CreditOmni: Outcome-Constrained Segmental Credit Assignment For Multimodal Reasoning

Anonymous authors
Paper under double-blind review

Abstract

Multimodal reasoning requires models to interpret audiovisual evidence and use it to support their final answers. Reinforcement learning can improve this ability with verifiable answer rewards, but these offer limited guidance on evidence and reasoning quality. In particular, responses with the same answer score may differ in the quality of their evidence and reasoning. In this paper, we propose CreditOmni, a framework with two components. (1) Outcome-constrained segmental credit assignment distinguishes responses with the same answer score by evaluating evidence and reasoning separately. Process scores adjust the magnitudes of answer-derived advantages while preserving their signs. (2) Two-stage group relative policy optimization follows supervised fine-tuning and proceeds from segmental learning to full-sequence refinement. Experiments on IntentBench, Daily-Omni, and WorldSense show that CreditOmni achieves the best overall performance among the evaluated open-source 7B models. It outperforms Qwen2.5-Omni-7B by 7.52, 13.37, and 3.24 percentage points, respectively. Our code is available at https://anonymous.4open.science/r/creditomni/.

1 Introduction

Multimodal reasoning requires integrating visual, auditory, and linguistic cues to understand complex scenes and make evidence-based judgments (Xu et al., 2025; Chen et al., 2026). Models must identify question-relevant observations and explain how they support a conclusion. Evidence description and reasoning analysis therefore serve distinct roles, even when they lead to the same final answer. Reinforcement learning with verifiable rewards has improved reasoning in language and video models (Guo et al., 2025; Feng et al., 2025). Yet identical answer scores can accompany evidence descriptions and reasoning of different quality. Process supervision provides finer feedback by evaluating intermediate reasoning steps (Lightman et al., 2024; Wang et al., 2024), and multimodal process rewards extend this assessment to perceptual consistency (Luo et al., 2025). These signals make process quality available for training, but their usefulness also depends on how they are assigned within a response.

Mixing answer and process rewards before group normalization can shift update directions away from those determined by answer rewards alone. Evidence and reasoning tokens share the same advantage coefficient, even when the two segments differ in quality. To address these issues, we propose CreditOmni, a framework for outcome-constrained segmental credit assignment in multimodal reasoning. CreditOmni organizes each response into context, think, and answer (CTA) segments. It evaluates evidence and reasoning separately across three dimensions: social intent, temporal-causal consistency, and emotional and epistemic state. Answer advantages determine update directions, while process scores adjust advantage magnitudes for evidence and reasoning. Training begins with cold-start supervised fine-tuning (SFT) to establish the structured response format. Two stages of group relative policy optimization (GRPO; Shao et al., 2024) then progress from segment-level learning to full-sequence optimization guided by answer and format rewards. Figure 1 summarizes the motivation and overall design of CreditOmni.

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: **Motivation and overview of CreditOmni.** The six modules connect multimodal task demands and the limitations of answer-only feedback to CreditOmni's design. Three process rewards evaluate evidence and reasoning separately. Credit assignment preserves answer-derived update directions. Training combines structured initialization with two GRPO stages.

Our contributions are summarized as follows:

- We propose CreditOmni, a framework for outcome-constrained segmental credit assignment that combines verifiable answer advantages with separate evaluations of evidence and reasoning.
- We introduce a two-stage reinforcement learning pipeline with three process rewards. They evaluate evidence accuracy and inferential support, covering characters' goals, event order, explanatory evidence, state attribution, and access to information.
- Extensive experiments confirm that CreditOmni outperforms the compared open-source 7B baselines in overall performance on IntentBench, Daily-Omni, and WorldSense. Ablation studies further show that removing any individual process reward reduces Stage I overall scores on all three benchmarks.

2 Related Work

Multimodal Understanding And Reasoning. LLaVA-OneVision unifies single-image, multi-image, and video tasks to study transfer across visual inputs (Li et al., 2024a). Qwen2.5-Omni jointly supports text, image, audio, and video inputs (Xu et al., 2025). Their capabilities are exam-

ined from complementary perspectives: MVBench covers dynamic video understanding (Li et al., 2024b); MovieGraphs, DramaQA, and Social-IQ examine character relationships, hierarchical stories, and social situations (Vicol et al., 2018; Choi et al., 2021; Zadeh et al., 2019). MUSIC-AVQA, Daily-Omni, and WorldSense extend evaluation from sound–visual object associations to audiovisual understanding across diverse scenarios (Li et al., 2022; Zhou et al., 2025; Hong et al., 2026). Built on Qwen2.5-Omni, CreditOmni combines audiovisual evidence with question information for multimodal understanding and reasoning across these scenarios.

Reinforcement Learning For Reasoning. DeepSeekMath introduces GRPO, using relative rewards among responses to the same question for policy optimization (Shao et al., 2024). DeepSeek-R1 uses verifiable outcomes for mathematical and coding tasks (Guo et al., 2025), while Video-R1 applies rule-based reinforcement learning to video reasoning (Feng et al., 2025). Process supervision adds intermediate feedback: human step annotations train process reward models (Lightman et al., 2024), and Math-Shepherd constructs step labels automatically for verification and reinforcement learning (Wang et al., 2024). In multimodal tasks, URSA incorporates perception and reasoning assessments into GRPO training (Luo et al., 2025); VisualPRM instead uses step-level scores to select answers at inference (Wang et al., 2026). AVATAR combines historical-sample reuse with reasoning-position advantage shaping for audiovisual reasoning (Kulkarni & Fazli, 2026). OmniVideo-R1 combines question-oriented evidence localization and modality information fusion for reinforcement learning (Chen et al., 2026). CreditOmni compares same-question candidates and evaluates evidence and reasoning separately, using verifiable answer feedback to constrain segmental credit assignment.

3 CreditOmni

3.1 Method Overview

We study multimodal reasoning with audiovisual inputs and answer-annotated training data. Input x comprises video, audio, a question, and its options or label vocabulary. The response $\widetilde{y}_i$ has three segments: context describes evidence c_i, think develops reasoning h_i, and answer gives the final answer a_i.

Training begins with cold-start supervised fine-tuning (SFT) to establish the CTA response format. We fine-tune Qwen2.5-Omni-7B (Xu et al., 2025) on annotated samples from Video-R1 (Feng et al., 2025) and Social-IQ 2.0 (Wilf et al., 2023). The resulting policy π^{SFT} initializes the first of two GRPO stages. In Stage I, a fixed OmniJudge evaluates each candidate's evidence and reasoning separately. These evaluations are combined with answer advantages for outcome-constrained segmental credit assignment. The policy loss is averaged within each segment before the segment losses are combined. Stage II starts from the resulting policy π^{I} and optimizes complete responses using a joint answer-and-format reward. The stages differ in the scope of feedback and the aggregation of losses, as detailed in Sections 3.2 and 3.3, respectively.

Both reinforcement learning stages use GRPO, with their shared optimization framework defined below. For each question, the old policy π_{old} samples G candidate responses, with $G = 8$ in our experiments. Let T_i denote the number of tokens in response i, A_i its group-relative advantage, and β the KL regularization coefficient. With full-response averaging and KL regularization, the base GRPO loss (Shao et al., 2024) is

$$\mathcal{L}^{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\left[\ell_{i,t}^{\text{pg}}(\theta; A_i) + \beta k_{i,t}(\theta)\right]. \tag{1}$$

In this objective, all tokens in a response share the same advantage A_i. Normalizing the answer reward r^{ans} within its group gives the answer advantage:

$$A^{\text{ans}} = \frac{r^{\text{ans}} - \bar{r}_g^{\text{ans}}}{\sigma_g^{\text{ans}} + \varepsilon_r}, \qquad \varepsilon_r > 0. \tag{2}$$

Here, $\bar{r}_g^{\text{ans}}$ and σ_g^{ans} are the group mean and population standard deviation of answer rewards. The numerical stabilizer ε_r is a positive constant fixed before training. Responses with rewards above the group mean receive positive advantages, whereas those below it receive negative advantages.

For a single response, the policy loss uses the current-to-old probability ratio w_t with clipping threshold ε. We estimate the KL divergence between the current policy and the fixed reference policy π_{ref} using k_t:

$$
\begin{aligned}
w_t(\theta) &= \frac{\pi_\theta(\widetilde{y}_t \mid x, \widetilde{y}_{<t})}{\pi_{\text{old}}(\widetilde{y}_t \mid x, \widetilde{y}_{<t})}, \\
q_t^{\text{ref}}(\theta) &= \frac{\pi_{\text{ref}}(\widetilde{y}_t \mid x, \widetilde{y}_{<t})}{\pi_\theta(\widetilde{y}_t \mid x, \widetilde{y}_{<t})}, \\
k_t(\theta) &= q_t^{\text{ref}}(\theta) - \log q_t^{\text{ref}}(\theta) - 1, \\
\ell_t^{\text{pg}}(\theta; A) &= -\min(w_t(\theta)A, \text{clip}(w_t(\theta), 1-\varepsilon, 1+\varepsilon)A)\,.
\end{aligned}
\tag{3}
$$

These definitions provide the optimization basis for both training stages, with implementation conventions specified in Appendix C.3. Figure 2 shows the complete CreditOmni workflow, from multimodal inputs and cold-start initialization to the two reinforcement learning stages.

