 2026; Araujo et al., 2026). Constructing such data requires annotations consistent with audio-visual content and adequate coverage of infrequent categories in long-tailed distributions (Cai et al., 2026; Chen et al., 2026c). These requirements and multimodal training costs limit scalability (Guan et al., 2026). Meanwhile, text-only reasoning training has improved performance on separate visual and audio tasks (Liu et al., 2025b; Bousselham et al., 2026; Rouditchenko et al., 2025). However, transfer to these tasks does not establish its effectiveness for joint audio-visual reasoning, which requires integrating complementary evidence across modalities and over time. This motivates us to ask: Can Omni-LLM post-training be organized to target perception and reasoning separately while reducing reliance on native audio-visual training?

In Section 3, we investigate this possibility through a diagnostic study. Across nine Omni-LLMs, multi-hop reasoning remains difficult despite correct answers to all corresponding single-hop perception questions. A diagnostic case is shown in Figure 1(a). Beyond this performance gap, we examine whether perception and reasoning admit different optimization directions. Projected gradient updates retain most of the target loss reduction while substantially reducing changes to the other objective. These responses support partial decoupling in the local optimization of the two mean-loss objectives, motivating training strategies with different emphases on perception and reasoning.

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.

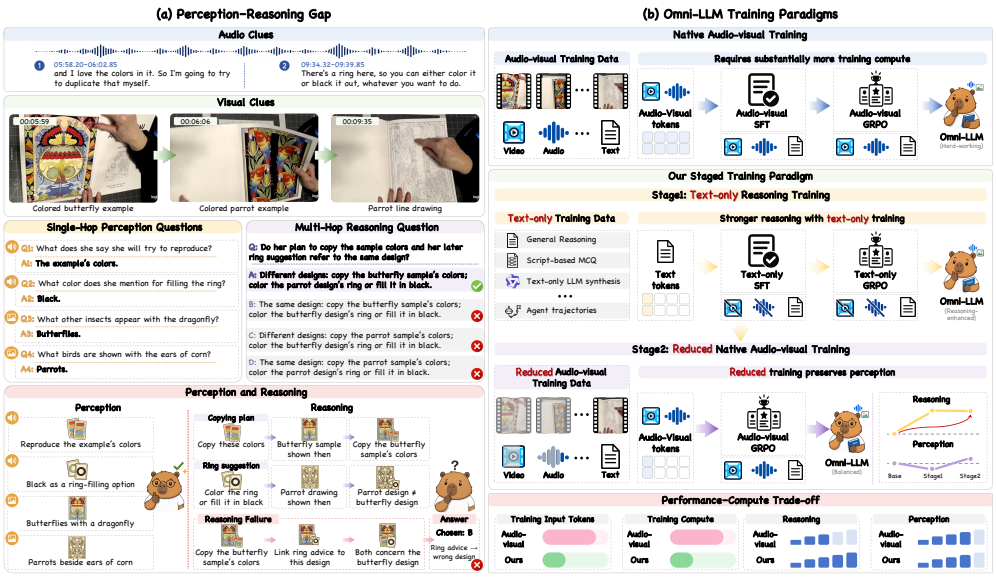


Figure 1: (a) A diagnostic case of multi-hop reasoning failure despite correct answers to corresponding single-hop questions. (b) Text-only training provides the main reasoning optimization, followed by reduced-data native audio-visual RL to restore perception while retaining most reasoning gains.

Building on these findings, we explore organizing Omni-LLM training around the capabilities to be improved, rather than downstream input modalities. We examine text-only reasoning training across data sources. On Qwen2.5-Omni-7B, SFT followed by RL using the best-performing text source improves all nine reasoning scores, raising their average by 9.59 points and geometric mean by 25.83% over the base model. It outperforms the complete native audio-visual route with 56.6% fewer GPU-hours. Gains extend across the evaluated Qwen2.5-Omni, Qwen3-Omni (Xu et al., 2025b), and Gemma 4 (Team et al., 2026) models of different scales. Crucially, a text-only LLM generates fictional scenario descriptions and reasoning questions and answers for SFT and RL, without native audio-visual data in construction or training. Training on these fully synthetic data raises the average reasoning score by 7.70 points and geometric mean by 21.01% over the same base model.

However, text-only reasoning training degrades perception. We therefore propose the text-centric post-training paradigm in Figure 1(b): text-only SFT followed by RL provides the main reasoning optimization, and reduced-data native audio-visual RL then refines perception. On Qwen2.5-Omni-7B, refinement uses about 90% fewer input tokens than full-data audio-visual RL. It restores perception above the base level while retaining 93.5% of the best-performing text-only pipeline's gain in the geometric mean of reasoning scores, addressing the performance–compute trade-off.

Our contributions are as follows:

- We conduct a diagnostic study revealing a gap between single-hop perception and multi-hop reasoning and suggesting partial decoupling in their local optimization.
- We systematically evaluate text-only reasoning training across data sources, model scales, and families, demonstrating omni-modal reasoning gains even with fully synthetic text requiring no audio-visual data in construction or training. Text-only training outperforms the complete native audio-visual route with lower training compute.
- We propose and validate a text-centric post-training paradigm: text-only training provides the main reasoning optimization, and reduced-data native audio-visual RL then restores perception while retaining most of the reasoning gains.

2 Related Work

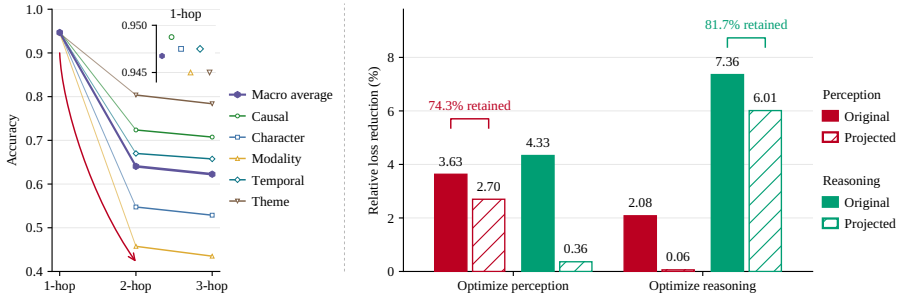
Text-only reasoning transfer. We study joint audio-visual reasoning across modalities and time, whereas text-only transfer studies address visual (Li et al., 2025; Liu et al., 2025b; Bousselham et al.,

2026) or audio tasks (Rouditchenko et al., 2025). Our text-only post-training strengthens reasoning within Omni-LLMs without external inference-time reasoners (Zhang et al., 2025). Unlike text-derived multimodal supervision (Yu et al., 2025; Cheng et al., 2026), our fully synthetic setting uses textual SFT and RL data generated by a text-only LLM without native audio-visual inputs.

Perception and reasoning in multimodal models. Cross-modal evaluations reveal perception and integration limitations (Wang et al., 2026a;b; Mukherjee et al., 2026). Methods use multimodal training (Liu et al., 2026c; Jia et al., 2026; Liu et al., 2026a;b) or active inspection (Xu et al., 2026; Chen et al., 2026b). VLM post-training balances perception and reasoning (Peng et al., 2026; Chen et al., 2026a; Wu et al., 2026). Our diagnostics motivate text-only post-training that improves joint audio-visual reasoning without native audio-visual training data.

3 Preliminary Study

For this diagnostic study, we construct questions from annotations of real audio-visual data and evaluate nine Omni-LLMs on the corresponding inputs. Single-hop perception questions assess the direct extraction of information. From the common subset answered correctly by all nine models, we construct 2-hop and 3-hop reasoning questions that combine the corresponding facts across audio and vision in two and three steps, respectively (Figure 1(a)). Details are provided in Appendix A.1.



(a) Accuracy across hops.

(b) Gradient projection.

Figure 2: Perception–reasoning gap and local update responses.

3.1 The Gap Between Perception and Reasoning

Figure 2a reports macro-averaged accuracy across models, with 94.7% on the original 1-hop set. Despite correct answers to every corresponding single-hop question, accuracy on the resulting multi-hop questions is only 64.1% for 2-hop and 62.3% for 3-hop. Across all five domains, these results reveal a gap between single-hop perception and multi-hop reasoning: recognizing individual facts does not ensure that models can combine them to reach the correct answer.

3.2 Partial Decoupling in Perception and Reasoning Optimization

Beyond the performance gap, we measure improvement on one objective while constraining changes to the other. In Qwen2.5-Omni-3B (Xu et al., 2025a), we define L_E and L_R as the mean cross-entropy losses for predicting the correct answer token on single-hop perception questions and 2-hop and 3-hop reasoning questions, respectively. Under five-fold cross-validation by video, each split defines these objectives on its four training folds. We compute gradients for the Thinker's Transformer decoder layers and final normalization layer, with all other parameters fixed.

Motivated by orthogonal gradient projection (Huang et al., 2025; Qiu et al., 2026), we compare original gradient updates with updates constrained to leave the other training-fold mean loss unchanged to first order. At the initial checkpoint θ_0, let $g_t = \nabla L_t(\theta_0)$ for distinct $t, s \in \{E, R\}$ and define

$$r_{t\perp s} = g_t - \frac{g_s^\top g_t}{\|g_s\|_2^2} g_s, \qquad \theta'_{t\perp s} = \theta_0 - \eta \frac{r_{t\perp s}}{\|r_{t\perp s}\|_2}. \tag{1}$$

The original update is $\theta'_t = \theta_0 - \eta g_t / \|g_t\|_2$. Both updates start from θ_0, with parameter changes of the same Euclidean norm η. We pool held-out loss reductions across folds and express them relative to each objective's pooled initial loss; positive values indicate decreases and negative values indicate increases. Implementation details and supplementary analyses appear in Appendix A.2.

Figure 2b shows that the projected reasoning update reduces reasoning loss by 6.01%, retaining 81.7% of the original target loss reduction. Perception loss decreases by 0.06%, versus 2.08% with the original update. Conversely, the projected perception update retains 74.3% of the original target loss reduction, while reasoning loss decreases by 0.36%, versus 4.33% with the original update. These responses support partial decoupling in the local optimization of the two mean-loss objectives, motivating post-training with different emphases on perception and reasoning.

4 Study Design

Building on the preliminary findings in Section 3, we conduct a systematic study to develop a post-training paradigm that improves omni-modal reasoning with lower training compute. To this end, we organize our study around the following five research questions (RQs):

- **(RQ1) Reasoning Gains and Training Compute.** To what extent does text-only reasoning training improve omni-modal reasoning performance across diverse tasks and models, and how much training compute does it require?
- **(RQ2) Choice of Training Data.** How does the choice of training data affect the reasoning gains from text-only reasoning training on omni-modal tasks?
- **(RQ3) Synthetic Training Data.** Can text-only reasoning training still yield omni-modal reasoning gains when all training data are synthesized by a text-only LLM?
- **(RQ4) Perceptual Effects.** How does text-only reasoning training affect the perceptual capabilities of Omni-LLMs relative to native audio-visual training?
- **(RQ5) Training Paradigm.** How should text-centric post-training be staged to improve omni-modal reasoning while maintaining perceptual capabilities?

4.1 Training Setup

Models. We conduct our main experiments on Qwen2.5-Omni-7B (Xu et al., 2025a). To examine gains across model scales and families, we apply the same text-only training pipeline to Qwen2.5-Omni-3B (Xu et al., 2025a), Gemma-4-E4B-IT, and Gemma-4-E2B-IT (Team et al., 2026). We also evaluate text-only SFT on Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b).

Training details. Text-only and native audio-visual post-training start from the same base model and follow SFT with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). We adopt the best-performing text-only configuration from the RQ2 comparison as the default, using Script-based MCQs in both stages and CoT supervision for SFT. Table 1 lists data types and sources for each RQ. An overview of the training setup is provided in Appendix B.1.

