AVATAR: Reinforcement Learning to See, Hear, and Reason Over Video

Supplementary Material

1. GRPO for Multimodal Reasoning

Recent works increasingly use RL methods, particularly GRPO [16, 19, 21], introducing new challenges related to training efficiency and stability. Wei et al. propose a "cold start" strategy via Supervised Fine-Tuning (SFT) before RL [20], while other approaches improve data efficiency through direct RL with test-time scaling [18]. Concurrent advances enhance GRPO through curriculum learning [12], variance-aware data selection [10], hint-based guidance for difficult samples [8], and combining GRPO and DPO [11]. Key limitations remain, including the "overthinking" issue, addressed by regulating reasoning length [17, 21], and perception errors mitigated via grounded reasoning approaches [6, 15] and new RL objectives such as PAPO's Implicit Perception Loss [19].

Audio-visual reasoning that incorporates audio often involves trade-offs between temporal coverage and spatial resolution. These challenges have been addressed through two-system architectures [29], explicit contextual rewards [28], and task-specific improvements [12, 22]. AVATAR introduces a unified framework that integrates a stratified replay buffer for curriculum learning with Temporal Advantage Shaping (TAS) for credit assignment. This approach improves reasoning and grounding within a single model, eliminating the need for two-system architectures [29] or explicit visual coordinate generation [15].

2. Ablation Studies

RQ1: Does TAS empirically focus the learning signal in practice? We validate that TAS successfully adopts this parabolic-shape in practice. We perform an analysis on the Video-Holmes [3] benchmark, calculating the average per-token loss (the learning signal) against its normalized position. Figure 1 shows the result. The baseline GRPO (blue dashed line) exhibits a relatively flat loss distribution. This is the empirical evidence of *uniform credit assignment*: the learning signal is diluted by being applied unfocused and equally to all tokens, regardless of their importance. In contrast, AVATAR w/ TAS (orange solid line) demonstrates a clear parabolic-shaped loss profile. The shaped advantage $A_{i,t}^{TAS}$ amplifies the learning signal at the critical *planning* ($\tilde{t} \in [0, 0.2]$) and *synthesis* ($\tilde{t} \in [0.8, 1.0]$) phases. This proves TAS is a functioning mechanism: it forces the optimizer to learn most aggressively from errors in the most important parts of the reasoning chain.

RQ2: How do AVATAR's components impact audio-visual reasoning? The impact is evident on reasoning tasks from the WorldSense benchmark (Table 1). The highest improvement is on the *Hallucination* task, a gain primarily driven by our off-policy architecture, which builds a more robust and grounded model by forcing repeated engagement with difficult samples from the hard tier. The effect of TAS is most clear on tasks requiring temporal logic, such as *Temporal Prediction*, which involves forecasting future events, and *Audio Change*, which requires identifying discrete shifts in the audio stream. The U-shaped weighting of TAS is uniquely suited for these problems, by modifying credit for initial tokens, it forces accurate grounding of the video's initial state, while modifying credit for final tokens rewards correct synthesis of how that state evolved over time.

RQ3: What is the impact of the on-policy/off-policy sample ratio? To determine the optimal balance between on-policy and off-policy data, we analyze performance across different sample ratios within a fixed group size of eight (Figure 2). We observe a consistent trend across all bench-

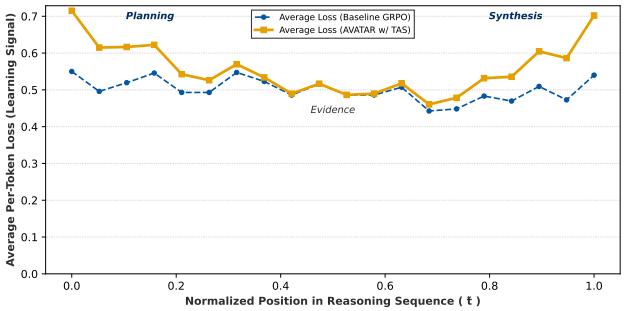
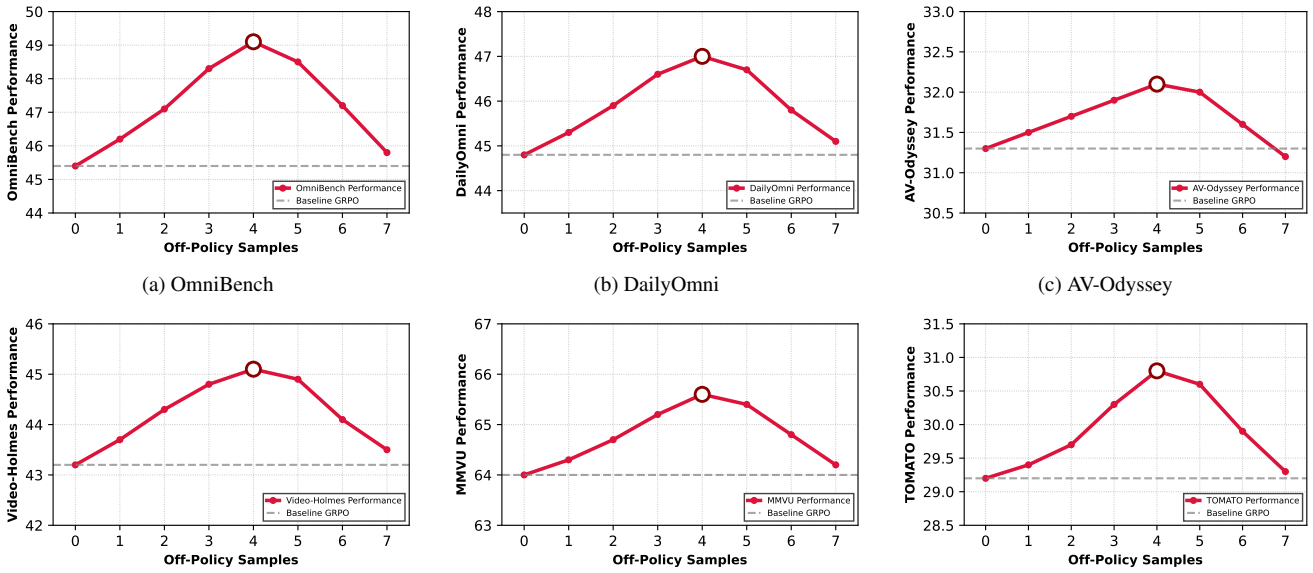