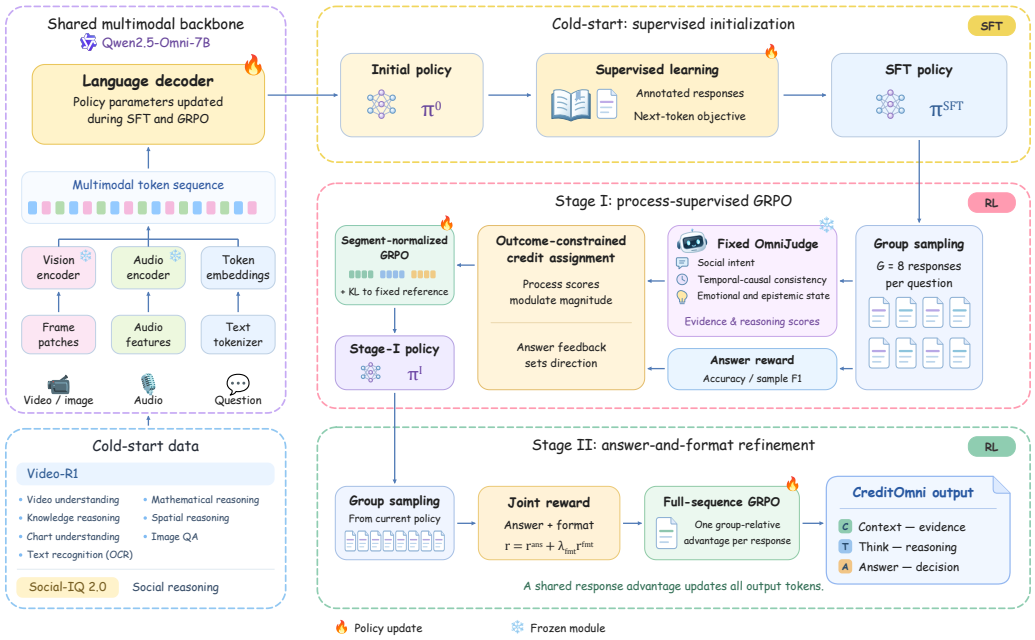


Figure 2: **Overview of CreditOmni training.** Cold-start SFT is followed by two GRPO stages. **(1) Cold-start SFT:** the model learns to generate CTA-formatted responses from annotated samples. **(2) Stage I:** eight candidates per question receive separate context and think scores across three process dimensions. Process feedback modulates their advantage magnitudes without changing the answer-advantage sign, while answer directly uses the answer advantage. Optimization is normalized by segment. **(3) Stage II:** answer and format rewards guide full-sequence optimization.

3.2 Stage I: Outcome-Constrained Credit Assignment

Stage I uses answer feedback to determine the advantage direction and process feedback to allocate credit separately to evidence and reasoning. We first introduce the motivation and design, then define the rewards, derive the segment advantages, and present the segment-normalized objective.

3.2.1 Motivation And Design

Table 1: **Illustrative example of reward mixing.** The centered signal is the mixed reward minus the group mean. Within one question group, candidate 2 is incorrect, but its mixed reward (-0.60) exceeds the group mean ($\bar{R} = -0.70$), yielding a positive centered signal ($+0.10$). Normalization by a positive denominator preserves this sign.

Candidate	Answer	Mixed reward	Centered signal
1	Correct	$+0.20$	$+0.90$
2	**Incorrect**	-0.60	$+0.10$
3	Incorrect	-1.20	-0.50
4	Incorrect	-1.20	-0.50

Table 1 illustrates how mixing answer and process rewards before group normalization can give an incorrect response a positive advantage. Using answer rewards alone also leaves candidates with equal answer scores indistinguishable in their evidence and reasoning quality. CreditOmni therefore evaluates answer correctness and process quality separately, with independent scores for context and think. Answer rewards alone determine the group-relative advantage and anchor the signs of all segment advantages; process scores adjust the context and think magnitudes within sign-preserving bounds, while answer retains its original advantage. Figure 3 illustrates this division of roles.

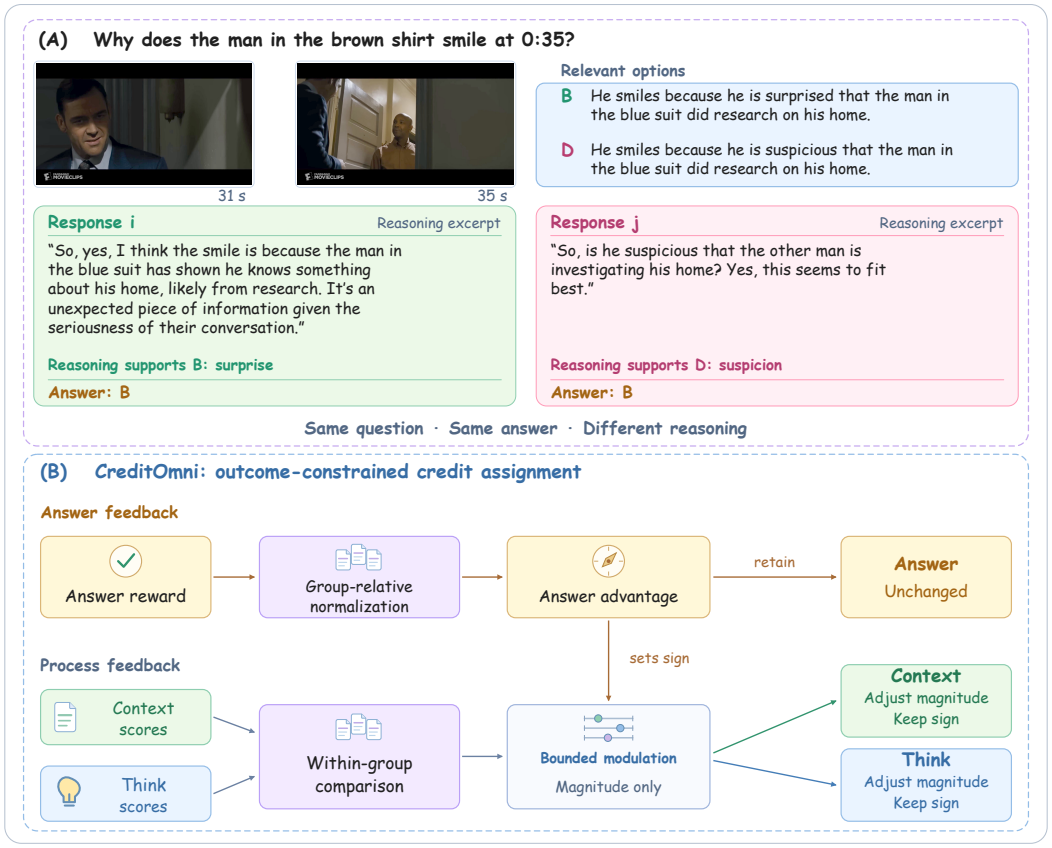


Figure 3: **Outcome-constrained credit assignment in CreditOmni.** (A) Two responses to the same question give the same correct answer but differ in reasoning. (B) Answer advantages determine update directions. Process scores modulate the magnitudes of context and think advantages within sign-preserving bounds. The answer segment retains the original answer advantage.

3.2.2 Reward Design

Format Reward. A valid CTA response contains three nonempty segments with unique opening and closing tags in the required order. The format reward is

$$r^{\text{fmt}} = \begin{cases} 1, & \text{if the CTA format is valid,} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Stage I uses format compliance as a group eligibility condition; Stage II includes this reward in its training signal.

Answer Reward. For a predicted option $\widehat{a}$ and reference answer $a^\star$, or predicted and reference label sets $\widehat{E}$ and $E^\star$, we use

$$r^{\text{ans}} = \begin{cases} 1, & \text{single-choice and } \widehat{a} = a^\star, \\ 0, & \text{single-choice and } \widehat{a} \neq a^\star, \\ F_1(\widehat{E}, E^\star), & \text{multilabel.} \end{cases} \tag{5}$$

Sample-level F1 is twice the intersection size divided by the sum of the predicted and reference label-set sizes. Both empty sets receive one, and exactly one empty set receives zero. Answer rewards provide verifiable outcome feedback in both stages.