4.2 Evaluation Setup

Benchmarks. For RQ1–RQ3, we evaluate omni-modal reasoning on Daily-Omni (Zhou et al., 2025), Omni-Cloze (Ma et al., 2026), OmniVideoBench (Li et al., 2026a), SocialOmni (Xie et al., 2026), Video-Holmes (Cheng et al., 2025), Video-MME-v2 (Fu et al., 2026), and WorldSense (Hong et al., 2026). For RQ4, we evaluate perceptual capabilities on AVHBench (Sung-Bin et al., 2025), MUSIC-AVQA v1/v2 (Li et al., 2022; Liu et al., 2024), AVQA (Yang et al., 2022), and NExT-QA (Xiao et al., 2021). For RQ5, we jointly evaluate reasoning and perception using these benchmarks.

Baselines. We compare with the original Qwen2.5-Omni-7B (Xu et al., 2025a) and six representative omni-modal reasoning baselines: EchoInk-R1-7B (Xing et al., 2025), AVATAR-7B (Kulkarni & Fazli, 2026), Omni-R1 (Zhong et al., 2025), HumanSense (Qin et al., 2026), HumanOmniV2-7B (Yang et al., 2025b), and OmniVideo-7B (Cai et al., 2026).

Table 1: Data types and sources for text-only and audio-visual training.

Stage	Data type	Data sources	RQ
Text-only training			
SFT	General reasoning	SmallThoughts (Wu et al., 2025); Mixture-of-Thoughts (Face, 2025); Bespoke-Stratos-17k (Labs, 2025); OpenThoughts-114k (Guha et al., 2026); OpenR1-Math-220k (Lozhkov et al., 2025); OpenMathReasoning (Moshkov et al., 2025)	RQ2
	Script-based CoT	OmniVideo-100K (Cai et al., 2026)	RQ1, RQ2 RQ4, RQ5
	Contextual QA	CoT-Drop (Liu et al., 2025a); CoT-Collection (Kim et al., 2023); CoT-Scientific-RAG-Reasoning (Kumar, 2025)	RQ2
GRPO	Script-based MCQ	OmniVideo-100K (Cai et al., 2026)	RQ1, RQ2 RQ4, RQ5
	Temporal QA	Video-T3 QA (Li et al., 2025)	RQ2
	General & math QA	Nemotron-CrossThink (Akter et al., 2026)	RQ2
	Contextual QA	HotpotQA (Yang et al., 2018); MuSiQue (Trivedi et al., 2022); 2WikiMultiHopQA (Ho et al., 2020)	RQ2
SFT+GRPO	Text-only LLM synthesis	Qwen3.7-Max[1]	RQ3
Audio-visual training			
SFT	Audio-visual reasoning	AVRT-20K (Araujo et al., 2026); IntentTrain (Yang et al., 2025b)	RQ1 RQ4, RQ5
GRPO	Audio-visual MCQ	OmniVideo-100K (Cai et al., 2026)	RQ1 RQ4, RQ5

Metrics. We report each benchmark's official metric on a 0–100 scale. **Macro Avg.** is the unweighted mean of scores, computed separately for reasoning and perception. **Geometric Mean Improvement** is the geometric mean of ratios to the corresponding base scores, minus one and expressed as a percentage. For RQ1 and RQ5, we report training input tokens and training compute in GPU-hours (number of GPUs $\times$ training time in hours) separately for each training stage.

5 Findings

We summarize the findings from our five RQs and their implications for Omni-LLM training:

- **(RQ1)** Text-only reasoning training yields broad omni-modal reasoning gains with less training compute than the complete native audio-visual route.
- **(RQ2)** Script-based MCQs yield the strongest and broadest reasoning transfer.
- **(RQ3)** Training on data synthesized entirely by a text-only LLM also improves omni-modal reasoning, without native audio-visual data in construction or training.
- **(RQ4)** Text-only reasoning training incurs perceptual degradation.
- **(RQ5)** A text-centric training paradigm addresses the performance–compute trade-off by using text-only training for reasoning optimization and reduced-data native audio-visual refinement to maintain perception while retaining the main reasoning gains.

5.1 RQ1: Reasoning Gains and Training Compute

Text-only reasoning training yields substantial and broad omni-modal reasoning gains. In Table 2, text-only SFT followed by GRPO raises Qwen2.5-Omni-7B's Macro Avg. from 39.80 to 49.39, a 9.59-point gain. Its Macro Avg. and geometric mean improvement of $+25.83\%$ both rank first. Trained only on text, it exceeds EchoInk-R1-7B and the complete native audio-visual route by 4.05 and 2.02 Macro Avg. points, respectively. It ranks first on all three SocialOmni subtasks, Video-MME-v2, and WorldSense, and second on Daily-Omni and OmniVideoBench. All nine scores improve over the base model, whereas every external baseline declines on at least one.

[1] https://www.alibabacloud.com/help/en/model-studio/qwen3-7-max

Table 2: Omni-modal reasoning performance and training compute. Parenthesized values show absolute changes from the base model. **Bold** and underlined scores denote the best and second-best results, respectively.

Model	Omni-modal Reasoning									Aggregate		Training Budget	
	Daily-Omni	Omni-Cloze	Omni Video	SocialOmni			Video-Holmes	Video-MME-v2	World Sense	Macro Avg.	Geo. Mean Improv. (%)	Input Tok. (B) ↓	GPU-h ↓
				Who	When	How							
Qwen2.5-Omni-7B	62.07	21.90	29.30	36.75	57.42	56.67	24.40	24.31	45.40	39.80	0.00	*N/A*	
External Baselines													
EchoInk-R1-7B (Xing et al., 2025)	61.07	27.30	33.50	41.85	59.81	70.45	40.77	27.72	45.60	45.34	+16.51	0.072	191.2
AVATAR-7B (Kulkarni & Fazli, 2026)	56.14	25.94	31.60	36.25	60.29	67.26	45.10	20.47	46.00	43.23	+9.37	0.419	385.9
Omni-R1 (Zhong et al., 2025)	55.56	26.28	32.10	33.30	62.68	72.55	32.12	25.04	41.40	42.34	+7.03	0.055	576.0
HumanSense (Qin et al., 2026)	56.31	26.98	31.70	38.70	66.51	75.86	33.15	24.33	42.20	43.97	+10.69	0.208	190.3
HumanOmniV2-7B (Yang et al., 2025b)	58.47	**29.38**	30.50	40.20	57.89	66.03	42.90	24.93	47.10	44.16	+13.65	4.506	2164.1
OmniVideo-7B (Cai et al., 2026)	**68.34**	25.40	**38.30**	35.45	41.15	50.40	38.00	24.54	45.10	40.74	+5.38	1.511	201.0
Training Variants													
Audio-visual SFT	54.64	28.32	32.90	37.25	63.64	56.76	36.74	26.96	43.40	42.29	+9.67	0.092	12.2
+ GRPO	62.99	26.48	36.70	41.40	64.59	69.76	**46.65**	30.20	47.60	47.37	+21.96	1.214	642.3
Text-only SFT	52.72	25.25	31.40	39.30	53.59	66.07	27.71	22.34	40.90	39.92	+1.55	0.043	10.2
+ **GRPO (Ours)**	65.75	27.17	37.00	**41.95**	**68.90**	**79.93**	44.58	**30.66**	**48.60**	**49.39**	**+25.83**	0.463	274.0
	(+3.68)	(+5.27)	(+7.70)	(+5.20)	(+11.48)	(+23.26)	(+20.18)	(+6.35)	(+3.20)	(+9.59)			

Table 3: Generalization of text-only reasoning training across model scales and families. Parenthesized values show absolute changes from the corresponding base model.

Model	Omni-modal Reasoning									Aggregate		Training Budget	
	Daily-Omni	Omni-Cloze	Omni Video	SocialOmni			Video-Holmes	Video-MME-v2	World Sense	Macro Avg.	Geo. Mean Improv. (%)	Input Tok. (B) ↓	GPU-h ↓
				Who	When	How							
Qwen2.5-Omni-3B	56.22	23.51	27.30	36.50	50.24	51.23	25.26	21.35	43.50	37.23	0.00	*N/A*	
Text-only SFT	51.04	23.52	32.20	32.55	55.50	58.93	32.01	20.46	40.00	38.47	+3.44	0.128	10.1
+ **GRPO**	64.16	24.40	36.00	37.10	57.42	71.88	42.95	25.17	45.30	44.93	+20.39	0.463	259.3
	(+7.94)	(+0.89)	(+8.70)	(+0.60)	(+7.18)	(+20.65)	(+17.69)	(+3.82)	(+1.80)	(+7.70)			
Gemma-4-E4B-IT	50.54	25.16	28.10	35.30	52.63	69.17	32.55	21.81	35.80	39.01	0.00	*N/A*	
Text-only SFT	48.37	23.86	30.60	33.05	59.33	76.47	31.95	20.37	35.20	39.91	+0.43	0.042	12.9
+ **GRPO**	57.31	25.77	33.10	36.40	61.24	79.73	42.84	21.30	39.20	44.10	+11.50	0.572	289.0
	(+6.77)	(+0.61)	(+5.00)	(+1.10)	(+8.61)	(+10.56)	(+10.29)	(−0.51)	(+3.40)	(+5.09)			
Gemma-4-E2B-IT	48.20	26.45	28.00	34.35	51.67	66.18	30.00	19.56	33.50	37.55	0.00	*N/A*	
Text-only SFT	47.45	22.76	29.00	33.05	57.89	72.75	28.14	18.69	35.80	38.39	-0.05	0.042	9.8
+ **GRPO**	51.96	24.03	30.50	34.15	62.20	73.28	33.70	20.09	36.90	40.76	+6.73	0.565	215.2
	(+3.76)	(−2.42)	(+2.50)	(−0.20)	(+10.53)	(+7.10)	(+3.70)	(+0.53)	(+3.40)	(+3.21)			
Qwen3-Omni-30B	71.85	33.50	38.40	69.25	63.64	46.35	50.40	27.00	54.00	50.49	0.00	*N/A*	
Text-only SFT	64.66	36.08	41.30	66.15	59.81	73.28	54.00	29.62	52.74	53.07	+6.06	0.135	31.1
	(−7.19)	(+2.58)	(+2.90)	(−3.10)	(−3.83)	(+26.93)	(+3.60)	(+2.62)	(−1.26)	(+2.58)			

The reasoning gains extend across model scales and families. In Table 3, text-only SFT followed by GRPO raises Qwen2.5-Omni-3B's Macro Avg. from 37.23 to 44.93, with a geometric mean improvement of $+20.39\%$. It improves all nine scores over the base model, including gains of 20.65 and 17.69 points on SocialOmni How and Video-Holmes, respectively. On Gemma-4-E4B-IT and Gemma-4-E2B-IT, Macro Avg. rises by 5.09 and 3.21 points, and eight and seven scores improve, respectively. For the latter, GRPO improves all nine scores over SFT despite uneven SFT gains. Even on the stronger Qwen3-Omni-30B-A3B-Instruct, text-only SFT alone raises the Macro Avg. from 50.49 to 53.07. These results support text-only training across scales and families.

Text-only reasoning training achieves a superior performance–compute trade-off. The complete text-only route in Table 2 improves Macro Avg. (49.39 versus 47.37) with fewer GPU-hours (284.17 versus 654.52) than the native audio-visual route. Across SFT and GRPO, it processes 0.505B versus 1.306B training input tokens. This reduces input tokens by 61.3% and training compute by 56.6%. It achieves the highest Macro Avg. among compared methods with fewer GPU-hours than measured for AVATAR-7B, Omni-R1, and HumanOmniV2-7B. Thus, text-only training outperforms the complete native audio-visual route with less training compute.

5.2 RQ2: Choice of Training Data

Script-based MCQ yields stronger omni-modal reasoning transfer. In Table 4, GRPO on Script-based MCQs achieves the highest Macro Avg. after SFT on either General reasoning data or Script-based MCQs with CoT supervision, reaching 48.91 and 49.39, respectively. Contextual QA in both phases yields 42.97. Both MCQ configurations improve all nine scores over the base

Table 4: Effect of training data choice on omni-modal reasoning. **Bold** and underlined scores denote the best and second-best results, respectively.