Figure 1. **TAS Empirically Focuses the Learning Signal.** We plot the average per-token loss (the learning signal) vs. normalized position across the Video-Holmes [3] benchmark. The **Baseline GRPO** (blue dashed line) shows a flat, unfocused signal, indicative of **uniform credit assignment**. **AVATAR w/ TAS** (orange solid line) successfully focuses the learning signal, amplifying it on the critical *planning* and *synthesis* phases.

Table 1. **Ablation on key reasoning tasks from the WorldSense benchmark**. Halluc: Hallucination, TP: Temporal Prediction, AC: Audio Change

Config	Halluc	TP	AC
Baseline GRPO	35.6	53.6	32.5
AVATAR w/o Off-Policy	40.1 (+4.5)	54.8 (+1.2)	34.1 (+1.6)
AVATAR w/o TAS	48.9 (+13.3)	56.1 (+2.5)	35.8 (+3.3)
AVATAR (Full)	**51.2** (+15.6)	**57.2** (+3.6)	**37.3** (+4.8)



(a) OmniBench (b) DailyOmni (c) AV-Odyssey

(d) Video-Holmes (e) MMVU (f) TOMATO

Figure 2. **Performance vs. on-policy/off-policy sample ratio across six benchmarks**. **AVATAR achieves optimal performance with** 4-4 **split** (4 on-policy, 4 off-policy samples), demonstrating that balanced mixing prevents both policy drift from excessive off-policy data and sample inefficiency from pure on-policy training.

Table 2. **Ablation comparing the impact of SFT and AVATAR integration**. Models trained with SFT achieve **stronger cross-modal grounding** and **reward alignment**, while removing SFT leads to significant degradation in both GRPO and AVATAR.

Setting	Audio-Visual Benchmarks		
	OmniBench	DailyOmni	AV-Odyssey
Qwen2.5 Omni	44.2	44.0	29.8
GRPO (w/o SFT)	41.2 (-3.0)	40.8 (-3.2)	27.1 (-2.7)
GRPO (w/ SFT)	46.0 (+1.8)	45.5 (+1.5)	30.2 (+0.4)
AVATAR (w/o SFT)	43.8 (-0.4)	43.5 (-0.5)	28.9 (-0.9)
AVATAR (w/ SFT)	**49.1** (+4.9)	**47.0** (+3.0)	**32.1** (+2.3)

marks: performance peaks when using a balanced 4-4 split (4 on-policy, 4 off-policy samples). Using fewer off-policy samples (0-3) fails to fully mitigate the data inefficiency of a purely on-policy approach. On the other hand, over-reliance on off-policy samples (5-7) degrades performance due to instability, as the policy drifts too far from the older, stale data in the replay buffer. The 4-4 split therefore represents an empirical sweet spot for AVATAR, maximizing the data efficiency gains of our off-policy architecture while maintaining the training stability necessary for effective learning.

RQ4: Effect of SFT on Learning Stability. Table 2 shows that incorporating SFT before GRPO provides essential structural grounding that improves convergence and generalization. Without SFT, models lack cross-modal alignment, resulting in unstable rewards and degraded performance across all benchmarks. Initializing GRPO from an SFT-trained policy yields more stable learning and moderate gains, while AVATAR achieves the highest improvements. Hence, SFT establishes a reliable foundation for stable advantage estimation and TAS-based reward shaping, enabling GRPO to refine rather than relearn multimodal reasoning.