Process Rewards. In Stage I, OmniJudge uses Qwen-Omni-Plus to transcribe the audio content of each video. GPT-5.6 Sol then evaluates context and think separately using sampled video frames and human-reviewed transcripts. We use three process rewards, corresponding to social intent, temporal-causal consistency, and emotional and epistemic state. For each response, scores on each evaluated dimension range from 0 to 1. Higher context scores indicate more accurate evidence descriptions, whereas higher think scores indicate better-supported inferences.

Social Intent Reward. Context is evaluated for accurate descriptions of relevant actions and utterances, attributed to the appropriate people. Think is evaluated for how this evidence supports a person's purpose or intent, distinguishing observable attention to an object from an inferred preference or intention.

Temporal-Causal Consistency Reward. Context is evaluated for event order and the time period specified by the question. Think is evaluated for supported temporal and causal judgments, distinguishing earlier from later states and causal relations from co-occurrence.

Emotional And Epistemic State Reward. Context is evaluated for emotional cues and the information available to each person. Think is evaluated for how these cues support the target person's emotion, belief, or knowledge, with emotional cues and access to information attributed to that person. Emotion and epistemic state are assessed as two aspects of one reward where relevant.

Appendix A illustrates these reward criteria through examples.

3.2.3 Relative Scoring And Advantage Modulation

Using the rewards defined above, we construct training advantages for the three response segments. Stage I retains groups with valid CTA segments and variation in answer rewards. This excludes groups with uniformly correct or incorrect responses under binary accuracy, or identical F1 scores. For each retained response, Equation (2) gives the answer advantage A^{ans}. Its sign distinguishes correct from incorrect responses under binary rewards. Under F1 rewards, it indicates performance relative to the group mean. Within each retained group, process scores are compared separately for context and think to obtain the relative signals used in the modulation below. Algorithm 1 summarizes the procedure; Appendix C.3 details group eligibility.

Relative Process Scores. We compare context ($S = C$) against context and think ($S = T$) against think within each question group, over the three process dimensions $d \in \mathcal{D} = \{\text{si}, \text{tc}, \text{ek}\}$ defined in Section 3.2.2. Let $\bar{s}^S_{i,d}$ be candidate i's mean evaluator score for segment S on dimension d under the repeated-scoring protocol. Pairwise comparisons assign one point for a win, half a point for a tie, and zero for a loss. Centering this ranking score yields

$$b^S_{i,d} = \frac{G}{G-1}\left[\frac{1}{G-1}\sum_{\substack{j=1\\ j\neq i}}^{G}\psi_d\big(\bar{s}^S_{i,d} - \bar{s}^S_{j,d}\big) - \frac{1}{2}\right], \tag{6}$$

where ψ_d uses a tie tolerance of 0.05 for every dimension (Appendix B.3). Each candidate pair contributes one point in total, so the uncentered group mean is $1/2$. The resulting $b^S_{i,d}$ equals the candidate's ranking score minus the mean ranking score of the other candidates. All ties yield zero.

We combine the dimensions using fixed nonnegative weights:

$$B^S_i = \frac{1}{Z}\sum_{d\in\mathcal{D}} \alpha_d b^S_{i,d}, \qquad Z = \sum_{d\in\mathcal{D}} \alpha_d > 0, \qquad S \in \{C, T\}. \tag{7}$$

Only enabled dimensions applicable to the current question are scored. Disabled or inapplicable dimensions contribute zero with no weight redistribution. Context and think share the weights and use their own scores. This weighted average preserves the bound $|B^S_i| < 1$, as required by the sign-preserving credit assignment. When no dimension is evaluated, both process scores are zero.

Outcome-Constrained Modulation. For each response, we use its relative process scores B^S to assign

$$A^S = A^{\text{ans}} + \eta_S |A^{\text{ans}}| B^S, \quad S \in \{C, T\}; \qquad A^A = A^{\text{ans}}, \tag{8}$$

where $0 \le \eta_C, \eta_T < 1$. With nonzero modulation, a higher process score strengthens a positive advantage or reduces the magnitude of a negative advantage. Context and think receive separate adjustments according to their own quality, while answer retains the original answer advantage.

Preserving The Answer-Derived Direction. For $A^{\text{ans}} \neq 0$, bounded process scores give

$$\frac{A^S}{A^{\text{ans}}} = 1 + \eta_S \operatorname{sgn}(A^{\text{ans}}) B^S \in [1 - \eta_S, 1 + \eta_S]. \tag{9}$$

The lower bound is positive, so each segment advantage retains the sign of A^{ans}. If $A^{\text{ans}} = 0$, its segment advantages are also zero. Appendix B details dimension selection, the scoring prompt, and the repeated-scoring protocol.

3.2.4 Segment-Normalized Optimization

Full-response averaging assigns a segment a policy-loss coefficient proportional to its token count. To set segment weights independently of length, we average the policy loss within each segment before combining the three means. Let $\mathcal{T}^S_i$ denote the disjoint token sets for $S \in \mathcal{S} = \{C, T, A\}$, covering the complete response. Stage I minimizes

$$\mathcal{L}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\left[\sum_{S\in\mathcal{S}}\frac{\omega_S}{|\mathcal{T}^S_i|}\sum_{t\in\mathcal{T}^S_i}\ell^{\text{pg}}_{i,t}(\theta; A^S_i) + \frac{\beta}{T_i}\sum_{t=1}^{T_i} k_{i,t}(\theta)\right]. \tag{10}$$

Here $\omega_S \ge 0$, $\sum_S \omega_S = 1$, and $\beta \ge 0$. Each segment mean receives its fixed weight ω_S, while the KL term remains averaged over the full response. The reference policy is frozen at π^{SFT}, and the batch objective averages over accepted groups. Algorithm 1 and Appendix C.3 specify policy updates and token handling.

3.3 Stage II: Answer-And-Format Refinement

Starting from π^{I}, Stage II refines answer quality and CTA format compliance with the joint reward

$$r^{\text{II}} = r^{\text{ans}} + \lambda_{\text{fmt}} r^{\text{fmt}}, \qquad \lambda_{\text{fmt}} > 0. \tag{11}$$

Both rewards follow Section 3.2.2 and are evaluated independently. Stage II also trains on responses that do not satisfy CTA format requirements.

The joint reward is normalized once within each group, giving one advantage shared by all response tokens. Optimization follows the full-response GRPO objective in Equation (1), with a frozen copy of π^{I} as the reference policy and without process scores or segment weights. Algorithm 2 gives the complete procedure, and Appendix D specifies its fixed coefficients.

4 Experiments

We use 84,000 training examples for cold-start supervised fine-tuning and 2,500 examples for subsequent GRPO training. We train on eight NVIDIA A800 graphics processing units (GPUs) using bfloat16 (BF16) precision and DeepSpeed ZeRO-2. The global SFT batch contains 32 examples, and each GRPO batch contains four complete question groups totaling 32 responses. Appendix D lists learning rates and reward parameters. All Stage I variants start from the same SFT initialization. The full Stage I model uses all three process rewards and initializes Stage II, which produces the final model used in benchmark comparisons.

We evaluate model performance on the official evaluation sets of IntentBench, Daily-Omni, and WorldSense. Table 2 compares overall scores on the three benchmarks.

Table 2: Overall scores (%) on IntentBench, Daily-Omni, and WorldSense. Citations indicate score sources; "-" denotes an unreported score. * marks audio transcripts on IntentBench. IntentBench Overall is weighted by subset size, using sample-F1 for Emotion and accuracy otherwise. CogGuide (7B+) is a multi-model framework; its reported scores use $\mathrm{CG}_{\mathrm{Qwen}}$ + HumanOmniV2 + Qwen3. CreditOmni is the final Stage II model.