Training Configuration	Omni-modal Reasoning									Aggregate	
	Daily-Omni	Omni-Cloze	Omni Video	SocialOmni			Video-Holmes	Video-MME-v2	World Sense	Macro Avg.	Geo. Mean Improv. (%)
				Who	When	How					
Qwen2.5-Omni-7B	62.07	21.90	29.30	36.75	57.42	56.67	24.40	24.31	45.40	39.80	0.00
SFT: General reasoning	55.72	26.58	31.40	37.35	56.94	69.02	37.51	25.49	45.90	42.88	+9.88
+ **GRPO:** Script-based MCQ	**66.58**	**28.56**	36.90	41.45	63.64	75.42	**47.25**	30.41	**50.00**	48.91	+25.67
+ **GRPO:** Temporal QA	55.22	27.01	34.00	38.60	57.89	76.09	38.54	25.75	44.90	44.22	+12.96
+ **GRPO:** General & math QA	55.56	26.88	33.70	38.65	63.64	74.59	38.16	25.92	44.70	44.64	+13.73
SFT: Script-based CoT (MCQ)	52.72	25.25	31.40	39.30	53.59	66.07	27.71	22.34	40.90	39.92	+1.55
+ **GRPO:** Script-based MCQ	65.75	27.17	**37.00**	**41.95**	**68.90**	**79.93**	44.58	**30.66**	48.60	**49.39**	**+25.83**
+ **GRPO:** Temporal QA	57.06	25.88	33.57	37.90	57.42	71.08	36.25	22.81	42.09	42.67	+8.48
+ **GRPO:** General & math QA	55.81	25.97	31.56	37.55	55.50	67.93	35.49	25.22	42.72	41.97	+7.58
SFT: Contextual QA	52.88	26.46	32.80	38.25	54.07	70.92	27.82	21.44	39.90	40.50	+2.50
+ **GRPO:** Contextual QA	55.14	26.47	31.20	37.15	59.81	73.98	35.44	25.24	42.30	42.97	+9.20

model, whereas other complete configurations improve the Macro Avg. but not all scores. Tracking entities and event states and integrating evidence across segments may explain these broader gains in omni-modal reasoning. Further results on data mixtures appear in Appendix C.1.

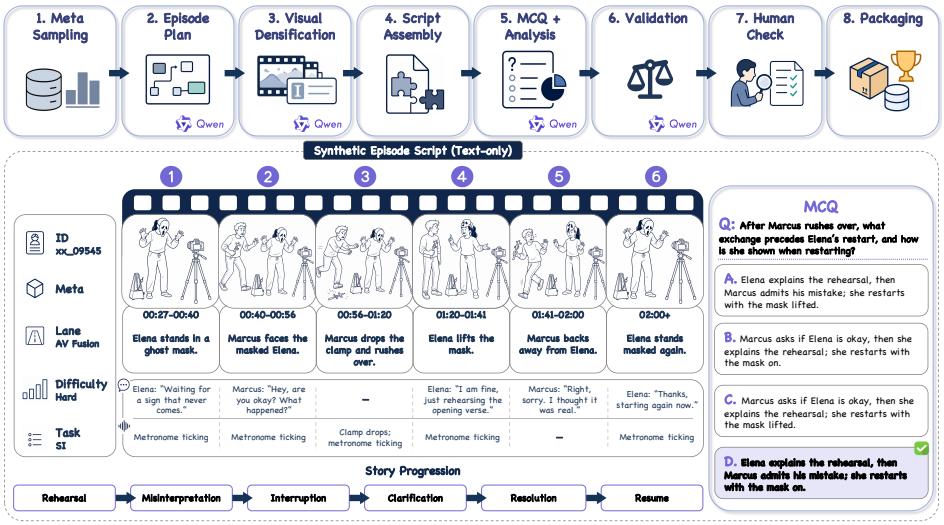


Figure 3: Text-only LLM synthesis of scripts, MCQs, and reasoning explanations, without audio-visual data in construction or training. The illustration shows a fictional script, not generated video.

5.3 RQ3: Synthetic Training Data

Text-only LLM synthesis extends text-only training to data construction. Building on RQ1, we test text-only reasoning training on SFT and GRPO data synthesized entirely by a text-only LLM. Neither data construction nor training uses any native audio-visual data. Figure 3 shows generation from specified scenarios, task types, and modality requirements. The LLM creates fictional episode scripts with visual descriptions, speech, and non-speech sounds, maintaining entity and temporal consistency across segments. MCQs require entity tracking, event-state updates, and cross-segment evidence integration; joint audio-visual questions require complementary audio and visual evidence. Scripts, MCQs, and reasoning explanations undergo validation and human review.

Text-only LLM synthesis yields omni-modal reasoning gains. In Figure 4, training on synthesized text yields a 7.70-point Macro Avg. gain for Qwen2.5-Omni-7B, from 39.80 to 47.50. Gains on Qwen2.5-Omni-3B, Gemma-4-E4B-IT, and Gemma-4-E2B-IT are 6.59, 3.93, and 2.79 points, respectively. The Qwen models improve nine and eight scores over their base models and remain within 1.89 and 1.11 Macro Avg. points of Script-based MCQ. Crucially, these gains require no na-

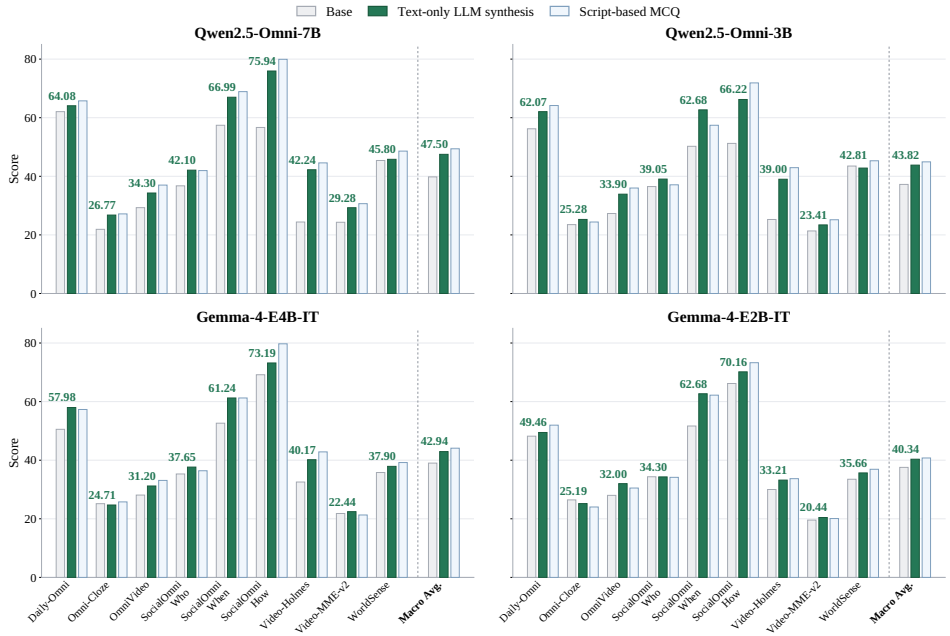


Figure 4: Omni-modal reasoning with training data synthesized entirely by a text-only LLM for both SFT and GRPO. Results cover four models across different scales and families.

Table 5: Perceptual effects of text-only and audio-visual training. Parenthesized values show absolute changes from the base model.

Model	Perceptual Capabilities							Aggregate	
	AVHBench			MUSIC-AVQA		AVQA	NExT-QA	Macro Avg.	Geo. Mean Improv. (%)
	Overall	Audio-driven Video Hallucination	Video-driven Audio Hallucination	v1	v2	Which / Come From / Happening / Where	Descriptive		
Qwen2.5-Omni-7B	75.55	80.70	73.00	71.22	69.10	88.79	82.21	77.37	0.00
Audio-visual SFT	78.12	82.60	75.91	75.91	74.62	90.70	83.13	80.50	+4.22
+ GRPO	78.31	81.87	76.55	73.15	70.32	90.58	82.99	79.07	+2.21
	(+2.76)	(+1.17)	(+3.55)	(+1.93)	(+1.22)	(+1.79)	(+0.78)	(+1.70)	
Text-only SFT	75.78	80.30	73.55	65.09	62.93	83.75	79.22	73.35	-5.37
+ GRPO (Ours)	73.94	79.42	71.23	69.60	68.99	87.68	81.85	76.41	-1.25
	(−1.61)	(−1.28)	(−1.77)	(−1.62)	(−0.11)	(−1.11)	(−0.36)	(−0.96)	

tive audio-visual data in construction or training. This opens a path to scaling up reasoning training data without collecting or annotating corresponding audio-visual samples.

5.4 RQ4: Perceptual Effects

Text-only reasoning training incurs perceptual degradation. In Table 5, text-only SFT followed by GRPO reduces Qwen2.5-Omni-7B's perceptual Macro Avg. from 77.37 to 76.41, with a 1.25% decrease in the geometric mean of scores. All seven perceptual scores decline by 0.11–1.77 points. In contrast, native audio-visual SFT followed by GRPO improves all seven scores and raises the Macro Avg. by 1.70 points to 79.07. The text-only route's 9.59-point reasoning gain in RQ1 accompanies a 0.96-point perceptual Macro Avg. decline. This pattern is consistent with the partial decoupling of perception and reasoning suggested by our preliminary study. Nevertheless, text-only reasoning training does not fully replace native audio-visual training for preserving perception.

5.5 RQ5: Training Paradigm

RQ1–RQ4 motivate a text-centric training paradigm. Building on the reasoning gains in RQ1–RQ3 and perceptual degradation in RQ4, we investigate text-centric post-training. Once reasoning has been strengthened through text-only training, we test whether reduced-data native audio-visual

Table 6: Reasoning, perception, and training budgets under staged training. A→V and V→A denote audio-driven video and video-driven audio hallucination. Parentheses show absolute changes from the base model. Budgets are per audio-visual stage; reduced-data budgets are shared across initializations.

(a) Omni-modal Reasoning

Training Configuration		Daily-Omni	Omni-Cloze	Omni Video	SocialOmni			Video-Holmes	Video-MME-v2	World Sense	Macro Avg.	Geo. Mean Improv. (%)
Text-only	Audio-visual				Who	When	How					
Qwen2.5-Omni-7B		62.07	21.90	29.30	36.75	57.42	56.67	24.40	24.31	45.40	39.80	0.00
Reduced-data Audio-visual SFT												
–	SFT	51.38	25.86	31.90	37.90	56.46	68.52	34.08	20.69	40.10	40.77	+3.56
SFT	SFT	55.14	25.74	37.40	37.30	59.81	71.59	35.49	23.38	41.60	43.05	+9.60
SFT + GRPO	SFT	65.33	27.06	37.40	42.10	63.64	70.56	45.18	28.09	47.40	47.42	+21.71
Reduced-data Audio-visual GRPO												
SFT	GRPO	62.74	26.75	37.47	40.80	66.51	72.35	43.22	27.98	47.10	47.21	+20.81
SFT + GRPO	GRPO	64.75	27.08	37.17	41.15	66.51	75.00	45.45	30.16	49.02	48.48	+24.15
		(+2.68)	(+5.18)	(+7.87)	(+4.40)	(+9.09)	(+18.33)	(+21.05)	(+5.85)	(+3.62)	(+8.67)	

(b) Perceptual Capabilities

Training Configuration		AVHBench			MUSIC-AVQA		AVQA	NExT-QA	Macro Avg.	Geo. Mean Improv. (%)
Text-only	Audio-visual	Overall	A→V	V→A	v1	v2	Which / Come From / Happening / Where	Descriptive		
Qwen2.5-Omni-7B		75.55	80.70	73.00	71.22	69.10	88.79	82.21	77.37	0.00
Reduced-data Audio-visual SFT										
–	SFT	76.25	80.83	73.98	74.50	71.64	90.67	83.35	79.28	+2.53
SFT	SFT	76.31	81.10	73.94	74.41	71.67	90.64	83.20	79.25	+2.49
SFT + GRPO	SFT	74.87	80.02	72.32	73.14	70.42	90.39	83.84	78.53	+1.49
Reduced-data Audio-visual GRPO										
SFT	GRPO	75.84	80.65	73.46	74.19	71.30	89.14	82.70	78.63	+1.73
SFT + GRPO	GRPO	76.06	81.60	73.32	73.22	70.49	89.69	83.06	78.50	+1.51
		(+0.51)	(+0.90)	(+0.32)	(+2.00)	(+1.39)	(+0.90)	(+0.85)	(+1.13)	

(c) Training Budget

Metric	Full-data Audio-visual		Reduced-data Audio-visual	
	SFT	GRPO	SFT	GRPO
Input Tok. (B) ↓	0.092	1.214	0.002	0.122
GPU-h ↓	12.2	642.3	0.4	82.7

training is sufficient to preserve perception. Specifically, we apply audio-visual SFT or GRPO after either text-only SFT or the complete text-only SFT–GRPO pipeline, with reduced-data audio-visual SFT alone as a control. Table 6 reports reasoning, perception, and training budgets.