RQ5: AVATAR is Generalizes Beyond Video. We train Qwen2.5-VL-7B [1] on the Vision-R1-RL dataset [9] using both standard GRPO and AVATAR, and evaluate on four multimodal math reasoning benchmarks: MathVerse [26], MathVista [13], MMK12 [14], and R1-OneVision-Bench [25]. As shown in Table 3, AVATAR generalizes effectively beyond video, outperforming reasoning models such as OpenVL-Thinker [5] and VLAAThinker [2], with gains of +9.9 on MMK12 and +3.6 on MathVerse. We further observe a distinct ablation pattern compared to audio-visual tasks, where Hinting and TAS emerge as the primary contributors to performance. This difference arises because math reasoning relies on multi-step dependencies, rather than the sparse modality signals of video alignment. In this setting, a single intermediate error can break the entire solution, making TAS (which prevents attention drift) and Hinting (which provides strategic guidance for approaching complex problems) critical for success.

RQ6: Hinting Unlocks the Hard Tail of the Distribution. Figure 3 quantifies the impact of the hinting mechanism on problem solvability during training. The baseline GRPO

Table 3. **Evaluation on multimodal math reasoning.** AVATAR demonstrates strong task-agnostic capabilities by achieving strong results across all four benchmarks.

Model	MathVerse	MathVista	MMK12	R1-OV Bench
State-of-the-Art Methods				
OpenVLThinker-7B [5]	45.8	70.0	53.5	34.7
VLAAThinker-VL-7B [2]	48.2	68.0	51.7	38.4
Baseline				
Qwen2.5-VL-7B	46.0	67.1	47.9	34.9
+ GRPO	48.1	68.7	56.3	36.8
+ **AVATAR**	**49.6** (+3.6)	69.8 (+2.7)	**57.8** (+9.9)	**38.9** (+4.0)
Ablation Studies				
w/o Replay Buffer (on-policy)	49.2 (+3.2)	**70.4** (+3.3)	57.0 (+9.1)	38.5 (+3.6)
w/o TAS (uniform credit)	48.4 (+2.4)	69.2 (+2.1)	56.2 (+8.3)	37.5 (+2.6)
w/o Hinting	48.1 (+2.1)	68.8 (+1.7)	55.6 (+7.7)	37.0 (+2.1)

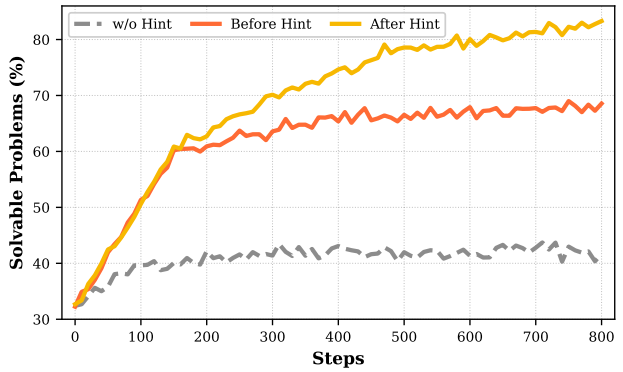


Figure 3. **Hinting unlocks solvability.** While GRPO (grey) stagnates, AVATAR (orange) learns steadily.

(grey) stagnates early ($\approx 40\%$), failing to learn from complex failures. AVATAR without hinting (orange) improves solvability to $\approx 68\%$ via off-policy replay but eventually plateaus. Enabling hinting (yellow) initiates a second phase of growth, increasing solvability to over $\mathbf{80}\%$. This demonstrates that hints effectively unlock the hardest tail of the distribution, namely, problems that otherwise yield zero reward, by converting them into informative learning signals.

RQ7: Efficacy on Reasoning-Heavy Video Benchmarks. To validate AVATAR on tasks demanding complex temporal logic and domain-specific knowledge, we evaluate the Qwen2.5-Omni-7B model across four reasoning-intensive datasets: Video-MMMU [7], VSI-Bench [24], SciVideoBench [4], and Video-TT [27] (Table 4). While standard GRPO yields only marginal improvements over the base model due to uniform credit assignment failing on long-horizon tasks, AVATAR consistently delivers robust performance. It achieves absolute gains of $+2.6$ on Video-MMMU and $+2.2$ on SciVideoBench over the base model, outperforming GRPO across all metrics. These results confirm that our stratified replay and temporal advantage shaping successfully sustain exploration on challenging reasoning paths where standard on-policy RL collapses.