Model	IntentBench	Daily-Omni	WorldSense
Closed-source models			
Claude 3.5 Sonnet (Hong et al., 2026)	-	-	34.8
GPT-4o (Yang et al., 2025; Hong et al., 2026)	59.98*	-	42.6
OpenAI o1 (think) (Yang et al., 2025)	66.69*	-	-
Gemini 1.5 Pro (Hong et al., 2026)	-	-	48.0
Gemini 2.0 Flash (Zhou et al., 2025)	-	67.84	-
Gemini 2.0 Flash Lite (Zhou et al., 2025)	-	61.32	-
Gemini 2.5 Pro (think) (Yang et al., 2025)	67.15	-	-
Open-source models			
Unified-IO-2-8B (Yang et al., 2025)	-	28.24	-
Unified-IO-2 XXL (Yang et al., 2025)	-	-	25.9
VideoLLaMA2 (Yang et al., 2025)	-	35.17	25.4
MiniCPM-o (Yang et al., 2025)	54.51	53.13	-
VITA-1.5 (Yang et al., 2025)	54.17	-	36.9
Ola (Yang et al., 2025)	57.41	49.87	-
Qwen2.5-Omni-3B (Yang et al., 2025)	-	40.52	-
Qwen2.5-Omni-7B (Yang et al., 2025)	64.20	47.45	45.4
HumanOmniV2 (Yang et al., 2025)	69.33	58.47	47.1
EchoInk-R1 (Kulkarni & Fazli, 2026)	63.6	46.2	45.7
Omni-R1 (Kulkarni & Fazli, 2026)	63.5	46.8	44.1
AV-Reasoner (Kulkarni & Fazli, 2026)	59.5	53.8	44.6
AVATAR (Kulkarni & Fazli, 2026)	63.9	47.0	46.0
Qwen3-Omni-30B-A3B (think) (Chen et al., 2026)	68.5	75.8	48.0
Qwen3-Omni-30B-A3B (instruct) (Chen et al., 2026)	65.7	63.6	54.0
CogGuide (7B+) (Shou et al., 2025)	71.18	62.49	48.36
CreditOmni-7B	**71.72**	**60.82**	**48.64**

Table 2 shows that CreditOmni leads the evaluated open-source 7B models on all three benchmarks. Compared with Qwen2.5-Omni, it improves scores by **7.52**, **13.37**, and **3.24** percentage points on IntentBench, Daily-Omni, and WorldSense, respectively.

4.1 Ablation Studies

Table 3: **Training comparisons and Stage I ablations across three benchmarks.** Scores are percentages; parenthesized changes are percentage points relative to CreditOmni-I (Full). Stage I variants share the same SFT initialization. Reward ablations score and check only remaining enabled, applicable dimensions, with unchanged reliability criteria. Removed dimensions contribute zero; the aggregation denominator is fixed.

Configuration	IntentBench	Daily-Omni	WorldSense
A. Training comparisons			
SFT-only	64.95	49.54	44.86
Vanilla GRPO (answer-only)	66.74	53.63	46.03
CreditOmni-I (Full)	71.19	60.40	48.42
CreditOmni-II (Final)	**71.72**	**60.82**	**48.64**
B. Stage I: process-reward ablations			
w/o Social intent	70.51 (−0.68)	57.89 (−2.51)	47.45 (−0.97)
w/o Temporal-causal consistency	69.63 (−1.56)	58.40 (−2.00)	47.70 (−0.72)
w/o Emotional and epistemic state	70.00 (−1.19)	58.98 (−1.42)	47.86 (−0.56)
C. Stage I: credit assignment			
Fixed-coefficient fusion	70.70 (−0.49)	59.06 (−1.34)	47.95 (−0.47)
Shared-advantage assignment	68.70 (−2.49)	57.23 (−3.17)	47.13 (−1.29)

Table 3 shows gains from full Stage I over answer-only GRPO and further gains from Stage II. Removing any process reward lowers performance. Fixed-coefficient fusion replaces $\eta_S|A^{\text{ans}}|$ with $\lambda_C = \lambda_T = 0.25$, leaving the answer segment unchanged. Shared-advantage assignment applies $\sum_{S\in\mathcal{S}} \omega_S A^S$ to all segments, preserving their signs. Process scoring and loss aggregation remain unchanged, with no additional centering. Both variants underperform full Stage I.

The experiments highlight the importance of both the information conveyed by process feedback and its allocation across response segments. Reward ablations assess the contributions of individual process-quality dimensions, whereas credit-assignment variants examine how these signals are used during training. These findings support the joint design of process rewards and segment-level credit assignment for multimodal supervision.

5 Early Explorations And Design Insights

Developing CreditOmni involved extensive experimentation with reinforcement learning for large multimodal models. Attempts that fell short of expected gains informed our understanding of model optimization, reasoning supervision, and training practice. Appendix E examines two representative questions through experimental observations and methodological analyses: effective supervision for multimodal reasoning and feedback quality versus resource costs. These efforts offer an interesting glimpse into the training and optimization of large multimodal models, providing design insights for further investigation and refinement.

6 Conclusion

We present CreditOmni, a multimodal post-training framework that incorporates evidence and reasoning quality into model learning. Evaluations on IntentBench, Daily-Omni, and WorldSense show consistent gains over the base model and the highest overall scores among the evaluated open-source 7B models. These findings demonstrate the value of using the response process as a source of training supervision. CreditOmni offers a practical approach to learning from how multimodal evidence supports a conclusion, extending supervision beyond the correctness of the final answer.

AI Use Statement

In this study, OmniJudge uses Qwen-Omni-Plus for audio transcription and GPT-5.6 Sol for process scoring during Stage I training. Human reviewers check the transcripts and correct errors before scoring. GPT-5.6 Sol separately evaluates candidate evidence descriptions and reasoning using sampled video frames, reviewed transcripts, and the question.

References

Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, Biao Wang, Qinglin Lu, and Ruqi Huang. OmniVideo-R1: Reinforcing audio-visual reasoning with query intention and modality attention. *arXiv preprint arXiv:2602.05847v2*, 2026. URL https://arxiv.org/abs/2602.05847v2.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. DramaQA: Character-centered video story understanding with hierarchical qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1166–1174, 2021. doi: 10.1609/aaai.v35i2.16203.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3310. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8eb3976840e08b80dda9667562574246-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

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*, 2026.

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*, 2026. URL https://arxiv.org/abs/2508.03100.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a. URL https://arxiv.org/abs/2408.03326.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19108–19118, 2022.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b. URL https://arxiv.org/abs/2311.17005.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Ruilin Luo, Zhuofan Zheng, Lei Wang, Yifan Wang, Xinzhe Ni, Zicheng Lin, Songtao Jiang, Yiyao Yu, Chufan Shi, Ruihang Chu, Jin Zeng, and Yujiu Yang. Unlocking multimodal mathematical reasoning via process reward model. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1664.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. *arXiv preprint arXiv:1910.02054*, 2019. URL https://arxiv.org/abs/1910.02054.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Zhou-Peng Shou, Zhi-Qiang You, Fang Wang, and Hai-Bo Liu. CogGuide: Human-like guidance for zero-shot omni-modal reasoning. *arXiv preprint arXiv:2509.06641*, 2025. URL https://arxiv.org/abs/2509.06641.

Paul Vicol, Makarand Tapaswi, Lluís Castrejón, and Sanja Fidler. MovieGraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8581–8590, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Vicol_MovieGraphs_Towards_Understanding_CVPR_2018_paper.html.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024. doi: 10.18653/v1/2024.acl-long.510.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. VisualPRM400K: An effective dataset for training multimodal process reward models. In *International Conference on Learning Representations*, 2026.

Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-IQ 2.0 Challenge: Benchmarking multimodal social understanding. GitHub repository, 2023. URL https://github.com/abwilf/Social-IQ-2.0-Challenge.

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*, 2025. doi: 10.48550/arXiv.2503.20215. URL https://arxiv.org/abs/2503.20215.

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*, 2025. URL https://arxiv.org/abs/2506.21277.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-Omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025. URL https://arxiv.org/abs/2505.17862v1.

A Questions And Comparisons Of Model Outputs

Vanilla GRPO

(a) Qualitative Case: Emotional and Epistemic State



00:43 00:48 00:50

Question. How does the woman interact with the man in the black sweater?

Options. A. She is friendly with him. B. The woman's tone is dismissive towards him. C. They have a confrontational interaction with each other. D. She is calm with him.

Gold: D. Model answer: B (incorrect).

(b) Response.

<context>

In the video, the woman displays rapid movements which indicate dissatisfaction or anxiety, with body language that suggests urgency. The man in the black sweater is calm with a steady expression, showing neither tension nor anxiety. The setting is an indoor office environment, conveying a professional relationship. Although both parties are dressed casually, the woman's anxious body language and the man's relaxed demeanor create a subtle conflict. In the audio, the woman's tone is hurried and tense, with an upward intonation. The man in the black sweater responds in a calm and gentle tone, with a slowing pace to soothe her. The conversation reveals that the woman is anxious about being overlooked in the work pitch, mentioning, "Kevin forgot I was even working on this pitch," and she feels the pressure to "get this thing approved." The man's response is "I'll keep my eyes out for you," with no comforting nature in his voice.