Staged training improves reasoning while maintaining perception. Table 6 shows higher reasoning Macro Avg. scores from the complete text-only SFT–GRPO pipeline than from text-only SFT under either audio-visual refinement method. With this initialization, audio-visual GRPO achieves the highest reasoning Macro Avg. among the staged configurations at 48.48 versus 47.42 for audio-visual SFT, with nearly identical perception averages of 78.50 and 78.53. This refinement uses about 90% fewer training input tokens than full-data native audio-visual GRPO.

Compared with the complete text-only route in Tables 2 and 5, this configuration improves the perception Macro Avg. by 2.09 points while reducing reasoning by 0.91 points. It retains 93.5% of the text-only reasoning gain measured by geometric mean improvement (24.15% versus 25.83% before refinement). It exceeds the full-data native audio-visual SFT–GRPO route by 1.11 reasoning Macro Avg. points, with a 0.57-point lower perception average. Together, these results establish our recommended text-centric post-training paradigm for Omni-LLMs: text-only SFT followed by GRPO for the main reasoning optimization, then reduced-data native audio-visual GRPO for perceptual refinement. This paradigm retains the main reasoning gains while restoring perception above the base level, addressing the performance–compute trade-off.

6 Conclusion

Building on our diagnostics, we studied text-centric post-training for Omni-LLMs. Text-only training improves reasoning across data sources, model scales, and families, even with fully synthetic text requiring no audio-visual data in construction or training. On Qwen2.5-Omni-7B, our best-performing text-only pipeline outperforms the complete native audio-visual route with less compute. Reduced-data native audio-visual RL then restores perception while retaining most reasoning gains. This text-centric paradigm organizes post-training around the capabilities to be improved: text for reasoning optimization and native audio-visual supervision for perceptual refinement.

AI USE STATEMENT

We used large language models for diagnostic question construction, synthetic data generation and validation, and benchmark evaluation, as detailed in Appendices A.1 and B. AI tools were additionally used for language editing and limited code assistance. All research decisions, analyses, and conclusions were made and verified by the authors. The authors take responsibility for the content and conclusions of this work.

ETHICS STATEMENT

This study uses existing datasets and benchmarks alongside fictional text-only training scenarios generated without native audio-visual inputs. Synthetic data undergo programmatic and model-based validation with manual spot checks. Generated data and model outputs may reflect biases and inaccuracies in the underlying models and datasets.

REPRODUCIBILITY STATEMENT

Training and evaluation settings are described in Sections 4.1 and 4.2. Appendices A, B, and D provide diagnostic procedures, implementation details, and prompts, respectively. To facilitate direct replication of our results, we will release our source code, the generated dataset, and trained model checkpoints upon publication.

REFERENCES

Syeda Nahida Akter, Shrimai Prabhumoye, Matvei Novikov, Seungju Han, Ying Lin, Evelina Bakhturina, Eric Nyberg, Yejin Choi, Mostofa Patwary, Mohammad Shoeybi, et al. Nemotron-crossthink: Scaling self-learning beyond math reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 984–1002, 2026.

Edson Araujo, Saurabhchand Bhati, M Jehanzeb Mirza, Brian Kingsbury, Samuel Thomas, Rogerio Feris, James R Glass, and Hilde Kuehne. Avrt: Audio-visual reasoning transfer through single-modality teachers. *arXiv preprint arXiv:2604.16617*, 2026.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. Vold: Reasoning transfer from llms to vision-language models via on-policy distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26209–26218, 2026.

Xinyue Cai, Chaoyou Fu, Yi-Fan Zhang, Ran He, and Caifeng Shan. Omnivideo-100k: A dataset for audio-visual reasoning through structured scripts and evidence chains. *arXiv preprint arXiv:2606.14702*, 2026.

Joao V Cavalcanti, Laurent Lessard, and Ashia C Wilson. Adaptive backtracking line search. *arXiv preprint arXiv:2408.13150*, 2024.

Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, et al. Omnixr: Evaluating omni-modality language models on reasoning across modalities. In *International Conference on Learning Representations*, volume 2025, pp. 68100–68118, 2025.

Shuang Chen, Hangyu Guo, Zhaochen Su, Yafu Li, Jiacheng Chen, Yulun Wu, Weijie Wang, ZhiYuan Feng, Xiaoye Qu, and Yu Cheng. Revisual-r1: Advancing multimodal reasoning from optimized cold start to staged reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 32869–32888, 2026a.

Yu Chen, Caorui Li, Ziyu Xiong, Yidong Wang, Mingqi Gao, Shuman Liu, Biao Liu, Chunfeng Yang, Anxiang Zeng, Haibo Zhang, et al. Omnireasoner: Thinking with long audio-video via native tool use. *arXiv preprint arXiv:2607.19339*, 2026b.

Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, et al. Omnivideo-r1: Reinforcing audio-visual reasoning with query intention and modality attention. *arXiv preprint arXiv:2602.05847*, 2026c.

Dongjie Cheng, Yongqi Li, Zhixin Ma, Hongru Cai, Yupeng Hu, Wenjie Wang, Liqiang Nie, and Wenjie Li. Omni-r1: Towards the unified generative paradigm for multimodal reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38907–38924, 2026.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Yifan Dai, Zhenhua Wu, Bohan Zeng, Daili Hua, Jialing Liu, Bozhou Li, Yuran Wang, Chengzhuo Tong, Hao Liang, Xiaochen Ma, et al. Latentomni: Rethinking omni-modal understanding via unified audio-visual latent reasoning. *arXiv preprint arXiv:2605.22012*, 2026.

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, et al. Nemotron 3 nano omni: efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

Yuwen Du, Rui Ye, Shuo Tang, Xinyu Zhu, Yijun Lu, Yuzhu Cai, and Siheng Chen. Openseeker: Democratizing frontier search agents by fully open-sourcing training data. *arXiv preprint arXiv:2603.15594*, 2026.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL `https://github.com/huggingface/open-r1`.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pp. 3762–3773. PMLR, 2020.

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

John Fox and Georges Monette. cv: An r package for cross-validating regression models. *Journal of Statistical Software*, 116:1–41, 2026.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

Yiran Guan, Sifan Tu, Dingkang Liang, Linghao Zhu, Jianzhong Ju, Zhenbo Luo, Jian Luan, Yuliang Liu, and Xiang Bai. Thinkomni: Lifting textual reasoning to omni-modal scenarios via guidance decoding. *arXiv preprint arXiv:2602.23306*, 2026.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. In *International Conference on Learning Representations*, volume 2026, pp. 108059–108130, 2026.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 2026.

Chengyue Huang, Junjiao Tian, Brisa Maneechotesuwan, Shivang Chopra, and Zsolt Kira. Directional gradient projection for robust fine-tuning of foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 60524–60544, 2025.

Yiduo Jia, Muzhi Zhu, Hao Zhong, Mingyu Liu, Yuling Xi, Hao Chen, Bin Qin, Yongjie Yang, Zhenbo Luo, and Chunhua Shen. Omnijigsaw: Enhancing omni-modal reasoning via modality-orchestrated reordering. *arXiv preprint arXiv:2604.08209*, 2026.

Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12685–12708, 2023.

Yogesh Kulkarni and Pooyan Fazli. Avatar: Reinforcement learning to see, hear, and reason over video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7912–7922, 2026.

Abhinav Kumar. CoT-Scientific-RAG-Reasoning. Hugging Face dataset, December 2025. URL https://huggingface.co/datasets/abhinavdread/CoT-Scientific-RAG-Reasoning/tree/9a883557fd4b5b958c1c8b8b431ba08f6ae39c42. Revision 9a88355; accessed 2026-08-25.

Bespoke Labs. Bespoke-stratos: The unreasonable effectiveness of reasoning distillation. https://www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoning-distillation, 2025. Accessed: 2025-01-22.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *International Conference on Learning Representations*, volume 2026, pp. 138214–138236, 2026a.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19086–19096. IEEE, 2022.

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. Temporal reasoning transfer from text to video. In *International Conference on Learning Representations*, volume 2025, pp. 74070–74102, 2025.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Haoxuan Li, Hao Wang, Shijian Wang, Guanting Dong, Jiajie Jin, Yinuo Wang, Yuan Lu, et al. Omnigaia: Towards native omni-modal ai agents. *arXiv preprint arXiv:2602.22897*, 2026b.

Che Liu, Lichao Ma, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Xuerui Yang, and Fei Tian. Boosting omni-modal language models: Staged post-training with visually debiased evaluation. *arXiv preprint arXiv:2605.12034*, 2026a.

Hongcheng Liu, Pingjie Wang, Heyang Liu, Zhiyuan Zhu, Yusheng Liao, Yanfeng Wang, and Yu Wang. Anchornet: Adaptive anchor token enhancement in video-grounded dialogue generation. *IEEE Journal of Selected Topics in Signal Processing*, 2026b.

Hongcheng Liu, Yuhao Wang, Zhe Chen, Pingjie Wang, Zhiyuan Zhu, Yixuan Hou, Yanfeng Wang, and Yu Wang. Cross-modal coreference alignment: Enabling reliable information transfer in omni-LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26430–26453, San Diego, California, United States, July 2026c. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1217. URL https://aclanthology.org/2026.acl-long.1217/.

Jiang Liu, Jialian Wu, Xiaodong Yu, Yusheng Su, Prakamya Mishra, Gowtham Ramesh, Sudhanshu Ranjan, Chaitanya Manem, Ximeng Sun, Ze Wang, et al. Instella: Fully open language models with stellar performance. *arXiv preprint arXiv:2511.10628*, 2025a.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, et al. X-reasoner: Towards generalizable reasoning across modalities and domains. *arXiv preprint arXiv:2505.03981*, 2025b.

Xiulong Liu, Zhikang Dong, and Peng Zhang. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4478–4487, 2024.

Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Guilherme Penedo, Edward Beeching, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. Openr1-math-220k. `https://huggingface.co/datasets/open-r1/OpenR1-Math-220k`, 2025.

Fanbin Lu, Zhisheng Zhong, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Arpo: End-to-end policy optimization for gui agents with experience replay. *arXiv preprint arXiv:2505.16282*, 2025.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *International Conference on Learning Representations*, volume 2026, pp. 132277–132304, 2026.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025.

Swapnanil Mukherjee, Agyeya Negi, Tanuja Ganu, and Ponnurangam Kumaraguru. C^3po: Evaluating cross-modal composition and counterfactual performance in omnimodal models. *arXiv preprint arXiv:2608.05381*, 2026.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities with two-stage rule-based rl. *Frontiers of Computer Science*, 2026.

Zheng Qin, Ruobing Zheng, Yabing Wang, Tianqi Li, Yi Yuan, Jingdong Chen, and Le Wang. Humansense: From multimodal perception to empathetic context-aware responses through reasoning mllms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24973–24981, 2026.