Table 4. **AVATAR on reasoning-heavy video benchmarks.**

Model	Video-MMMU	VSI-Bench	SciVideoBench	Video-TT
Baseline				
Qwen2.5-Omni-7B	46.8	25.4	29.2	41.8
+ GRPO	48.1	25.9	30.3	43.0
+ **AVATAR**	**49.4** (+2.6)	**26.8** (+1.4)	**31.4** (+2.2)	**44.2** (+2.4)

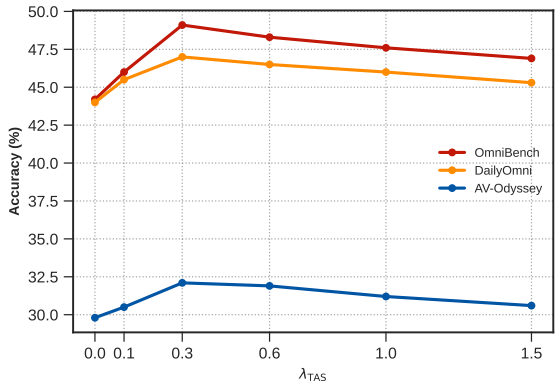
RQ8: Impact of λ_{TAS} on Learning Stability. As illustrated in Figure 4, model performance exhibits a clear three-phase trend with respect to λ_{TAS}. At low values ($\lambda_{\text{TAS}}<0.1$), the reward shaping effect is weak, leading to slower convergence and unstable gains. As λ_{TAS} increases to a moderate level ($\lambda_{\text{TAS}}\approx 0.3$), performance peaks across both audio-visual and video reasoning benchmarks, reflecting a balanced trade-off between stability and exploration. However, at higher values ($\lambda_{\text{TAS}}>1.0$), the model becomes over-regularized, reducing adaptability and leading to gradual accuracy decline.

RQ9: How do isolating rewards impact performance? Table 5 shows the cumulative impact of each reward component in AVATAR's training pipeline. The baseline, which includes only accuracy and format rewards, offers limited learning signals, especially on complex reasoning tasks such as Video-Holmes and MMVU, where binary feedback proves insufficient. Introducing self-rewarding (R_{self}) in Stage 2 enables consensus-based learning from the model's own outputs, yielding moderate improvements on audio-visual tasks that benefit from internal consistency signals (e.g., $+0.7$ on OmniBench, $+1.3$ on MMVU). Adding the stepwise judge reward (R_{judge}) in Stage 3 results in larger gains by assessing intermediate reasoning steps, particularly enhancing performance on fine-grained localization benchmarks (e.g., $+0.8$ on OmniBench, $+1.3$ on MMVU). Finally, incorporating VCRS further stabilizes training by maintaining non-zero advantage baselines via moving averages, allowing the full reward setup to achieve the highest observed performance (e.g., $+0.8$ on OmniBench, $+0.9$ on MMVU).

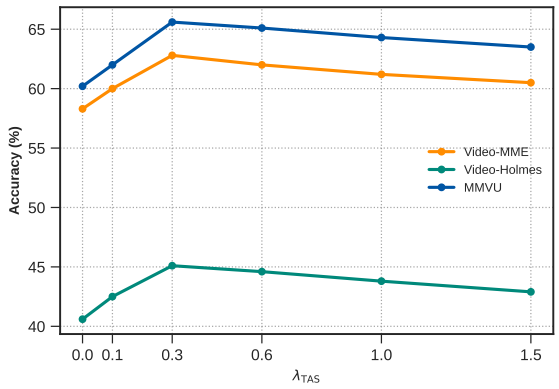
RQ10: How do training dynamics differ across benchmarks? The training curves in Figure 5 reveal distinct learning patterns that directly validate our core contributions. Baseline GRPO shows a consistent decline across all benchmarks due to the vanishing advantage problem: when encountering difficult samples where all responses receive similar rewards, the learning signal collapses to zero, stalling further learning. As a result, GRPO quickly plateaus and fails to move beyond solving only the simplest queries.
In contrast, AVATAR shows a *dip-and-recovery pattern* that reflects the interplay of its key components. The initial drop in performance, especially on challenging benchmarks like OmniBench, Video-Holmes, and MMVU, occurs when the stratified replay buffer transitions from easy to hard samples, exposing the model to its most difficult failures. This performance dip would persist without TAS; however, the U-

Table 5. **Ablation studies of the reward strategy across three training stages**. Stage 1 uses R_{acc} and R_{format} for basic accuracy and format compliance. Stage 2 adds R_{self} (majority vote consensus) to enable more consistent learning. Stage 3 incorporates R_{judge} (stepwise VLM evaluation) alongside previous rewards, providing more detailed reasoning feedback. VCRS maintains stable advantage baselines via moving averages, preventing vanishing advantages. Each stage yields consistent performance gains, with the largest improvements on reasoning-heavy benchmarks (Video-Holmes: $+4.5$, MMVU: $+5.4$), where structured feedback is most beneficial.