</context>

<think>

Okay, let me figure out how the woman is interacting with the man in the black sweater. First, I'm looking at the video. Her body language – all those rapid movements – definitely suggests she's not calm or friendly. She looks quite anxious, maybe even angry. She's not really making eye contact, and her gestures seem more like she's emphasizing something to her boss rather than truly communicating. The man, in contrast, looks very calm and steady. The office setting seems pretty standard, just a typical work interaction. Now, let me listen to the audio. Her voice is the key here. It's hurried, and she sounds tense, with an upward intonation. That usually means someone isn't pleased or isn't being friendly. The man, on the other hand, speaks calmly and gently, with a slower pace. It seems he's trying to soothe her. I can also hear what they're saying. She's definitely upset about her work pitch, saying something like, "Kevin forgot I was even working on this pitch, and now I'm doing all the work." She's not really asking him for advice, just expressing her frustration. His response, "I'll keep my eyes out for you," sounds more like he's being dismissive, not like he's engaging on an equal footing. So, putting it all together – her quick movements and tense voice, his calm, somewhat dismissive reaction – it really seems like there's a power dynamic at play or maybe she's upset over the work situation. Option B, "dismissive," feels right for him, while she's clearly not being calm or friendly herself. I don't think it's a "confrontation" either because there's no real opposition in his calm demeanor. Her anxiety is more directed at the situation itself.

</think>

<answer> B **</answer>**

Figure 4: **Emotion attribution error case.** (a) Video frames and question. (b) Vanilla GRPO response. The question and option B concern the woman's attitude, but the inference highlighted in red attributes dismissiveness to the man and uses it to select B. Emotional and Epistemic State Reward checks whether emotional or attitudinal evidence is assigned to the correct participant.

Ours (CreditOmni)

(a) Qualitative Case: Emotional and Epistemic State



Question. How does the woman interact with the man in the black sweater?

Options. A. She is friendly with him. B. The woman's tone is dismissive towards him. C. They have a confrontational interaction with each other. D. She is calm with him.

Gold: D. **Answer:** D.

(b) CreditOmni response.

<context>

The video shows a woman sitting on a sofa with a laptop in front of her. When a man in a black sweater approaches, she looks up at him. He then sits beside her, and they talk face to face. She maintains a steady seated posture and a fairly serious expression, looking at him while he speaks. In the audio, the man first asks, "Hey, can we talk?" The woman briefly replies, "Sure." He then apologizes while she looks at him and continues listening.

</context>

<think>

I'm watching the video, focusing on the two people specified in the question. The question asks how the woman interacts with the man in the black sweater, so I need to focus on their conversation beside the sofa in the latter part of the clip. I should not use her earlier conversation with another man, or her expressions of work-related frustration in other shots, to characterize this interaction. In this scene, the woman is initially using her laptop on the sofa. When the man in the black sweater approaches and asks whether they can talk, she looks up at him and replies, "Sure." He then sits down and apologizes, while she remains seated and looks at him as she listens. These observations indicate that she accepts the conversation. Her initial use of the laptop does not, by itself, mean that she is unwilling to engage with him. If I consider "friendly," agreeing to talk does not necessarily mean that she is being warm or affectionate. Her expression remains fairly serious, and the available visual evidence does not clearly show an enthusiastic interaction. So "calm" describes her present behavior more closely than "friendly." "Confrontational" is not the best description either. His apology indicates that the conversation concerns something that happened earlier, but the apology alone does not establish that she is confronting him now. I need to look at her current response: she agrees to talk and continues listening while he speaks, without overtly confrontational gestures. As for "dismissive," a serious expression or a brief reply is not sufficient evidence. She looks up at him, responds to his request, and continues participating in the conversation. These concrete actions provide a firmer basis for interpretation than treating a single expression as contempt. In particular, I should not transfer her earlier dissatisfaction with her work situation into evidence that she is dismissive toward this man. So the final judgment should concern the woman's own behavior, not whether the man is gentle, apologetic, or trying to reassure her. Considering that she accepts the conversation, maintains a steady seated posture, and listens to him, "She is calm with him" best fits the interaction among the four options. Therefore, I choose D.

</think>

<answer> D **</answer>**

Figure 5: **CreditOmni analysis for the question in Figure 4.** (a) Video frames and question. (b) CreditOmni response. The inference highlighted in green grounds the woman's attitude in the target interaction, without using her earlier frustration or the man's behavior as evidence of her attitude toward him. The comparison highlights correct participant attribution.

Vanilla GRPO

(a) Qualitative Case: Social Intent



00:02.47 00:07.47 00:12.48

Question. Why do the men not look at each other while talking in the car?

Options. A. The driver must focus on another more important task. B. They are excited to get the American candy. C. Yes, they are friendly and fun with each other. D. He spins the map a lot of times.

Gold: A. Model answer: D (incorrect).

(b) Response.

<context>

The visual content shows two individuals inside a vehicle, and they appear to be talking to each other. There are specific moments where they stop talking and look at something else, focusing their attention on objects within the vehicle rather than interacting with each other directly.

</context>

<think>

I notice the two individuals in the car. Their conversation is punctuated by moments of looking away, directing their attention internally to the vehicle rather than engaging with each other. I consider the options. Option A suggests the driver is preoccupied, but focusing solely on driving wouldn't explain the driver talking and the passenger not looking. Option B, which mentions excitement about American candy, doesn't connect with the visual lack of eye contact between them. Option C is too generalized about them being friendly, without explaining their behavior. The behavior aligns with option D, which mentions spinning a map, as people often gaze elsewhere when focused on a task like this, showing they're engaged with something external, not each other. It seems they're more interested in the map than in conversing with each other.

</think>

<answer> D **</answer>**

Figure 6: **Social intent error case.** (a) Video frames and question. (b) Vanilla GRPO response. The incorrect candidate interprets the lack of eye contact as both men preferring the map over conversation and selects D. The inference highlighted in red turns the direction of attention into a shared preference, overlooking the driving task. Social Intent Reward checks whether evidence supports the attributed preference or intent.

Ours (CreditOmni)

(a) Qualitative Case: Social Intent and Task Priorities



00:02 00:07 00:12

Question. Why do the men not look at each other while talking in the car?

Options. A. The driver must focus on another more important task. B. They are excited to get the American candy. C. Yes, they are friendly and fun with each other. D. He spins the map a lot of times.

Gold: A. **Answer: A.**

(b) CreditOmni response.

<context>

The video shows two men inside a car, engaged in conversation. One man is at the steering wheel, while the passenger unfolds and turns a large map. The driver sometimes glances toward the passenger or the map, but during the driving conversation he generally looks ahead. The passenger alternates between looking at the map and looking toward the driver. They continue talking without maintaining constant eye contact. The dialogue concerns their journey and the distance still to travel.

</context>

<think>

As I observe the scene, I see the two men inside a car, suggesting they are on a journey, possibly a trip along a road. The map and their conversation about the remaining distance support this interpretation. Their roles are different: one man is driving, while the other is a passenger handling the map. The driver's body language shows him directing much of his attention toward the road ahead during the driving conversation, a common behavior when someone is driving. This does not mean that both men look forward throughout the video. The driver does glance sideways, and the passenger looks down at the map and turns toward him. What matters is that the driver cannot maintain sustained eye contact with the passenger while attending to the road. This gaze toward the road rather than continuously at the passenger suggests that the driver is prioritizing driving for safety. It does not mean that the conversation itself lacks direct interaction. They are still responding to each other, and their shared journey gives them something to discuss even when their gaze is directed elsewhere. Thus, the reason they are not continuously looking at each other while talking is that the driver must attend to the road, with his attention divided between conversing and driving. The passenger's map handling is part of the scene, but simply describing the map being turned does not explain the driver's task priority. Their friendliness and the later interest in candy also do not provide that explanation. This conclusion aligns with option A: the driver must focus on another, more important task.

</think>

<answer> A </answer>

Figure 7: **CreditOmni analysis for the question in Figure 6.** (a) Video frames and question. (b) CreditOmni response. The inference highlighted in green connects reduced eye contact to the driver's task priority, rather than a preference for the map or a lack of conversational engagement.

Vanilla GRPO

(a) Qualitative Case: Temporal-Causal Reasoning



00:06 00:09 00:24

Question. How did the man feel when the woman hesitated to say yes?

Options. A. He was disappointed and embarrassed. B. The man was calm and collected. C. He felt excited and happy. D. He felt scared and nervous.

Gold: D. **Model answer: C (incorrect).**

(b) Response.

<context>

In the video, the man is wearing a suit and appears tense. When the woman hesitates to respond, he immediately shows a look of surprise with his eyes wide open. Then he closes his eyes and slowly shakes his head, with his gaze lowered (eyes downcast), displaying a complex expression. His body posture is forward-leaning, indicating an intention to move closer. The woman looks happy and relaxed, in stark contrast to the man's internal state. The background is an outdoor night scene, and the lighting focuses on the man, highlighting his emotional fluctuations. In the conversation, before the woman hesitation, the man's tone is urgent and high-pitched, conveying tension and anticipation. After asking a question, the woman's response is brief ("Yes"), her tone noticeably lower. The subsequent audio clearly captures the man's excited and happy exclamations "Yes! I knew it!" followed by laughter and cheers from the surrounding crowd, indicating an intense positive emotional outburst.