Haomiao Qiu, Miao Zhang, Ziyue Qiao, Weili Guan, Min Zhang, and Liqiang Nie. Splitlora: Balancing stability and plasticity in continual learning through gradient space splitting. In *International Conference on Learning Representations*, volume 2026, pp. 48458–48480, 2026.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-r1: Do you really need audio to fine-tune your audio llm? In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Kim Sung-Bin, Oh Hyun-Bin, Lee Jung-Mok, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations*, volume 2025, pp. 24244–24271, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, et al. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *arXiv preprint arXiv:2511.11793*, 2025.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Xingrui Wang, Jiang Liu, Chao Huang, Xiaodong Yu, Ze Wang, Ximeng Sun, Jialian Wu, Alan Yuille, Emad Barsoum, and Zicheng Liu. Xmodbench: Benchmarking cross-modal capabilities and consistency in omni-language models. In *International Conference on Learning Representations*, volume 2026, pp. 19508–19521, 2026a.

Yaoting Wang, Ziyi Zhang, Wenming Tu, Shaoxuan Xu, Wenjie Du, Cheng Liang, Weijun Wang, Yuanchao Li, Guangyao Li, Hao Fei, et al. Avi-bench: Toward human-like audio-visual intelligence of omni-mllms. *arXiv preprint arXiv:2606.07643*, 2026b.

Yucheng Wang, Yifan Hou, Aydin Javadov, Mubashara Akhtar, and Mrinmaya Sachan. Compose and fuse: Revisiting the foundational bottlenecks in multimodal reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 11272–11301, 2026c.

Juncheng Wu, Hardy Chen, Haoqin Tu, Xianfeng Tang, Freda Shi, Hui Liu, Hanqing Lu, Cihang Xie, and Yuyin Zhou. From seeing to thinking: Decoupling perception and reasoning improves post-training of vision-language models. *arXiv preprint arXiv:2605.20177*, 2026.

Yifan Wu, Jingze Shi, Bingheng Wu, Jiayi Zhang, Xiaotian Lin, Nan Tang, and Yuyu Luo. Concise reasoning, big gains: Pruning long reasoning trace with difficulty-aware prompting. *arXiv preprint arXiv:2505.19716*, 2025.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Tianyu Xie, Jinfa Huang, Yuexiao Ma, Rongfang Luo, Yan Yang, Qingchuan Ma, Wang Chen, Yuhui Zeng, Yixuan Zou, Zhiqiang Lu, et al. Socialomni: Benchmarking audio-visual social interactivity in omni models. *arXiv preprint arXiv:2603.16859*, 2026.

Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 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. *arXiv preprint arXiv:2605.28192*, 2026.

Chen Yang, Ran Le, Yun Xing, Zhenwei An, Zongchao Chen, Wayne Xin Zhao, Yang Song, and Tao Zhang. Toolmind technical report: A large-scale, reasoning-enhanced tool-use dataset. *arXiv preprint arXiv:2511.15718*, 2025a.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3480–3491, 2022.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025b.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Ruilin Yao, Bo Zhang, Jirui Huang, Xinwei Long, Yifang Zhang, Tianyu Zou, Shili Xiong, Yi Rong, Yufei Wu, Shichao Su, et al. Lens: Multi-level evaluation of multimodal reasoning with large language models. In *International Conference on Learning Representations*, volume 2026, pp. 2627–2654, 2026.

Xiaomin Yu, Wenjie Zhang, Ziyue Qiao, Chengwei Qin, and Hui Xiong. Text-only data synthesis for vision language model training. *arXiv preprint arXiv:2503.22655*, 2025.

Weijia Zhang, Zijia Liu, Haoru Li, Haoqi Chen, and Jiaxuan You. Seeingeye: Agentic information flow unlocks multimodal reasoning in text-only llms. *arXiv preprint arXiv:2510.25092*, 2025.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. 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, Main Conference, pp. 45266–45297. Curran Associates, Inc., 2025. doi: 10.52202/085713-1508. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4086fe59dc3584708468fba0e459f6a7-Paper-Conference.pdf.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

A Diagnostic Analysis

A.1 Diagnostic Question Construction and Evaluation

This section details the construction and evaluation of the diagnostic questions used in Section 3 to examine perception and reasoning in Omni-LLMs.

Question taxonomy. We categorize questions by the number of reasoning steps specified during construction. Single-hop perception questions assess the direct extraction of information from audio-visual inputs. Multi-hop questions combine facts from corresponding single-hop QA pairs across audio and vision in two or three steps. They must connect entities, events, attributes, or relations rather than merely group independent questions from the same video or topic. Both question types cover five domains: Causal, Character, Modality, Temporal, and Theme.

Data construction. We use FineVideo's audio-visual data and annotations (Farré et al., 2024) to generate QA pairs with Qwen3.7-Max and convert them into four-option MCQs. All nine models independently answer the original set of approximately 4,000 single-hop questions, which remains the full 1-hop evaluation set in Section 3. Only questions answered correctly by all nine models are used to construct multi-hop questions. Each multi-hop question combines facts from these single-hop QA pairs, supported by relevant audio or visual segments. Together with the question and stated task rules, these facts must uniquely determine the answer without unstated specialized knowledge. We remove ambiguous, underspecified, or unsupported questions, as well as cases solvable from a single cue or commonsense even with answer options. Together, the original single-hop and resulting multi-hop sets draw on 2,703 videos; the latter contains approximately 5,000 questions, equally divided between 2-hop and 3-hop.

Evaluation protocol. Table 7 lists the evaluated models and their accuracy at each hop count. Models receive the question and its four options with the corresponding audio-visual inputs. For each model, a multi-hop question and its corresponding single-hop questions use the same audio-visual material and preprocessing settings, with each question answered in a separate call. Multi-hop evaluation provides no preceding single-hop questions, answers, evidence hints, or conversation history. All models use the original single-hop set and the final multi-hop sets defined above. Accuracy is the fraction of correctly answered questions within each set. For Figure 2a, domain scores are averaged across models, and the macro-average assigns equal weight to all five domains.

Table 7: Accuracy (%) by model and question hop count in the diagnostic study.

Model	1-hop	2-hop	3-hop
AVATAR-7B (Kulkarni & Fazli, 2026)	93.4	62.0	60.5
HumanSense (Qin et al., 2026)	94.6	64.6	65.2
Omni-R1 (Zhong et al., 2025)	94.0	65.0	62.5
EchoInk-R1-7B (Xing et al., 2025)	95.5	71.0	68.4
HumanOmniV2-7B (Yang et al., 2025b)	96.3	71.8	72.2
OmniVideo-7B (Cai et al., 2026)	95.1	69.5	67.5
Qwen2.5-Omni-3B (Xu et al., 2025a)	90.9	37.4	30.0
Qwen2.5-Omni-7B (Xu et al., 2025a)	93.8	60.0	58.0
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)	98.5	75.2	76.0

A.2 Local Optimization Analysis

Experimental setup. We use Qwen2.5-Omni-3B with the full question sets defined above. Each objective averages the cross-entropy loss for predicting the correct answer token, using probabilities normalized over the full vocabulary. We compute gradients for the Thinker's Transformer decoder layers and final normalization layer, with all other parameters fixed. Gradients are averaged across questions before normalization.

We partition the data into five video-level folds, keeping all questions from each video in the same fold (Fox & Monette, 2026). For each split, gradient estimation for constructing update directions and update-norm calibration use only the four training folds. All updates start from the same initial

checkpoint and are evaluated on the held-out fold, which does not inform update construction or step-size selection (Cawley & Talbot, 2010).

First-order analysis. Gradient projection is used to constrain parameter updates in multi-objective fine-tuning and continual learning (Farajtabar et al., 2020; Chaudhry et al., 2018; Huang et al., 2025; Qiu et al., 2026). Within each split, the following analysis uses the training-fold objectives and gradients, with the split index omitted for clarity. We write an update as $-\eta u$, where $u = g_t/\|g_t\|_2$ for the original update and $u = r_{t\perp s}/\|r_{t\perp s}\|_2$ for the projected update in Equation 1. Both directions have unit norm, so η is the Euclidean norm of the parameter change. The first-order change in the other training objective is $-\eta g_s^\top u$; Equation 1 sets this term to zero. For $c = g_E^\top g_R/(\|g_E\|_2\|g_R\|_2)$, the maximum first-order decrease in the target loss per unit update norm satisfies

$$\max_{\|u\|_2\le 1,\ g_s^\top u=0} g_t^\top u = \|r_{t\perp s}\|_2, \qquad \frac{\|r_{t\perp s}\|_2}{\|g_t\|_2} = \sqrt{1-c^2}. \tag{2}$$

Here u ranges over feasible directions. For any such u, $g_t^\top u = r_{t\perp s}^\top u$; Cauchy–Schwarz gives the bound. When $r_{t\perp s} \neq 0$, equality is attained at $u = r_{t\perp s}/\|r_{t\perp s}\|_2$, whose unit norm matches the update in Equation 1. Across the five splits, the mean gradient cosine similarity is 0.5391, and the mean of the split-specific first-order retention ratios is 84.23% at the same update norm. Held-out loss responses are measured after applying each split-specific update; training-gradient orthogonality does not impose a zero first-order change on held-out objectives.

We compare random perception halves E_1/E_2 with a size-matched E_1/R pair, projecting in both directions. The full reasoning set contains equal numbers of 2-hop and 3-hop questions. Sampling is stratified by hop count and question type, with independent held-out perception splits. We pool held-out directional derivatives across folds by question count before computing projected-to-original ratios. Table 8 shows similar retention on the two perception halves, whereas cross-task projections show higher target retention and lower retention on the other objective.

Table 8: First-order retention (%) on held-out objectives, averaged over three seeds. Values are projected-to-original ratios of predicted loss decrease.

Update	Target retention	Other retention
E_1 projected ($\perp E_2$)	56.6	53.6
E_2 projected ($\perp E_1$)	60.9	58.5
E_1 projected ($\perp R$)	71.8	5.7
R projected ($\perp E_1$)	85.0	13.1

Table 9: Loss responses at two update norms, with R also reported separately for 2-hop and 3-hop questions. The higher-step E/R responses correspond to Figure 2b. All values are percentages; positive reductions indicate loss decreases.

Step	Update	E reduction	R reduction			Target retention
			Overall	2-hop	3-hop	
Lower	E original	+1.675	+1.910	+1.867	+1.985	100.00
Lower	E projected	+1.130	+0.006	+0.008	+0.003	67.48
Lower	R original	+1.176	+3.331	+2.718	+4.397	100.00
Lower	R projected	-0.040	+2.860	+2.077	+4.224	85.89
Higher	E original	+3.632	+4.333	+4.201	+4.563	100.00
Higher	E projected	+2.699	+0.362	+0.466	+0.181	74.30
Higher	R original	+2.080	+7.358	+5.694	+10.254	100.00
Higher	R projected	+0.060	+6.013	+4.150	+9.255	81.72

Step sizes. Within each split, we calibrate the higher update norm on 128 questions drawn from the four training folds: 64 perception questions and 64 reasoning questions, equally divided between 2-hop and 3-hop. Starting from an initial candidate norm of approximately 1.44×10^{-2}, we halve the candidate norm until both original update directions satisfy the Armijo condition on their respective calibration objectives, using the calibration losses and directional derivatives with $\sigma = 0.1$ (Cavalcanti et al., 2024). This search selects $\eta_{\text{high}} \approx 7.20 \times 10^{-3}$ in all five splits. We then set $\eta_{\text{low}} = \eta_{\text{high}}/2 \approx 3.60 \times 10^{-3}$ to assess sensitivity to the update norm. In each split, we verify that this lower norm also satisfies the Armijo condition for both original directions on their respective calibration objectives. All four update directions use both norms, with each update starting from θ_0.

Table 10: Loss reductions (%) by held-out video-level fold at both update norms.