Ablation Configuration		Audio-Visual Benchmarks			Video Reasoning Benchmarks		
Group	**Setting**	**OmniBench**	**DailyOmni**	**AV-Odyssey**	**Video-Holmes**	**MMVU**	**TOMATO**
	Baseline Qwen2.5 Omni	44.2	44.0	29.8	40.6	60.2	29.0
Reward Suite	$R_{\text{acc}} + R_{\text{format}}$	46.8 (+2.6)	45.2 (+1.2)	30.1 (+0.3)	42.3 (+1.7)	62.1 (+1.9)	28.9 (-0.1)
	$R_{\text{acc}} + R_{\text{format}} + R_{\text{self}}$	47.5 (+3.3)	45.8 (+1.8)	30.7 (+0.9)	43.1 (+2.5)	63.4 (+3.2)	29.4 (+0.4)
	$R_{\text{acc}} + R_{\text{format}} + R_{\text{self}} + R_{\text{judge}}$	48.3 (+4.1)	46.4 (+2.4)	31.4 (+1.6)	44.2 (+3.6)	64.7 (+4.5)	30.1 (+1.1)
	$R_{\text{acc}} + R_{\text{format}} + R_{\text{self}} + R_{\text{judge}} + \text{VCRS}$	**49.1** (+4.9)	**47.0** (+3.0)	**32.1** (+2.3)	**45.1** (+4.5)	**65.6** (+5.4)	**30.8** (+1.8)



(a) Audio-Visual Reasoning



(b) Video Reasoning

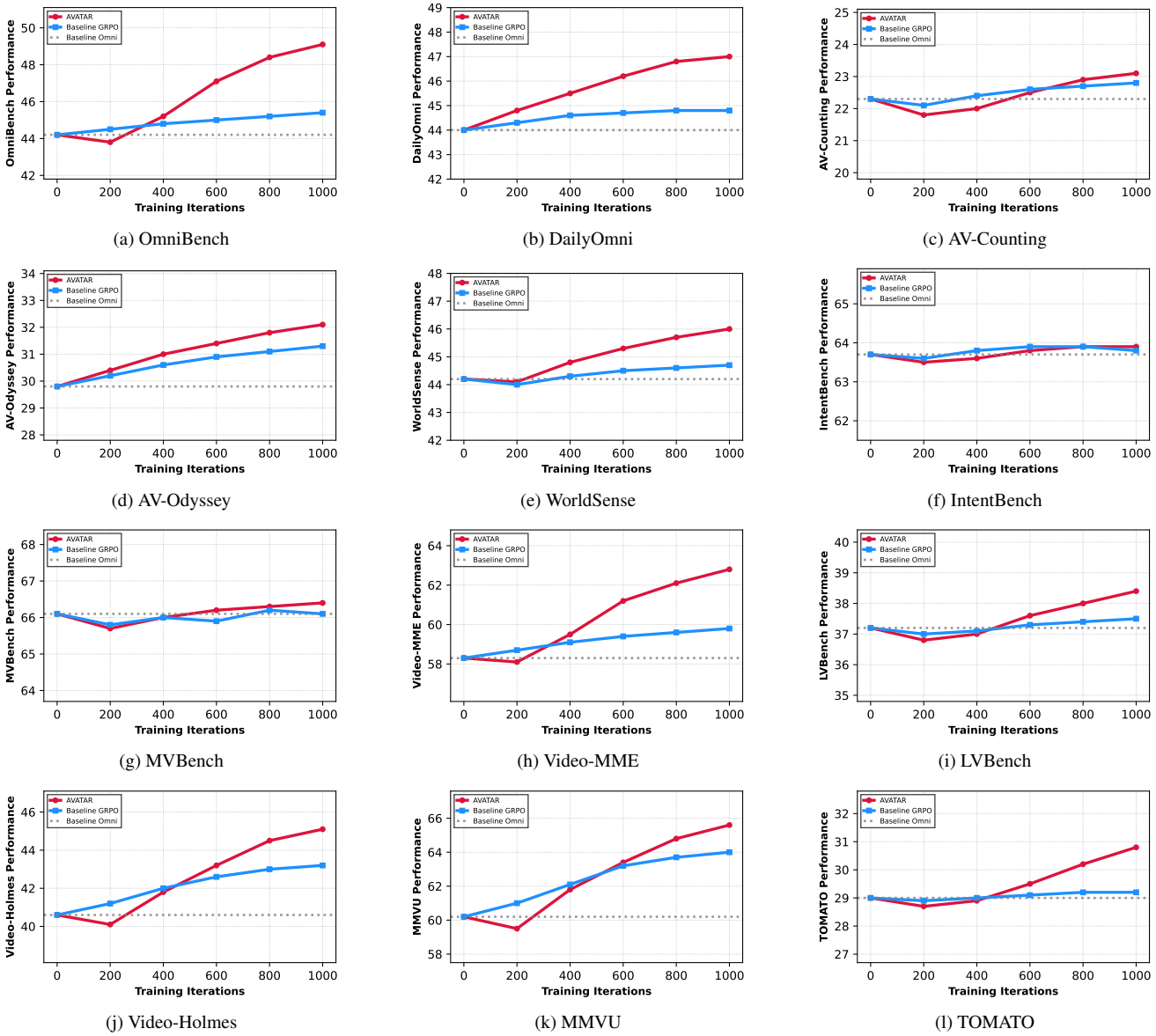
Figure 4. **Performance variation with TAS weighting** (λ_{TAS}). Both (a) **Audio-Visual** and (b) **Video Reasoning** benchmarks peak around λ_{TAS}=0.3, showing improved stability and reward shaping before slight decline at higher values.

shaped temporal weighting in TAS provides focused learning signals that extract meaningful gradients from these difficult examples. By emphasizing the planning and synthesis phases that uniform credit assignment would dilute, TAS enables recovery and subsequent improvement. The model's rebound and improved final performance confirm that our off-policy architecture transforms early failures into learning opportunities through repeated exposure, while TAS ensures that each replayed experience yields maximum value. These gains emerge within just 1000 iterations, underscoring AVATAR's superior sample efficiency and training effectiveness.