</context>

<think>

Oka… I'm trying to figure out the man's emotions when the woman hesitates. Watching the video, his tense appearance and the way he leans forward right after she doesn't immediately reply – that surprised, almost shocked look with his eyes wide, then closing them and lowering his gaze, really stands out. It feels like he's feeling let down, and perhaps a bit embarrassed by her hesitation, but he's definitely not just shocked. Then I listen to the audio, especially his voice before and right after she says "Yes." His tone is very high-pitched and urgent, which tells me he's highly strung, waiting. After she finally responds, I hear him say "Yes! I knew it!" really loudly and happily, and then there's cheers from the crowd. This sounds like an emotional outburst, more excited and happy than just simple relief from the pressure. So, considering all this: If the options were 'Embarrassed and Disappointed,' he definitely shows those, especially with the headshake and lowered gaze. But "only" those two might not cover his full emotional range. 'Breathless and Anticipation' – he seems breathless in the audio, sure, especially after that outburst, but anticipation felt more like the pre-hesitation tension, not the immediate reaction to her pause. 'Excited and Happy' – that really fits. I hear laughter and cheering, and his excited exclamations contradict the negative aspects I saw, like disappointment. His body language leans forward, eager, like he wants to express his emotions. 'Scared and Nervous' – I don't really see or hear fear or nervousness. The 'nervous' part could apply to his initial state, but his reaction to the 'hesitation' itself is more about disappointment and then elation. It's a complex moment. He's disappointed by her hesitation, maybe a bit embarrassed, but the core emotion that dominates his subsequent behavior and the situation is excitement and joy, especially after the tension. So, he's definitely excited and happy.

</think>

<answer> C **</answer>**

Figure 8: **Temporal-causal mismatch error case.** (a) Video frames and question. (b) Vanilla GRPO response. The question concerns his feelings while the woman hesitates, before accepting. The candidate distinguishes tension before her response from cheering afterward, but the inference highlighted in red uses subsequent excitement to override the earlier state and select C. Temporal-Causal Consistency Reward checks whether the evidence and explanation match the requested time window.

Ours (CreditOmni)

(a) Qualitative Case: Temporal Order and Emotional State



00:06 00:09 00:24

Question. How did the man feel when the woman hesitated to say yes?

Options. A. He was disappointed and embarrassed. B. The man was calm and collected. C. He felt excited and happy. D. He felt scared and nervous.

Gold: D. **Answer:** D.

(b) CreditOmni response.

<context>

The video shows a man proposing to a woman seated on a bench. He holds an open ring box toward her and watches her while waiting for a response. The woman smiles and looks down before answering. The surrounding situation is filled with anticipation, with the woman hesitating in her response. The man's attention remains on her, indicating that he is emotionally invested in the situation. The spoken proposal is followed by a pause before the woman accepts. The later shots show the couple celebrating and embracing. Those later reactions occur after her acceptance, rather than during the hesitation specified in the question.

</context>

<think>

As I observe the video, I see the man holding out the ring box and waiting for the woman's answer. His attention remains fixed on her response, and the pause leaves the outcome of the proposal unresolved. This suggests to me that he is not just momentarily excited; his emotional state includes the nervous anticipation of an answer that matters greatly to him. The woman's hesitation adds to the tension in the scene. Given the overall situation, it is reasonable to interpret him as anxious about whether she will accept. But hesitation is not the same as rejection. I should not conclude that he is disappointed and embarrassed simply because his expectations are not immediately met. His emotional investment and the uncertainty of the pause support nervousness more directly than a definite sense of being let down. The timing matters here. The video does later show excitement and happiness, but those reactions follow her acceptance. They cannot replace the judgment about how he feels while she is still hesitating. Likewise, holding the ring box and waiting does not by itself make him calm and collected; that posture needs to be interpreted in the context of the unresolved proposal. Therefore, the best-fitting description at this specific point in time is that he feels scared and nervous. Here, "scared" is most reasonably understood as apprehension about a possible rejection, not fear of physical danger. The answer concerns his uncertainty while waiting, rather than disappointment inferred from the delay or happiness observed afterward. I choose D.

</think>

<answer> D **</answer>**

Figure 9: **CreditOmni analysis for the question in Figure 8.** (a) Frames before and after acceptance, with the question. (b) CreditOmni response. The inference highlighted in green distinguishes nervous anticipation during hesitation from later celebration, keeping the judgment aligned with the queried time.

B OmniJudge Scoring Protocol And Human Review

B.1 API Scoring And Dimension Selection

OmniJudge uses Qwen-Omni-Plus to transcribe the audio content of each video once. Human reviewers check the transcripts and correct errors before scoring. The reviewed transcripts are cached for subsequent scoring calls. GPT-5.6 Sol then evaluates context and think separately using sampled video frames and the reviewed transcripts. The evaluation covers the three process dimensions defined in Section 3.2.2.

The enabled dimensions $\mathcal{D}_{\text{on}}$ are fixed before training. The question and media determine the applicable dimensions $\mathcal{D}_g^{\text{app}}$ before sampling. Both segments are evaluated on $\mathcal{D}_g^{\text{req}} = \mathcal{D}_{\text{on}} \cap \mathcal{D}_g^{\text{app}}$. Disabled or inapplicable dimensions contribute zero under the fixed aggregation denominator. If no dimension is evaluated, both process aggregates are zero and the GPT-5.6 Sol scoring calls are skipped.

B.2 Two-Pass Group Scoring

In each call, GPT-5.6 Sol evaluates all eight candidates using sampled video frames, human-reviewed transcripts, the question, and options. Each evaluated dimension yields separate scores and confidences for `context` and `think`: evidence accuracy for the former and evidential support for the latter. The input presents the frames and transcripts before the candidate text. It excludes gold answers, candidate `answer` fields, answer rewards, and SFT targets. We make two scoring calls, randomly shuffling the candidate order for each call. Both calls reuse identical sampled frames and cached, reviewed transcripts with the same fixed GPT-5.6 Sol evaluator.

After matching candidates across calls, we compare the two scores for each candidate, segment, and evaluated dimension. A group is retained only if every absolute score difference is at most 0.20 and every judgment reports sufficient evidence and confidence of at least 0.80 in both calls. Invalid or incomplete evaluations exclude the group. For each retained scoring item, we average the two scores before computing the relative process score in Equation (6).

B.3 Pairwise Score Construction

After restoring candidate order, let $s_{i,d}^{S,\ell} \in [0,1]$ denote the score for candidate i's segment S on dimension d in evaluator call $\ell \in \{1,2\}$. For retained groups, the score used in the main text is $\bar{s}_{i,d}^{S} = (s_{i,d}^{S,1} + s_{i,d}^{S,2})/2$. We fix the tie tolerance at $\xi_d = 0.05$ for every dimension before training. Define

$$\psi_d(z) = \begin{cases} 1, & z > \xi_d, \\ \frac{1}{2}, & |z| \le \xi_d, \\ 0, & z < -\xi_d, \end{cases} \qquad \rho_{i,d}^{S} = \frac{1}{G-1} \sum_{\substack{j=1 \\ j \neq i}}^{G} \psi_d\big(\bar{s}_{i,d}^{S} - \bar{s}_{j,d}^{S}\big)\,. \tag{12}$$

Each pair satisfies $\psi_d(z) + \psi_d(-z) = 1$, yielding a group mean ranking score of $1/2$. Centering gives $b_{i,d}^{S} = G(\rho_{i,d}^{S} - 1/2)/(G-1)$, the expanded form used in Equation (6).

B.4 Human Review Of Transcripts

Human reviewers compare Qwen-Omni-Plus transcripts with the original audiovisual input and correct errors before scoring. The reviewed transcripts are then supplied to GPT-5.6 Sol alongside sampled video frames. The scoring prompt follows the criteria in Section 3.2.2 and remains fixed throughout process scoring during training.

B.5 Scoring Prompt

The scoring prompt below follows the process-reward criteria in Section 3.2.2.

System prompt

You are a multimodal reasoning evaluator. Using the provided video frames, audio transcript, and question, score each candidate's Context and Think separately on the applicable dimensions below.
1. Social intent: Are the actions, utterances, and person attributions in Context accurate? Is the purpose or intent inferred in Think supported by evidence?
2. Temporal-causal consistency: Are the event order and time period in Context accurate? Are the temporal or causal inferences in Think supported?
3. Emotional and epistemic state: Are the emotional cues and information available to each person in Context accurate? Are the emotion, belief, or knowledge judgments in Think supported?
Assign a score using the five ranges below; decimal values within each range are allowed.