	E original		E projected		R original		R projected	
Step / held-out fold	L_E	L_R	L_E	L_R	L_E	L_R	L_E	L_R
Lower / 1	+1.181	+1.944	+0.934	+0.006	+0.834	+3.106	-0.042	+2.272
Lower / 2	+2.441	+1.589	+1.407	+0.006	+1.417	+3.252	-0.032	+2.854
Lower / 3	+1.474	+1.678	+0.956	+0.006	+1.183	+3.416	-0.060	+2.902
Lower / 4	+1.555	+1.850	+1.272	-0.002	+0.887	+2.789	-0.048	+2.655
Lower / 5	+1.749	+2.439	+1.132	+0.015	+1.490	+4.020	-0.020	+3.524
Lower / Pooled	+1.675	+1.910	+1.130	+0.006	+1.176	+3.331	-0.040	+2.860
Higher / 1	+2.794	+3.986	+2.342	+0.296	+1.739	+7.247	+0.076	+6.037
Higher / 2	+4.248	+4.072	+3.289	+0.354	+2.669	+7.180	+0.068	+5.959
Higher / 3	+3.748	+4.299	+2.877	+0.385	+1.848	+7.533	+0.031	+5.952
Higher / 4	+3.720	+4.505	+2.401	+0.331	+1.961	+6.991	+0.039	+5.529
Higher / 5	+3.628	+4.734	+2.544	+0.435	+2.208	+7.794	+0.091	+6.546
Higher / Pooled	+3.632	+4.333	+2.699	+0.362	+2.080	+7.358	+0.060	+6.013

After fixing both norms, we check sufficient decrease on each held-out fold for all four split-specific update directions at both norms, without modifying either norm. Let L_t^{val} denote the held-out objective and $h_t = \nabla L_t^{\mathrm{val}}(\theta_0)$. For $D_t = L_t^{\mathrm{val}}(\theta_0) - L_t^{\mathrm{val}}(\theta_0 - \eta u)$, the Armijo condition is $D_t \geq \sigma \eta h_t^\top u$, with $\sigma = 0.1$ and $h_t^\top u > 0$. Across the 40 checks (eight per split), $D_t/(\eta h_t^\top u)$ ranges from 0.328 to 0.512, satisfying the Armijo condition in every case. This establishes sufficient decrease on the held-out objectives without requiring accurate first-order predictions of the loss reduction.

Measured responses. For cross-validation, each question is evaluated only under the update constructed without its video-level fold. We pool question-level baseline losses and loss reductions across the five held-out folds before computing percentages, rather than averaging fold-level percentages. Writing the pooled mean loss reduction as $\bar{D}_t$ and the pooled initial mean loss as $\bar{L}_t(\theta_0)$, we report relative loss reduction as $100\bar{D}_t/\bar{L}_t(\theta_0)$ and target retention as $100\bar{D}_t^{\mathrm{projected}}/\bar{D}_t^{\mathrm{original}}$. For R, losses are pooled across 2-hop and 3-hop questions before computing percentages. Table 9 reports both update norms; its higher-step overall responses match Figure 2b. At the lower step, the projected perception update retains 67.48% of its target loss reduction and decreases reasoning loss by 0.006%. The projected reasoning update retains 85.89% of its target loss reduction and increases perception loss by 0.040%. Thus, the lower step also supports local partial decoupling of the two mean-loss objectives. Table 10 gives the held-out-fold breakdown.

B Training and Evaluation Details

B.1 Training Setup

Training data and preprocessing. Tables 11 and 12 summarize the training data. Text-only and audio-visual GRPO use the same questions in the main comparison.

For RQ3, Qwen3.7-Max generates fictional scripts and MCQs with answers and reasoning explanations from abstract scenarios, task types, and aggregate distribution statistics, without native audio-visual inputs (Figure 3). Scenarios cover daily life, work and craft, entertainment, sports, and social interaction. Table 14 reports sample counts by task category. Programmatic checks validate structure, timing, and answer formats; model-based checks assess consistency, answer uniqueness, and distractor plausibility. We manually review a subset of synthetic samples. Generation and validation prompts appear in Appendix D.2.

Table 11: Training sample counts for text-only and audio-visual training. "–" denotes an unused stage.

Model	**SFT**	**GRPO**
Script-based CoT (SFT) / MCQ (GRPO)		
Qwen2.5-Omni-7B	8,948	12,593
Qwen2.5-Omni-3B	8,948	17,524
Gemma-4-E4B-IT	8,948	15,744
Gemma-4-E2B-IT	8,948	15,744
Qwen3-Omni-30B-A3B-Instruct	29,747	–
Text-only LLM synthesis		
Qwen2.5-Omni-7B	9,000	11,807
Qwen2.5-Omni-3B	9,000	11,807
Gemma-4-E4B-IT	9,000	11,807
Gemma-4-E2B-IT	9,000	11,807
Audio-visual training		
Qwen2.5-Omni-7B	42,996	12,593

Table 12: Training data composition. Bold counts indicate configuration totals.

Text-only training (SFT)	Samples
General reasoning	**86,917**
SmallThoughts	9,575
Mixture-of-Thoughts	26,405
Bespoke-Stratos-17k	3,643
OpenThoughts-114k	11,247
OpenR1-Math-220k	14,520
OpenMathReasoning	21,527
Contextual QA	**40,000**
CoT-Drop	30,000
CoT-Collection	9,083
CoT-Scientific-RAG-Reasoning	917
Script-based CoT (MCQ + open-ended)	**40,484**
MCQ	8,948
open-ended	31,536
Script-based CoT (MCQ) + General reasoning	**95,865**
Script-based CoT (MCQ)	8,948
General reasoning	86,917

Text-only training (GRPO)	Samples
Contextual QA	**15,000**
HotpotQA	5,000
MuSiQue	5,000
2WikiMultiHopQA	5,000
Temporal QA	**10,000**
Video-T3 QA	10,000
General & math QA	**10,000**
Nemotron-CrossThink	10,000
Audio-visual training	
SFT: Audio-visual reasoning	**42,996**
AVRT-20K	40,109
IntentTrain	2,887
Audio-visual training (reduced data)	
SFT: Audio-visual reasoning	**685**
AVRT-20K	548
IntentTrain	137
GRPO: Audio-visual MCQ	**1,000**
OmniVideo-100K	1,000

Table 13: SFT and GRPO settings. GRPO batch sizes count completions; "–" marks unused stages.

Hyperparameter	Qwen2.5-Omni-7B	Qwen2.5-Omni-3B	Gemma-4-E4B-IT	Gemma-4-E2B-IT	Qwen3-Omni-30B-A3B-Instruct
SFT					
Learning rate	2×10^{-6}	5×10^{-6}	5×10^{-5}	5×10^{-5}	10^{-4}
Global batch	48	48	48	48	8
Epochs	1	3	1	1	1
GRPO					
Learning rate	10^{-5}	10^{-5}	10^{-5}	10^{-5}	–
Global batch	16	16	16	16	–
Epochs	1	1	1	1	–
Generations per prompt	8	8	8	8	–
Temperature	1.0	1.0	1.0	1.0	–
Top-p	0.99	0.99	0.95	0.95	–
Top-k	50	50	64	64	–
KL coefficient β	0.001	0.001	0.001	0.001	–
Clipping threshold ϵ	0.2	0.2	0.2	0.2	–
Repetition penalty	1.0	1.0	1.0	1.0	–

Training settings. We apply LoRA ($r = 16$, $\alpha = 32$) to the Thinker's Transformer decoder in Qwen-Omni and the language model in Gemma-4. Audio and vision encoders and multimodal projection modules remain frozen. SFT and GRPO use AdamW with linear learning-rate warmup over the first 5% of training steps, followed by cosine decay. Table 13 lists the hyperparameters.

GRPO uses a probability-ratio clipping interval of $[0.8, 1.2]$ with $\epsilon = 0.2$. Qwen2.5-Omni-7B uses the same SFT and GRPO hyperparameters for text-only and audio-visual training. For audio-visual training, we retain the audio track and sample frames at 1 FPS, with at most 180 frames per video and 200,704 pixels per frame. Other settings, including sequence-length limits, follow official base-model defaults.

Table 14: Synthetic training data by task category, with counts and percentages for each stage.

Task category	SFT	GRPO
Perceptual understanding	2,098 (23.31%)	2,706 (22.92%)
Contextual reasoning	2,426 (26.96%)	3,090 (26.17%)
Temporal and causal reasoning	2,231 (24.79%)	2,952 (25.00%)
Social understanding	2,245 (24.94%)	3,059 (25.91%)

Rewards. GRPO combines rule-based accuracy and format rewards as $R = R_{\text{acc}} + 0.1R_{\text{format}}$. The accuracy reward checks extracted final answers against references using task-specific verification, including option matching for MCQs and answer verification for mathematical problems. The format reward checks the required structure for reasoning and final answers.

Implementation and compute accounting. We train on eight NVIDIA A100 PCIe 80 GB GPUs. We reproduce external baselines with their official training datasets and original training stages, measuring GPU-hours on the same hardware. We count non-padding training input tokens using each

model's tokenizer and training template, including special tokens and expanded audio and visual token positions. $\mathcal{I}$ indexes sample occurrences across all epochs within each stage. L_i^{SFT} denotes the complete sequence length including assistant targets, and L_i^{prompt} denotes the prompt length before generation. With $G = 8$ completions per prompt, counts in billions are

$$T_{\mathrm{SFT}} = 10^{-9} \sum_{i \in \mathcal{I}} L_i^{\mathrm{SFT}}, \qquad T_{\mathrm{GRPO}} = 10^{-9} G \sum_{i \in \mathcal{I}} L_i^{\mathrm{prompt}}. \tag{3}$$

T_{GRPO} excludes generated completion tokens.

Training compute is $H_{\mathrm{GPU}} = \sum_j W_j t_j / 3600$ GPU-hours, where W_j and t_j denote the GPU count and wall-clock duration in seconds of interval j. GRPO intervals include online rollout generation and policy optimization.

B.2 Evaluation Setup

We follow each benchmark's official evaluation settings for decoding, maximum output length, and audio-visual sampling. For Gemma-4, we retain the benchmark protocols but use official model inference defaults and set the maximum output length to 1,024 tokens. Qwen3.7-Max assesses response quality on SocialOmni How and answers Omni-Cloze's cloze-style MCQs using captions generated by the evaluated model. For Video-MME-v2, we report average accuracy (Avg Acc). For perceptual evaluation, we restrict AVQA to the Which, Come From, Happening, and Where question types, and NExT-QA to the Descriptive subset. AVHBench Overall is the pooled accuracy across Audio-driven Video Hallucination and Video-driven Audio Hallucination. Macro Avg. is the unweighted mean of nine reasoning scores or five perception scores: AVHBench Overall, MUSIC-AVQA v1 and v2, AVQA, and NExT-QA. Geometric Mean Improvement uses the same score sets. Evaluation system prompts appear in Appendix D.1.

C Further Analyses

C.1 Training Data Mixtures

To complement RQ2, we mix Script-based MCQs with either open-ended questions or General reasoning data during SFT, using CoT supervision. The subsequent GRPO phase uses the same Script-based MCQ data for all configurations. Table 15 reports results for these mixtures, while Table 4 provides the reference results without data mixing.

Table 15: Effect of SFT data mixtures on omni-modal reasoning. OE denotes open-ended questions. Parenthesized values show absolute changes from the base model.