RQ11: Does AVATAR generalize to smaller models? To evaluate the generalizability of our approach across different model scales, we conduct experiments with a $3B$ parameter version of the base Qwen2.5 Omni [23] model. Tables 6 and 7 demonstrate that AVATAR consistently outperforms both the baseline and on-policy GRPO across all benchmarks, achieving improvements of $+3.4$ on OmniBench, $+2.9$ on Video-Holmes, and $+2.8$ on MMVU. These results validate that AVATAR's core components, the stratified replay buffer and TAS, effectively enhance multimodal reasoning capabilities regardless of language model scale. The consistent performance patterns across different model sizes further confirm that our off-policy architecture is broadly applicable, with particularly strong benefits observed on complex reasoning tasks, where AVATAR's targeted learning signals are most impactful in guiding model improvement.

2.1. Prompt Templates

To support AVATAR's training pipeline, we use two prompt templates for different components of our framework. The *hint generation prompt* (Figure 6) is used by our stratified replay buffer's hint mechanism when the policy becomes stuck in local optima on challenging samples. This prompt instructs a teacher model to generate strategic hints without revealing the final answer, helping the policy escape stagnation while maintaining the learning challenge. The *audio-visual localization judge prompt* (Figure 7) helps our stepwise reasoning judge in Stage 3 training by providing granular evaluation criteria across four dimensions: grounding of audio cues, identification of visual objects, spatial localization accuracy, and caption correctness. This offers detailed feedback on cross-modal reasoning quality rather



(a) OmniBench (b) DailyOmni (c) AV-Counting

(d) AV-Odyssey (e) WorldSense (f) IntentBench

(g) MVBench (h) Video-MME (i) LVBench

(j) Video-Holmes (k) MMVU (l) TOMATO

Figure 5. **Training curves across audio-visual and video reasoning benchmarks**. **AVATAR demonstrates superior sample efficiency and final performance**, particularly on challenging reasoning tasks. AVATAR's initial dip, followed by a strong recovery, validates the effectiveness of our off-policy architecture and TAS for credit assignment.

than binary success/failure signals.

2.2. Qualitative Analysis

Figure 8 illustrates the qualitative differences between baseline GRPO and AVATAR on audio-visual reasoning tasks. AVATAR demonstrates *superior cross-modal integration* by linking visual cues ("tense expression, his eyes darting around") with audio analysis ("hurried and tense tone when he speaks"), while baseline GRPO makes *disconnected observations*, such as "He looks worried which indicates stress" without linking modalities. The responses highlight the effectiveness of TAS, evident in the model's ability to *plan* by establishing comprehensive scene context, focus on relevant cues while avoiding *redundant pattern matching* (e.g., "Black clothing often symbolizes negativity"), and *synthesize* complex emotional dynamics rather than relying on *surface-level assumptions* ("something bad might be happening"). AVATAR also shows *improved temporal reasoning* by tracking emotional progression ("tone shifting from calm to anxious") and *enhanced contextual understanding* through precise dialogue interpretation ("Sorry, I have a train to catch" indicating abrupt departure). In contrast, baseline GRPO

Table 6. **Audio-visual benchmarks: $3B$ model comparison**. AVATAR 3B consistently outperforms both the baseline $3B$ Omni and baseline GRPO across all benchmarks, achieving improvements of $+3.4$ on OmniBench and $+2.6$ on AV-Odyssey. The results demonstrate that our off-policy architecture and TAS effectively enhance multimodal reasoning capabilities even with smaller model scales, validating the broad applicability of our approach. Performance improvements over baseline 3B Omni are shown in green.

	Audio-Visual Benchmarks					
Model	**OmniBench**	**DailyOmni**	**AV-Counting**	**AV-Odyssey**	**WorldSense**	**IntentBench**
Baseline $3B$ Omni	42.4	42.9	19.7	27.2	35.2	59.7
Baseline GRPO	43.1 (+0.7)	43.4 (+0.5)	20.0 (+0.3)	28.5 (+1.3)	35.8 (+0.6)	59.9 (+0.2)
AVATAR $3B$	**45.8** (+3.4)	**44.7** (+1.8)	**20.9** (+1.2)	**29.8** (+2.6)	**37.1** (+1.9)	**60.5** (+0.8)

Table 7. **Video understanding and reasoning benchmarks: $3B$ model comparison**. AVATAR $3B$ demonstrates superior performance across both general video understanding and complex reasoning tasks, with particularly strong gains on reasoning-heavy benchmarks like Video-Holmes ($+2.9$) and MMVU ($+2.8$). The consistent improvements over baseline GRPO highlight how our stratified replay buffer and TAS enhance learning efficiency even with reduced model capacity, confirming that our framework's benefits scale effectively across different model sizes. Performance improvements over baseline 3B Omni are shown in green.