Score	Criterion
0.8–1.0	Accurate key descriptions or inferences, sufficient support, and no substantive errors.
0.6–0.8	Mostly supported content, with local omissions or minor deviations.
0.4–0.6	Partly supported content, with clear omissions or errors that affect the judgment.
0.2–0.4	Only a small portion is supported; key descriptions or inferences contain major errors.
0–0.2	Little to no support or fundamental errors; 0 denotes entirely irrelevant or wholly unsupported content.

Give one brief reason for each score, identifying the key evidence or main error. Also return judgment confidence in [0,1] and whether the input permits a reliable evaluation (true or false). If the input is insufficient for reliable evaluation, return false and explain why. A clearly incorrect candidate may receive a low quality score, high confidence, and true.

User prompt

Sampled video frames: `{video_frames}`
Audio transcript: `{transcript}`
Question and options: `{question_and_options}`
Requested dimensions: `{dimensions}`
Candidates (ID, Context, Think): `{candidates}`
Return only JSON, organized by candidate ID, then Context/Think, then reward dimension. Each item contains `score`, `reason`, `confidence`, and `sufficient_evidence`.

C Segment Normalization And Two-Stage Update Algorithms

C.1 Stage I: Process-Modulated GRPO

Algorithm 1 presents the complete Stage I training procedure. Starting from the cold-start SFT policy π^{SFT}, it samples and filters candidates, computes answer advantages, modulates segment-level advantages, and updates the policy. Each training batch comprises B accepted candidate groups, with K_{update} parameter updates performed on the same candidates.

Algorithm 1 CreditOmni Stage I: Outcome-Constrained Segmental Optimization

Require: Initial policy π^{SFT}, training set $\mathcal{X}$, enabled dimensions $\mathcal{D}_{\text{on}}$; group size $G = 8$, batch size B, update counts N_{I} and K_{update}

Initialize $\pi_\theta \leftarrow \pi^{\text{SFT}}$ and freeze reference $\pi_{\text{ref}} \leftarrow \pi^{\text{SFT}}$
for $n = 1, \ldots, N_{\text{I}}$ **do**
  Fix $\pi_{\text{old}} \leftarrow \pi_\theta$; initialize batch $\mathcal{B} \leftarrow \emptyset$
  while $|\mathcal{B}| < B$ **do**
    Sample a question and G responses from π_{old}
    Parse CTA segments and answers; construct segment masks; compute r_i^{ans}
    if CTA/answer parsing or segment masks are invalid, or all answer rewards are equal **then**
      Reject the group and continue
    Compute A_i^{ans} using Equation (2)
    Select enabled, applicable dimensions $\mathcal{D}^{\text{req}}$; initialize all $b_{i,d}^S \leftarrow 0$
    if $\mathcal{D}^{\text{req}} \neq \emptyset$ **then**
      Score context and think twice with OmniJudge on $\mathcal{D}^{\text{req}}$
      if the group fails the reliability checks in Appendix B **then**
        Reject the group and continue
      Average score pairs; compute $b_{i,d}^S$ using Equation (6)
    Aggregate B_i^S and assign (A_i^C, A_i^T, A_i^A) using Equations (7)–(8)
    Add the group to $\mathcal{B}$
  for $k = 1, \ldots, K_{\text{update}}$ **do**
    Update π_θ using the batch mean of Equation (10)
return $\pi^{\text{I}} \leftarrow \pi_\theta$

C.2 Stage II: Answer-And-Format GRPO

Stage II initializes from π^{I} and freezes a copy as its reference policy. Before sampling each batch, it copies the current policy to π_{old} and keeps the old policy fixed throughout the batch's parameter updates. It uses the token policy loss and KL regularization surrogate in Equation (3), with reference $\pi_{\text{ref}}^{\text{II}}$ and clipping threshold $\varepsilon_{\text{clip}}^{\text{II}}$. Sampling and probability evaluation follow the same conventions as in Stage I. All output tokens share the advantage computed from the joint answer and format reward; token losses are averaged over the complete response:

$$\mathcal{L}_g^{\text{II}}(\theta) = \frac{1}{G} \sum_{i=1}^{G} \frac{1}{T_i} \sum_{t=1}^{T_i} \left[\ell_{i,t}^{\text{pg}}(\theta; A_i^{\text{II}}) + \beta_{\text{II}} k_{i,t}(\theta) \right]. \tag{13}$$

Here T_i counts returned tokens, including an EOS token when present; $0 < \varepsilon_{\text{clip}}^{\text{II}} < 1$ and $\beta_{\text{II}} \geq 0$. The batch objective averages these losses equally over B accepted candidate groups. Stage II does not require format compliance and does not use evaluator-based filtering. Malformed responses, unparseable answers, and outputs truncated before closing tags all contribute their actual returned tokens. If any candidate in a group returns no output tokens usable for loss computation, or all candidates receive the same answer reward, we discard the group and continue sampling. Format scores do not affect this reward-based filter.

Algorithm 2 CreditOmni Stage II: Joint Answer And Format Optimization

Require: Stage I policy π^{I}, training set $\mathcal{X}$; group size $G = 8$, batch size B, update counts N_{II} and K_{update}

Initialize $\pi_\theta \leftarrow \pi^{\mathrm{I}}$ and freeze reference $\pi^{\mathrm{II}}_{\text{ref}} \leftarrow \pi^{\mathrm{I}}$
for $n = 1, \ldots, N_{\mathrm{II}}$ **do**
  Fix $\pi_{\text{old}} \leftarrow \pi_\theta$; initialize batch $\mathcal{B} \leftarrow \emptyset$
  while $|\mathcal{B}| < B$ **do**
    Sample a question and G responses from π_{old}
    if any response has no output tokens usable for loss computation **then**
      Reject the group and continue
    Compute answer and format rewards independently; combine them using Equation (11)
    if all answer rewards are equal **then**
      Reject the group and continue
    Normalize joint rewards as in Equation (2), with $\varepsilon_{\mathrm{II}}$, to obtain A^{II}_i
    Add the group to $\mathcal{B}$
  for $k = 1, \ldots, K_{\text{update}}$ **do**
    Update π_θ using the batch mean of Equation (13)
return $\pi^{\mathrm{II}} \leftarrow \pi_\theta$

C.3 Eligibility And Token-Level Implementation

Group Eligibility. All candidates for a question share the reference answers $z^\star$. Answer and format rewards follow Section 3.2.2. Both stages retain groups only when answer rewards vary, whether measured by binary accuracy or F1. Stage I also requires valid CTA formatting, parseable answers, nonempty segments, and valid token masks with aligned offsets. Stage II requires usable output tokens and evaluates answer and format rewards independently: a failed check sets its corresponding reward to zero before the joint reward is normalized within a retained group. Algorithms 1 and 2 summarize the two stages.

Token Boundaries And Masks. We assign the T_i returned non-padding tokens to disjoint segment sets $\mathcal{T}_i^S$, where $S \in \{C, T, A\}$, without retokenization. These sets cover all returned tokens, including tags, whitespace, and an end-of-sequence (EOS) token when present; EOS belongs to `answer`. Segment boundaries lie at the first bytes of the opening `think` and `answer` tags. Each token is assigned according to its final decoded byte, and the masks $M^S_{i,t} = \mathbf{1}[t \in \mathcal{T}_i^S]$ record these assignments. Stage I rejects a group if any segment token set is invalid or empty, as specified in Algorithm 1. Stage II averages over the complete returned response.

Batch Updates. Each batch contains B accepted groups collected within N_{max} sampling attempts. We perform K_{update} parameter updates on this batch while keeping the old policy, rewards, process scores, masks, and stored old-policy probabilities fixed. The scheduler advances after every successful parameter update, and a completed batch counts as one effective update. Training stops if an optimization step fails or the sampling limit is reached before B groups are collected; in the latter case, no stage policy is returned. Appendix D lists the training hyperparameters.

KL Sampling. We compute the KL regularization term by averaging the surrogate in Equation (3) over retained samples from the old policy.

D Training Configuration

We train Qwen2.5-Omni-7B on eight NVIDIA A800 GPUs with BF16 precision and DeepSpeed ZeRO-2 (Rajbhandari et al., 2019). The visual and audio encoders and modality projectors remain frozen. All stages use AdamW with a cosine learning-rate schedule. Table 4 summarizes the stage-specific settings. SFT runs for two epochs. Each GRPO batch contains four question groups with eight responses per group, and we perform two parameter updates per batch. Stage I uses equally weighted process dimensions.

Table 4: Training configuration by stage. Segment weights follow the order context, think, and answer. "–" denotes a setting that does not apply to the stage.