Training Configuration	Omni-modal Reasoning									Aggregate	
	Daily-Omni	Omni-Cloze	Omni Video	SocialOmni			Video-Holmes	Video-MME-v2	World Sense	Macro Avg.	Geo. Mean Improv. (%)
				Who	When	How					
Qwen2.5-Omni-7B	62.07	21.90	29.30	36.75	57.42	56.67	24.40	24.31	45.40	39.80	0.00
SFT: Script-based CoT (MCQ+OE)	56.98 (−5.09)	27.12 (+5.22)	32.60 (+3.30)	37.65 (+0.90)	57.42 (+0.00)	75.00 (+18.33)	39.63 (+15.23)	24.76 (+0.45)	45.10 (−0.30)	44.03 (+4.23)	+12.20
+ **GRPO:** Script-based MCQ	64.16 (+2.09)	27.55 (+5.65)	36.70 (+7.40)	39.95 (+3.20)	58.85 (+1.43)	69.71 (+13.04)	43.33 (+18.93)	28.58 (+4.27)	47.50 (+2.10)	46.26 (+6.46)	+19.25
SFT: Script-based CoT (MCQ) + General reasoning	56.39 (−5.68)	26.01 (+4.11)	35.00 (+5.70)	36.40 (−0.35)	58.85 (+1.43)	71.88 (+15.21)	36.91 (+12.51)	24.40 (+0.09)	44.70 (−0.70)	43.39 (+3.59)	+10.62
+ **GRPO:** Script-based MCQ	65.66 (+3.59)	28.03 (+6.13)	37.80 (+8.50)	40.00 (+3.25)	61.72 (+4.30)	70.83 (+14.16)	47.52 (+23.12)	30.80 (+6.49)	48.70 (+3.30)	47.90 (+8.09)	+23.66

SFT data mixing does not improve final average performance. Mixing Script-based MCQs with open-ended questions raises the SFT Macro Avg. from 39.92 to 44.03. After GRPO, however, the mixture reaches 46.26, below the 49.39 obtained using MCQ data alone during SFT. Mixing with General reasoning data similarly raises the SFT Macro Avg. to 43.39, but the final score of 47.90 remains below the 49.39 and 48.91 obtained using the two SFT data types separately. Both mixtures nevertheless improve all nine evaluation dimensions over the base model. With the same subsequent GRPO training, SFT on Script-based MCQs alone achieves a higher reasoning Macro Avg. than either of the two tested SFT mixtures.

C.2 Agentic Reasoning Transfer

RQ1–RQ3 establish that text-only reasoning training transfers to native audio-visual reasoning tasks. Here, we examine whether training on text-only agent trajectories enables transfer to omni-modal agentic tasks that require multi-turn tool use and cross-modal evidence integration.

Experimental setup. We train Qwen2.5-Omni-7B with text-only SFT followed by GRPO, using agent training data from ARPO (Lu et al., 2025), MiroVerse (Team et al., 2025), OpenSeeker (Du et al., 2026), and ToolMind (Yang et al., 2025a). SFT uses 16,000 training samples over 2 epochs, while GRPO uses 8,000 training samples over 1 epoch. We follow the official OmniGAIA (Li et al., 2026b) evaluation setup and report Pass@1 (0–100) for Easy, Medium, Hard, and Overall. We compare with the original Qwen2.5-Omni-7B and OmniAtlas-SFT with and without OmniDPO (Li et al., 2026b).

Text-only reasoning training transfers to omni-modal agentic tasks. As shown in Table 16, text-only SFT followed by GRPO raises the OmniGAIA Overall score from 3.6 to 13.1, nearly matching the 13.3 achieved by native omni-modal training. The gains are particularly pronounced on the Medium and Hard subsets, where the text-only route reaches 15.0 and 6.4, respectively. Together, these results suggest that text-only reasoning training transfers to a more demanding form of omni-modal reasoning: agentic tasks requiring multi-turn tool use and cross-modal evidence integration. This result broadens the scope of text-only reasoning training from benchmark reasoning to agentic reasoning in Omni-LLMs.

Table 16: Agentic reasoning performance on OmniGAIA (Pass@1). Parenthesized values show absolute changes from the base model. **Bold** and underlined scores denote the best and second-best results, respectively.

Model	Difficulty Levels			Overall
	Easy	Medium	Hard	
Qwen2.5-Omni-7B	8.2	1.3	1.3	3.6
Omni-modal Training				
OmniAtlas-SFT	19.7	9.4	2.6	11.4
+ OmniDPO	**22.1**	11.3	3.9	**13.3**
Text-only Training				
Text-only SFT	16.4	8.1	**7.7**	10.8
+ GRPO (Ours)	14.8	**15.0**	6.4	13.1
	(+6.6)	(+13.7)	(+5.1)	(+9.5)

D Prompts

D.1 Training and Evaluation

Figures 5–8 show the Qwen-Omni and Gemma-4 prompts. The {input} placeholder denotes textual context, such as scripts, caption sequences, and passages, or video with or without audio; it is omitted when absent. Audio-visual training is used only for Qwen2.5-Omni-7B. Gemma-4 uses thinking mode with its native reasoning format. Evaluation uses the video-and-audio template, while NExT-QA uses video only. For SocialOmni How, we augment the official prompt with our reasoning instructions. Omni-Cloze uses its official prompts (Ma et al., 2026).

Qwen2.5-Omni and Qwen3-Omni

Training

System message

Reason and answer the question. Provide your reasoning between the <think> and </think> tags, and then give your final answer between the <answer> and </answer> tags.

User message

{input}
The question is: {question}
The options are:
A. {option_A}
B. {option_B}
...

(a) Training

Qwen2.5-Omni and Qwen3-Omni

Evaluation: video and audio

System message

Based on the given video and audio, reason and answer the question. Provide your reasoning between the <think> and </think> tags, and then give your final answer between the <answer> and </answer> tags.

User message

{input}
The question is: {question}
The options are:
A. {option_A}
B. {option_B}
...

(b) Video-and-audio evaluation

Figure 5: Training and video-and-audio evaluation prompts for Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, and Qwen3-Omni-30B-A3B-Instruct.

Qwen2.5-Omni and Qwen3-Omni

Evaluation: video

System message

Based on the given video, reason and answer the question. Provide your reasoning between the <think> and </think> tags, and then give your final answer between the <answer> and </answer> tags.

User message

{input}
The question is: {question}
The options are:
A. {option_A}
B. {option_B}
...

(a) Video-only evaluation

Qwen2.5-Omni and Qwen3-Omni

Evaluation: SocialOmni How

System message

Based on the given video and audio, reason and answer the question. Provide your reasoning between the <think> and </think> tags, and then give your final answer between the <answer> and </answer> tags.

User message

{input}
The question is: {question}
In the final answer, provide only a concise natural utterance without explanations.

(b) SocialOmni How

Figure 6: Video-only and SocialOmni How evaluation prompts for Qwen2.5-Omni and Qwen3-Omni.

Gemma 4

Training

System message

Select the correct option. In the final answer, output only its option label exactly as shown.

User message

{input}
The question is: {question}
The options are:
A. {option_A}
B. {option_B}
...

(a) Training

Gemma 4

Evaluation: video and audio

System message

Based on the given video and audio, select the correct option. In the final answer, output only its option label exactly as shown.

User message

{input}
The question is: {question}
The options are:
A. {option_A}
B. {option_B}
...

(b) Video-and-audio evaluation

Figure 7: Training and video-and-audio evaluation prompts for Gemma-4-E4B-IT and Gemma-4-E2B-IT.

Gemma 4

Evaluation: video

System message

Based on the given video, select the correct option. In the final answer, output only its option label exactly as shown.

User message

{input}
The question is: {question}
The options are:
A. {option_A}
B. {option_B}
...

(a) Video-only evaluation

Gemma 4

Evaluation: SocialOmni How

System message

Based on the given video and audio, answer the question.

User message

{input}
The question is: {question}
In the final answer, provide only a concise natural utterance without explanations.

(b) SocialOmni How

Figure 8: Video-only and SocialOmni How evaluation prompts for Gemma-4-E4B-IT and Gemma-4-E2B-IT.

D.2 Synthetic Data Generation

Figures 9 and 10 show the prompts for episode planning, visual expansion, MCQ and reasoning generation, and quality assessment. Named placeholders in braces denote variable content; JSON blocks illustrate output formats.

Episode planning

Create a compact plan for one fictional, physically realizable video based on the {scenario}, following {task_requirements} and {segment_headers}. The request contains only text. Never refer to a source video or dataset.

Return one JSON object with exactly these top-level fields:

```
{
  "title": "{title}",
  "summary": "{summary}",
  "entities": [
    {"name": "{entity_name}", "label": "people|object|animal|place",
      "description": "{visible_or_audible_details}"}
  ],
  "segments": [
    {
      "segment_index": 0,
      "visual_facts": ["{visual_fact}"],
      "speech": [
        {"speaker": "{entity_name}", "text": "{utterance}", "time_slot":
          "early|middle|late|full"}
      ],
      "sounds": [
        {"source": "{sound_source}", "sound": "{objective_sound_description}",
          "time_slot": "early|middle|late|full"}
      ],
      "state_changes": [
        {"entity": "{entity_name}", "before": "{observable_prior_state}", "after":
          "{observable_later_state}"}
      ]
    }
  ],
  "evidence_opportunities": [
    {
      "segment_indices": [0],
      "supported_task": "{task_type}",
      "audio_role": "{audio_evidence}",
      "visual_role": "{visual_evidence}"
    }
  ]
}
```

The program owns IDs, timestamps, task labels, option letters, and formatting. List speech events in their intended chronological order; the program assigns them non-overlapping time ranges within each segment. Copy every supplied segment_index exactly once and in order. Do not write a question, answer, option, distractor, quiz, rubric, or judge verdict.

Create the fictional episode from the specified abstract scenario and task type. Invent new people, objects, wording, quantities, and event details. Draft the episode from scratch, maintaining a coherent sequence of events.

The episode must always contain camera-observable action, natural speech with stable speaker binding, and at least one attributable non-speech sound. Dialogue should serve the activity rather than explain the scene to a future viewer. Avoid self-identification, staged demonstrations, poetic prose, surreal anomalies, and artificial twists unless the requested format requires them.

Use a short factual title and an objective summary. Never mention segments, visual density, speech density, modalities, evidence, anchors, narrative structure, targets, or generation constraints. A visual fact must not claim that something is silent, audible, heard, loud, or otherwise characterize sound.

Every speech speaker, including an off-screen narrator, caller, or producer, must first appear in entities under exactly the same speaker string. For first-person footage, add a named camera-holder entity instead of inventing an undeclared Narrator. Before returning JSON, check every speech speaker against the entity names exactly. Do not use `silent` or `silently` in visual facts; describe only the visible posture or motion.

Visual facts are concise planning notes, not full caption prose. Speech and sound strings are final and must not be regenerated later. State changes must be physically consistent. Evidence opportunities are optional natural incidents; use an empty list when the requested task is unsupported, and an empty string for an absent audio or visual evidence role. For future prediction, an evidence opportunity must contain at least two chronological segments: one or more premise segments followed by a distinct outcome segment. Do not place the premise and its conclusion in the same segment.

(a) Episode planning.

Visual expansion

Expand the visual facts in {episode_plan} into objective camera-observable descriptions, following {segment_constraints}.

Return one JSON object:

```
{
  "segments": [
    {
      "segment_index": 0,
      "visual_events": [
        {
          "time_slot": "early|middle|late|full",
          "align_to_audio_event_id": "{audio_event_id}",
          "text": "{visual_description}"
        }
      ]
    }
  ]
}
```

Return every requested segment exactly once and in the supplied order. The program assigns timestamps. Use an audio event id only when a visible action truly co-occurs with it; otherwise use null.

Describe only facts visible to a camera: entities, appearance, actions, objects, spatial relations, camera framing, and observable state changes. Audio timing is context for synchronization only. Do not quote, paraphrase, or infer speech; do not describe sounds, intentions, causes, emotions, hidden states, symbolism, or future questions. Do not introduce new entities, objects, text overlays, camera cuts, or anomalies. Match the requested total visual-event and prose-character budgets approximately. Character targets are soft and may vary by about 35%; reach them through chronological observable detail, spatial continuity, object state, body movement, and framing—not padding, adjective stacks, repetition, or invented micro-cuts.

Write like a factual human video caption, not like a compliance report. Do not use rubric or process language such as "observable state change", "spatial relationship", "documenting", or "visual field". Avoid stock claims that the camera maintains its framing, that objects remain stationary, that nothing enters the frame, or that lighting stays consistent unless that fact changes or disambiguates the event. Each sentence must add a concrete visible detail rather than merely assert continuity or absence.

The supplied {prior_entity_states} records the last known state accumulated from earlier segments, including earlier chunks. Preserve each state unless the current segment explicitly supplies a new state change. Never reset a door, container, object position, clothing item, or other tracked state merely because the earlier action is outside the current chunk.

The visual text must not contain auditory assertions such as audible, heard, speech content, spoken phrase, vocal cadence, dialogue, utterance, loud, silent, or quietly. Device labels such as an audio-level meter may be described only when visibly displayed. Describe visible mouth, hand, face, and object motion without claiming what is being said or heard. Follow each segment's own visual-event count and character range rather than using one repeated paragraph template.