	General Video Understanding			**Video Reasoning**		
Model	**MVBench**	**Video-MME**	**LVBench**	**Video-Holmes**	**MMVU**	**TOMATO**
Baseline $3B$ Omni	59.2	57.6	34.3	39.2	62.0	27.0
Baseline GRPO	59.8 (+0.6)	58.1 (+0.5)	34.7 (+0.4)	40.8 (+1.6)	63.2 (+1.2)	27.3 (+0.3)
AVATAR $3B$	**61.2** (+2.0)	**59.9** (+2.3)	**35.8** (+1.5)	**42.1** (+2.9)	**64.8** (+2.8)	**28.4** (+1.4)

tends to fall back on *generic genre classifications* ("suggests a thriller or drama genre"). These qualitative differences validate our framework's ability to enhance the quality of multimodal reasoning through targeted credit assignment and structured learning signals.

Hint Generation Prompt for Hard Samples

You are an expert AI tutor. Your task is to provide a concise, strategic hint to a student model that is struggling with a difficult audio-visual reasoning problem. The student model is consistently failing this problem and has stopped exploring new reasoning paths. Your hint should guide it out of a local minimum without solving the problem for it.
Given the following full problem context:
[User Query]: {The original question posed to the student model.}
[Audio Caption]: {A transcript of the relevant audio from the video clip.}
[Video Context]: {The full video clip associated with the query.}
[Ground Truth Answer]: {The correct final answer to the query.}

Your Task: Based on the full context and the ground truth, generate a single, short sentence that provides a high-level strategy for how to approach the problem.
Constraints:
- **DO NOT** reveal or allude to the final answer.
- **DO NOT** perform the reasoning steps yourself.
- Your hint must be a strategic suggestion (e.g., "focus on the audio first," "compare the objects on the left and right," "track the object's movement over time").

Example Hint: `first locate the object making the sound, then count`
Output: Provide ONLY the text of the hint.

Figure 6. The prompt template used to generate guiding hints for the main policy model.

Audio-Visual Localization Judge Prompt

You are a meticulous and precise Audio-Visual Grounding Evaluator. Your task is to provide a granular, stepwise evaluation of a model's attempt to localize an object based on audio and visual cues.
Given the following inputs, you will score the model's reasoning and final answer on several criteria.
[User Query]: {The user's original question to the model.}
[Audio Caption]: {A transcript of the relevant audio from the video clip.}
[Ground Truth Answer]: {The correct description and location of the target object.}
[Model's Reasoning]: {The reasoning process generated by the model inside its `<think>` tags.}
[Model's Final Answer]: {The final location and description from the model's `<answer>` tag.}

Your Task: Based on your analysis of all the provided information, you must evaluate the model's performance on the following four criteria. Provide your judgment as a single JSON object with a score from 0.0 (complete failure) to 1.0 (perfect) for each criterion.

1. **Audio Cue Grounding (`audio_grounding_score`):** Did the model's reasoning correctly identify the key descriptive words in the `[Audio Caption]` that point to the target? (e.g., did it correctly identify "clapping" as the key sound?).
2. **Visual Object Identification (`visual_id_score`):** Based on the audio cue, did the model's reasoning correctly identify the corresponding visual object in the scene? (e.g., after identifying "clapping," did it correctly identify the "audience" as the source?).
3. **Location Accuracy (`location_acc_score`):** How accurate is the final spatial location (e.g., "Left," "Center") provided in the `[Model's Final Answer]` when compared to the `[Ground Truth Answer]`?
4. **Caption Correctness (`caption_corr_score`):** How well does the model's final textual description of the object in its `[Model's Final Answer]` match the `[Ground Truth Answer]`?

Output Format: Return ONLY a single JSON object with your scores.
Example Output:

```
{''audio_grounding_score'': 0.9, ''visual_id_score'': 1.0,
''location_acc_score'': 1.0, ''caption_corr_score'': 0.8}
```

Figure 7. The prompt template for our VLM-based judge in Stage 3. It guides the judge to produce a granular, stepwise reward by scoring the policy model's reasoning on audio grounding, visual identification, location accuracy, and caption correctness.

Query: Describe the relationship between the audio and visual elements in this video.

Baseline GRPO: The video shows a man and woman at a train station. The man is wearing stripes and the woman looks worried. This suggests they might be in a relationship. The train station setting indicates travel. The woman's expression shows concern which could mean she's sad about departure. The man sits on a bench so he might be waiting for a train. The dialogue mentions catching a train which confirms the travel theme. The woman eats an apple which shows she's passing time. The close-up shot emphasizes emotion. The voices have different tones indicating contrasting moods. The striped shirt is distinctive making the man easily identifiable. Train stations are associated with goodbyes so this is likely a farewell scene.