Setting	SFT	Stage I	Stage II
Peak learning rate	2×10^{-5}	10^{-6}	5×10^{-7}
Global batch size	32	4×8 responses	4×8 responses
Updates per GRPO batch	–	2	2
Maximum generation length	2,048 tokens	2,048 tokens	2,048 tokens
Clipping threshold	–	0.2	0.2
KL coefficient	–	0.02	0.02
Segment weights	–	$(0.2, 0.3, 0.5)$	–
Pairwise tie tolerance	–	0.05	–
Format-reward coefficient	–	–	0.1
Group reward filter	–	Answer-reward variation	Answer-reward variation

E Early Explorations And Design Insights

E.1 Outcome Supervision And Reasoning Processes

Our early explorations also prompted us to consider how feedback on final outcomes can support the learning of reasoning capabilities. Certain tasks in mathematics and programming offer a useful perspective: when later steps depend on intermediate results and final outcomes can be reliably verified, the quality of the reasoning process affects the correctness of the final answer. Outcome rewards can therefore provide indirect supervision for the process that produces an answer.

Consider a reading problem from the GSM8K training set (Cobbe et al., 2021). A reader has a 120-page book and read 12 pages yesterday. Today, they read twice that amount and plan to read half of the remaining pages tomorrow. The question asks how many pages they will read tomorrow. The reference solution proceeds as follows.[1]

Table 5: Reference solution to a reading problem from the GSM8K training set.

Step	Calculation	Result
1	Pages read today: 12×2	24 pages
2	Total pages read: $12 + 24$	36 pages
3	Pages remaining: $120 - 36$	84 pages
4	Pages to read tomorrow: $84 \div 2$	**42 pages**

Along this solution path, intermediate results feed into subsequent calculations and jointly determine the final answer. When process quality is consistently associated with outcome correctness, group-relative outcome rewards may favor effective reasoning paths. The strength of this association may thus influence the effectiveness of outcome supervision and training efficiency.

Turning to multimodal reasoning, even when final answers can be verified, the same answer may arise from information-processing paths of differing quality. One response may accurately integrate relevant audiovisual cues, whereas another may reach the same answer despite missing information or insufficiently supported reasoning. A model may also exploit textual shortcuts, such as question wording, answer-option patterns, or learned language associations, without adequately using the relevant audiovisual evidence. These paths can receive identical answer rewards. Final outcomes alone therefore provide limited information about whether the model has made full use of the relevant evidence and whether its inferences were supported by that evidence.

This analysis motivates us to consider not only whether outcomes are verifiable, but also how effectively outcome feedback constrains the reasoning process. The resulting design insight is that reinforcement learning for large multimodal models should consider outcome correctness and process quality separately. Organizing process feedback around the capabilities required by the task can provide more targeted supervision for cross-modal information integration and reasoning.

[1] GSM8K training data.

E.2 Reliability And Practical Trade-Offs In Model-Assisted Supervision

In our early explorations, we used ChatGPT 5.6 to construct multiple versions of training annotations with different structures and used them to train large multimodal models. The two most representative structures were E/P/S process graphs and modality-separated process annotations. The former organize entities, perceptual facts, and inferences into interconnected nodes, whereas the latter organize audio and visual evidence separately. Figures 10 and 11 illustrate these two structures using the same actual training sample.

In our E/P/S explorations, we scored nodes individually to identify differences in quality across stages. However, during scoring, even when entity nodes correctly identified the people and their task roles, the scoring model still assigned lower scores based on details irrelevant to the task at hand, such as hair color and background content. This caused scores to diverge from the nodes' actual contributions to the task. As the annotation structure became more detailed, it also became harder to assess character correspondences, evidence links, and inferential dependencies comprehensively and consistently, which in turn affected training performance.

Our exploration of modality-separated annotations aimed to clarify information sources by distinguishing audio from visual evidence, making training and evaluation more targeted. However, during generation, the model frequently confused the two types of evidence, placing audio information in the visual evidence section or visual information in the audio evidence section. This was related both to possible confusion in modality attribution within the annotations and to limitations in the base model's ability to distinguish modalities. Separating the two types of evidence therefore did not produce the expected training outcomes and also affected the effectiveness of modality-specific scoring. This exploration shows that explicit section boundaries do not automatically establish clear modality boundaries; process structures need to match annotation quality and model capabilities.

Deploying the scoring model also involves resource costs. Local deployment requires additional GPU memory and computational resources, whereas external APIs reduce local hardware demands but introduce usage fees and latency.

Drawing on these experiences, CreditOmni organizes audiovisual information within a unified evidence segment, obtains segment-level evaluations via an API, and performs credit assignment under the constraint of verifiable outcomes. This is a practical choice that balances training performance, scoring reliability, and available resources. Further improvements should strengthen the task relevance and consistency of these evaluations, enabling model-assisted training to develop multimodal reasoning capabilities more reliably and efficiently.

E/P/S process annotation example

(a) Original video frames and question



Question: Why does the man imitate the woman?

A. The man imitates the woman to make a joke.

B. He wants to make her realize how ridiculous she sounds

C. The man imitates the woman's accent.

D. He wants to express his feelings about her texting

Reference answer: D

(b) E/P/S: 2 entity, 3 perception, and 5 inference nodes

Entity

[E01 entity; speaker_status=visible speaker] The man reacts to the woman's phone use.

[E02 entity; speaker_status=visible speaker] The woman repeatedly types on her phone.

Perception

[P01 perception; modality=visual; time=00:00-00:31; entities=E02] She continually looks down and types instead of engaging with him.

[P02 perception; modality=visual; time=00:11-00:31; entities=E01,E02] He copies her typing while staring pointedly at her.

[P03 perception; modality=visual; time=00:33-00:41; entities=E01,E02] After she notices, he gestures with visible frustration.

Inference and option mapping

[S01 evidence_synthesis <= (P01 AND P02 AND P03)] His imitation and frustration convey disapproval of her texting.

[S02 option_mapping <= (S01); option=A; polarity=contradicts] His frustrated delivery is not playful joking.

[S03 option_mapping <= (S01); option=B; polarity=insufficient] He may have this motive, but his imitation and frustration do not clearly establish it.

[S04 option_mapping <= (); option=C; polarity=insufficient] He does not imitate an accent.

[S05 option_mapping <= (S01); option=D; polarity=supports] The imitation expresses his feelings about her texting.

Response / final answer

He imitates her phone use to show his frustration and disapproval of her continual texting.

Final answer: D

Figure 10: E/P/S process annotations for an actual training sample. Two entity nodes identify the people, and three perception nodes record their relevant actions. An evidence-synthesis node integrates these facts; four option-mapping nodes then determine the answer. The frames above are taken from the original video at 00:05, 00:20, and 00:35. Red text highlights the connection between the integrated inference and the correct answer.

E/P/S: E = Entity (participants); P = Perception (audiovisual observations); S = Synthesis (evidence integration, inference, and option mapping).

Modality-separated process annotation example

(a) Original video frames and question



Question: Why does the man imitate the woman?

A. The man imitates the woman to make a joke.

B. He wants to make her realize how ridiculous she sounds

C. The man imitates the woman's accent.

D. He wants to express his feelings about her texting

Reference answer: D

(b) Modality-separated process annotations

Modality boundary

Full audio and video are available. I must use concrete visual evidence and concrete audio evidence together, and explain how both modalities support the answer.

Visual evidence

The uniformly sampled frames establish the depicted participants and interaction context, but do not independently verify a more specific facial-expression or action claim.

Audio evidence

1. At 00:35-00:43, the woman says, What? I was writing back to a hater. That's why my fingers are going a mile a minute.

2. This dialogue provides the context for the man's earlier actions, confirming that he was imitating her intense texting style.

Social reasoning

The man's exaggerated and performative typing is a form of imitation directed at the woman. When she explains that she was typing fast because she was replying to a "hater," it reveals the subject of their interaction. His imitation is not playful (ruling out A) but rather a pointed gesture to express his opinion or disapproval of her engaging with negativity on her phone. Therefore, his action is best understood as a way to express his feelings about her texting.

Hard negative exclusion

A. The man's expression is serious and annoyed, not playful, so it's unlikely he is trying to make a joke.

B. While this could be an underlying motive, it is not clearly established by the evidence. D is better supported by his imitation and frustration, which express his feelings about her texting.

C. There is no evidence in the audio or video that he is imitating an accent; his imitation is physical, related to her typing speed.

Final answer

D. He wants to express his feelings about her texting

Figure 11: Modality-separated process annotations for the same training sample. Visual and audio evidence are organized separately and combined with modality boundaries, social reasoning, hard-negative exclusion, and the final answer to form a complete response. The video frames, question, options, and reference answer are identical to those in Figure 10.