(b) Visual expansion.

Figure 9: Prompts for planning fictional episodes and expanding visual planning facts into descriptions.

Grounded MCQ and analysis

Create at most one grounded multiple-choice training item using the fictional script in {script_evidence}, following {task_requirements}.

Use the aggregate {distribution_statistics} for task types, answer formats, and durations only when compatible with the primary task and the available fictional evidence. The task and modality requirements always take precedence. Never mention source names, dataset names, task labels, or pipeline metadata in the question, options, or analysis.

Return one JSON object in either form:
{"status": "NONE", "reason": "{reason}"}

or:
```
{
  "status": "SUPPORTED",
  "evidence_segment_indices": [0],
  "question": "{question}",
  "premise_event": "{premise_event}",
  "correct_answer": "{answer_text}",
  "distractors": [
    {
      "text": "{distractor}",
      "why_plausible": "{plausibility_explanation}",
      "why_false": "{evidence_against_distractor}"
    }
  ],
  "analysis": {
    "connections": "{evidence_connections}",
    "correct_answer_explanation": "{answer_explanation}"
  }
}
```

Use exactly the requested number of answer choices: one correct answer and the required number of distractors. Do not assign option letters; the program shuffles options and builds the final letter-aware explanation. Never refer to the first, second, or numbered distractor. The correct answer must be uniquely true for the event or relation asked about. Distractors must be grammatically parallel, similar in specificity and length, and false as answers to the exact question rather than merely unmentioned. A temporal distractor may describe a real event elsewhere in the episode when the script explicitly places it outside the question's anchored moment; this is a valid hard negative, not a factual contradiction. Every option must refer to a semantically distinct answer; do not use a named person or object and a role-based description of that same entity as separate choices.

Follow the specified relationship between audio and visual evidence:

- Audio-Guided: use a distinctive speech or non-speech audio cue as the temporal anchor, then ask for a visible fact aligned with that cue.

- Vision-Guided: use a distinctive visible event, object, action, or readable text as the temporal anchor, then ask for an aligned speech or sound fact.

- Audio-Visual Fusion: require a relation or reasoning chain that materially combines both modalities.

For audio-visual fusion, the combined evidence must identify one answer, while each modality alone must leave at least two supplied choices plausible. For future prediction, treat {premise_segment_indices} as the observed context and the later {outcome_segment_indices} as held-out continuation evidence. Set question to an empty string and state only the observed event before the cutoff in premise_event. The program uses this premise to form a question about its direct conclusion or likely next event. The correct answer must describe an event after the observation cutoff. Return both premise and outcome indices in chronological order. Do not turn this task into concurrent speech retrieval or ordinary present-event localization. For other tasks, write a standalone question and leave premise_event empty. Connect the relevant evidence step by step and explain why it supports the correct answer. For unsupported tasks, return NONE with a brief reason. Do not force an unsupported question. The question and answer must not mention segments, captions, modalities, datasets, or the construction process.

(a) Question and reasoning generation.

Consolidated quality judge

Audit {candidate_mcq} using {script_evidence} and {task_requirements}. Do not repair or rewrite it.

Return one JSON object:
```
{
  "question_id": "{question_id}",
  "predicted_answer": "A|B|C|D|E|F|UNKNOWN",
  "keep": true,
  "checks": {
    "full_context_unique_and_grounded": true,
    "question_natural_and_task_aligned": true,
    "distractors_plausible_and_false": true,
    "no_summary_or_option_style_leakage": true,
    "episode_coherent_and_physical": true,
    "speech_natural_and_bound": true,
    "visual_and_sound_objective": true,
    "anchor_modality": "audio|visual|audio_visual|none",
    "answer_modality": "audio|visual|audio_visual|none",
    "audio_only_sufficient": false,
    "visual_only_sufficient": false,
    "future_answer_after_cutoff": true,
    "audio_only_plausible_answers": ["A", "B"],
    "visual_only_plausible_answers": ["A", "C"]
  },
  "reason": "{reason}"
}
```

Judge only supplied evidence. Audit the complete compact episode outline for physical continuity, objective visual/sound descriptions, attributable sound, and natural speech with stable speakers. UNKNOWN means the full context does not support one unique option. For audio-visual fusion, keep an item only if the combined evidence supports a unique answer and each modality alone leaves at least two listed options plausible.

Check the specified relationship between audio and visual evidence:

- Audio-Guided uses a distinctive speech or non-speech audio cue to locate a visible answer fact.

- Vision-Guided uses a distinctive visible cue to locate a speech or sound answer fact.

- Audio-Visual Fusion materially combines evidence from both modalities.

For future prediction, verify that the question states an observed premise, the correct answer is supported only by {held_out_outcome_context} after the observation cutoff, and the question asks for a direct conclusion or likely next event. Set future_answer_after_cutoff to true only when all three conditions hold. For other tasks, set it to false.

Report the actual anchor and answer modalities from the supplied evidence; do not merely copy the requested relationship or infer these roles from whether either modality alone is sufficient to answer the question.

A distractor may describe a fact that occurs elsewhere in the episode and still be false for the event or moment asked about. Treat such explicit temporal, identity, attribute, order, or cross-modal mismatches as valid hard negatives. For every task, require a coherent episode, natural task alignment, mutually distinct plausible false answers to the exact question, and no answer leakage. Two options that refer to the same person or object by a name and a description are not distinct. Give one concise, evidence-based reason; the keep decision and its reason must agree with every applicable check.

(b) Quality assessment.

Figure 10: Prompts for generating MCQs and reasoning explanations from fictional scripts and assessing synthetic sample quality.

E CASE STUDIES

Figures 11–14 show examples from our final Qwen2.5-Omni-7B model in RQ5, obtained through text-only reasoning training followed by reduced-data audio-visual refinement.

Video



00:06 00:15 00:23 00:29

Speech

00:00–00:10	A story to begin with ultimate glamour and color.
00:10–00:20	A dressing embedded with sheer love and affection, making the moment to remember and cherish.
00:20–00:30	Dedicated to ensure that you look the best on your most precious day, the wedding day.

Question

In what chronological order did these narration audio events occur relative to their corresponding visual preparation stages?

A. Wedding dedication speech (jewelry assistance) → Sheer love speech (headpiece placement) → Glamour narration (hairspray application)

B. Electronic music pulses (logo display) → Pop music (flower arrangement) → Processed voiceover (bracelet adjustment)

C. Glamour narration (headpiece placement) → Sheer love speech (flower arrangement) → Wedding dedication speech (jewelry assistance)

D. Sheer love speech (bracelet adjustment) → Wedding dedication speech (hairspray) → Glamour narration (logo display)

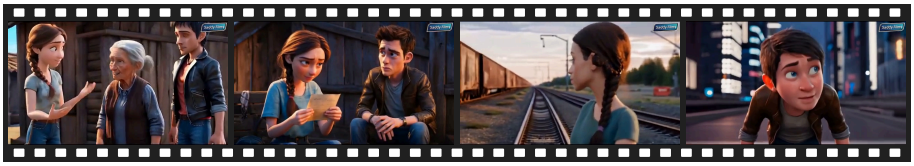
Model response

<think>
The chronological order of the narration audio events in relation to their corresponding visual preparation stages is as follows:
1. Glamour: The narration mentions glamour and color at the beginning, which corresponds to the woman having her hair styled with a headpiece.
2. Sheer love: The narration includes a mention of a dress embedded with sheer love, which corresponds to the placement of flowers in her hair.
3. Wedding dedication: The narration ends with a dedication to ensure the best on the most precious day, which corresponds to the final touches being made, including adjusting jewelry and a bracelet.

Option C correctly reflects this sequence.
</think>
<answer>
C
</answer>

Figure 11: Example from Daily-Omni.

Video



01:24 01:56 02:47 03:03

Speech

00:30–01:00	This is the village. I'm sure of it. She must be around here somewhere. Hi there. You seem new around here. Need any help? Yeah, actually, I'm looking for someone named Elena. Do you know her? Elena? She used to live here, but that was years ago.
01:30–02:00	She left a letter before she disappeared. This is letter? She left this at my home. Alex, if you're reading this, it means you've found your way to me. I've gone to the city to chase my dreams. I hope you understand. And maybe, just maybe, we'll meet again.
02:00–02:30	Then I have to go to the city. I have to find her. I think she'd like that. Thank you, Lara, for everything. Good luck, Alex. Find her. Come back safe, Alex. We've all been waiting. We are sisters. We were so happy together. Every moment filled with joy. But Elena always dreamed big. Beyond our little village. One day,

Question

What was the main reason why Alexander ultimately decided to go to the big city?

A. Lala suggested that he go to the city.

B. The old lady revealed Elena's address.

C. Tired of village life

D. The train just passes through the city.

E. Cities have more job opportunities.

F. Elena's letter hints that a reunion might be possible.

Model response

<think>The video depicts Alexander's journey as he searches for his friend Elena. The old lady reveals that Elena had always dreamed of the big city and left a letter before disappearing. When Alexander reads the letter, it mentions that she had gone to the city, hinting that a reunion might be possible. This revelation is the primary reason that prompts Alexander to go to the city to find Elena and possibly reunite with her. His decision follows the letter's hint of reunion, not dissatisfaction with village life, Lala's advice, or an address from the old lady. The train passing through the city is not a factor in Alexander's decision. Therefore, the correct answer is F.</think>
<answer>**F**</answer>

Figure 12: Example from Video-Holmes.

Video



00:04 00:48 01:15

Speech

00:00–00:10	portrait shoot. We're going to be shooting with the 56mm f1.2 portrait killer lens. Now, they
00:26–00:46	It's gonna be a nice little shoot. We're gonna be shooting in this record store that I found in downtown Denver. So I'm gonna meet up with one of the models there that I reached out to via Instagram. She's gonna meet us there. Her name is Karina. And yeah, we're gonna do a small little shoot with this bad boy. I'm also gonna be using the 18 to 55 millimeter
00:46–00:57	lens that is on my body of the XT 20 right now but primarily I want to shoot with this here and hopefully get some nice like bokeh some nice like sharp images

Question

What type of lens does the man primarily intend to use for today's filming, according to the video?

A. An 18-55mm zoom lens.

B. A 56mm f/3.2 portrait lens.

C. A 56mm f/1.2 portrait lens.

D. A 24mm f/1.2 portrait lens.

Model response

<think>
The man states, "today it's gonna be a nice little shoot, we're gonna be shooting in this record store that I found in downtown Denver. So I'm gonna meet up with one of the models there that I reached out to via Instagram. She's gonna meet us there. Her name is Karina, and yeah, we're gonna do a small little shoot with this bad boy. I'm also gonna be using the 18-55mm lens that it's on my body of the XT20 right now, but primarily I want to shoot with this here." This indicates that the 56mm f/1.2 portrait lens is the primary lens he intends to use for the specific purpose mentioned.
</think>
<answer>
C. A 56mm f/1.2 portrait lens.
</answer>

Figure 13: Example from WorldSense.

Video



00:05 00:29 00:55

Question

What kind of musical instrument is it?

A. accordion	B. acoustic_guitar	C. bagpipe
D. banjo	E. bassoon	F. cello
G. clarinet	H. congas	I. drum
J. electric_bass	K. erhu	L. flute
M. guzheng	N. piano	O. pipa
P. saxophone	Q. suona	R. trumpet
S. tuba	T. ukulele	U. violin
V. xylophone		

Model response

<think>
The video shows two individuals playing a plucked string instrument. The instrument has a light-colored body and a visible sound hole with a rosette. The person on the left is wearing a green shirt and is strumming the instrument, while the person on the right is wearing a plaid shirt and a cowboy hat, also playing a similar instrument. The guitar has a distinct shape and size, consistent with an acoustic guitar. The presence of a pickguard and the neck of the instrument further supports this identification. None of the other options in the list match the characteristics of the instrument shown.
</think>
<answer>
B: acoustic_guitar
</answer>

Figure 14: Example from MUSIC-AVQA v1.