AVATAR: A man in striped shirt and white gloves walks towards a train while a concerned woman watches. He sits on a bench appearing lost in thought, she eats an apple with a tense expression. He approaches her from behind suggesting urgency. The scene ends with a close-up of her face with train in background symbolizing separation. Audio shows her tone shifting from calm to anxious while his voice remains gentle and soothing. Dialogue snippets ("Is that you?", "Sorry, I have a train to catch," "Very funny") indicate identity check, abrupt departure, and sarcastic remark respectively, creating a tense, emotionally charged moment at the train station.

Query: What emotions are conveyed through both the visual and audio elements?

Baseline GRPO: The video shows a man in a black beanie. He looks worried which indicates stress. There's also a woman with a bun. She appears concerned so something bad might be happening. The man is driving a car and his posture suggests tension. Gas stations are often used in movies for dramatic scenes. The audio has music that sounds intense which creates suspense. The man says "I'm sorry" which means he feels bad. Black clothing often symbolizes negativity. The woman's expression shows she's upset about whatever situation they're in. Fast music usually indicates action or tension. The overall combination suggests a thriller or drama genre.

AVATAR: From the video, I can see the man in the black beanie and jacket has a tense expression, his eyes darting around, and he looks anxious. The woman with the bun appears concerned. In the car, he's driving with his head down and breathing heavily, which suggests guilt. At the gas station, his interaction with the woman still feels tense. The overall atmosphere of the video is described as tense and suspenseful. From the audio, the man's hurried and tense tone when he speaks, especially when he repeatedly says "I'm sorry," further highlights his guilt. The fast-paced, high-pitched music in the background also contributes to the tense mood.

Figure 8. **Qualitative comparison between Baseline GRPO and AVATAR**. AVATAR demonstrates better **cross-modal integration** and **structured reasoning**, while Baseline GRPO shows *surface-level observations* and *disconnected audio-visual analysis*. Green highlighting shows AVATAR's effective synthesis of multimodal cues, red highlighting reveals GRPO's uniform credit assignment limitations.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, and et al. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *Transactions on Machine Learning Research (TMLR)*, 2025. 2, 3

[3] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and et al. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. 1

[4] Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, and et al. Scivideobench: Benchmarking scientific video reasoning in large multimodal models. *arXiv preprint arXiv:2510.08559*, 2025. 3

[5] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and et al. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. In *Neural Information Processing Systems (NeurIPS)*, 2025. 2, 3

[6] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, and et al. Grit: Teaching mllms to think with images. In *Neural Information Processing Systems (NeurIPS)*, 2025. 1

[7] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, and et al. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. 3

[8] Qihan Huang, Weilong Dai, Jinlong Liu, Wanggui He, Hao Jiang, and et al. Boosting mllm reasoning with text-debiased hint-grpo. In *International Conference on Computer Vision (ICCV)*, 2025. 1

[9] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, and et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations (ICLR)*, 2026. 2

[10] Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, and et al. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*, 2025. 1

[11] Yunxin Li, Xinyu Chen, Zitao Li, Zhenyu Liu, Longyue Wang, and et al. Veripo: Cultivating long reasoning in video-llms via verifier-gudied iterative policy optimization. *arXiv preprint arXiv:2505.19000*, 2025. 1

[12] Lidong Lu, Guo Chen, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms. *arXiv preprint arXiv:2506.05328*, 2025. 1

[13] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, and et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[14] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, and et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 2

[15] Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, and et al. Grounded reinforcement learning for visual reasoning. In *Neural Information Processing Systems (NeurIPS)*, 2025. 1

[16] Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, and et al. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2025. 1

[17] Song Wang, Gongfan Fang, Lingdong Kong, Xiangtai Li, Jianyun Xu, and et al. Pixelthink: Towards efficient chain-of-pixel reasoning. *arXiv preprint arXiv:2505.23727*, 2025. 1

[18] Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and et al. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2025. 1

[19] Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, and et al. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations (ICLR)*, 2026. 1

[20] Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, and et al. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025. 1

[21] Wenyi Xiao, Leilei Gan, Weilong Dai, Wanggui He, Ziwei Huang, and et al. Fast-slow thinking for large vision-language model reasoning. In *Neural Information Processing Systems (NeurIPS)*, 2025. 1

[22] Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and et al. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025. 1

[23] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, and et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 4

[24] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and et al. Thinking in space: How multimodal large language models see, remember, and

recall spaces. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[25] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, and et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *International Conference on Computer Vision (ICCV)*, 2025. 2

[26] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, and et al. Mathverse: Does your multimodal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision (ECCV)*, 2024. 2

[27] Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and et al. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[28] Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, and et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025. 1

[29] Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, and et al. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. In *Neural Information Processing Systems (NeurIPS)*, 2025. 1