ESCAPE THE LANGUAGE PRIOR: MITIGATING LATE-STAGE MODALITY COLLAPSE IN AUDIO REASONING VIA MODALITY-AWARE POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio and omni-modal large language models exhibit impressive cross-modal reasoning capabilities. However, standard reinforcement learning methods such as GRPO do not explicitly weight individual tokens by their dependence on the non-text source modality. This exacerbates *late-stage modality collapse* during extended chain-of-thought generation, where models progressively abandon the primary source signal in favor of compressed textual priors, leading to confident but ungrounded hallucinations. To address this, we introduce Modality-Aware Policy Optimization (MAPO), a novel dual-branch reinforcement learning framework. First, MAPO dynamically concentrates the policy gradient on modality-critical tokens using a modality relevance mask, which is derived from the cross-modal differential entropy between an audio-ablated reference and the multimodal policy. Second, it integrates an auxiliary attention loss branch that applies a targeted, temporally scaled penalty to the model's internal attention distributions. This encourages sustained attention to the source modality throughout the reasoning trace. Evaluations on complex audio reasoning benchmarks demonstrate that MAPO substantially improves long-horizon reasoning fidelity and multimodal instruction following, achieving highly competitive performance on several key benchmarks among open-weights models. By relying strictly on native statistical signals rather than domain-specific inductive biases, MAPO offers a promising foundation for mitigating epistemic collapse across diverse multimodal systems.

1 INTRODUCTION

Audio and omni-modal large language models have demonstrated remarkable capabilities by integrating non-text modalities into complex reasoning systems. However, post-training these models using standard reinforcement learning algorithms (e.g., PPO (Schulman et al., 2017), DPO (Rafailov et al., 2024), GRPO (Shao et al., 2024)) exposes a critical structural inefficiency: their objectives do not explicitly weight individual tokens by their dependence on the source modality. In cross-modal generation, tokens are fundamentally not equal. Many serve purely as linguistic scaffolding that the model's internal language prior can accurately predict. Conversely, a critical subset of tokens relies entirely on the non-text source modality to capture essential acoustic attributes or events. Uniform token coefficients provide no explicit mechanism to prioritize positions requiring active cross-modal grounding.

Maintaining access to the source is especially important during extended cross-modal chain-of-thought (CoT) (Wei et al., 2023) reasoning. We find that, in this process, models can increasingly abandon the uncompressed primary source signal (e.g., raw audio) in favor of their own heavily compressed textual summaries, a pattern we refer to as *late-stage modality collapse*. Relying on these intermediate captions causes the model's reasoning to drift into ungrounded hallucinations. Crucially, as the model succumbs to its language prior and ignores the source modality, it becomes confidently disconnected. This epistemic collapse masks the model's true uncertainty, which makes standard statistical interventions such as entropy maximization ineffective at recovering lost grounding.

To address these intertwined challenges, we introduce **Modality-Aware Policy Optimization (MAPO)**, a novel RL framework designed to break the language prior and encourage sustained

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.

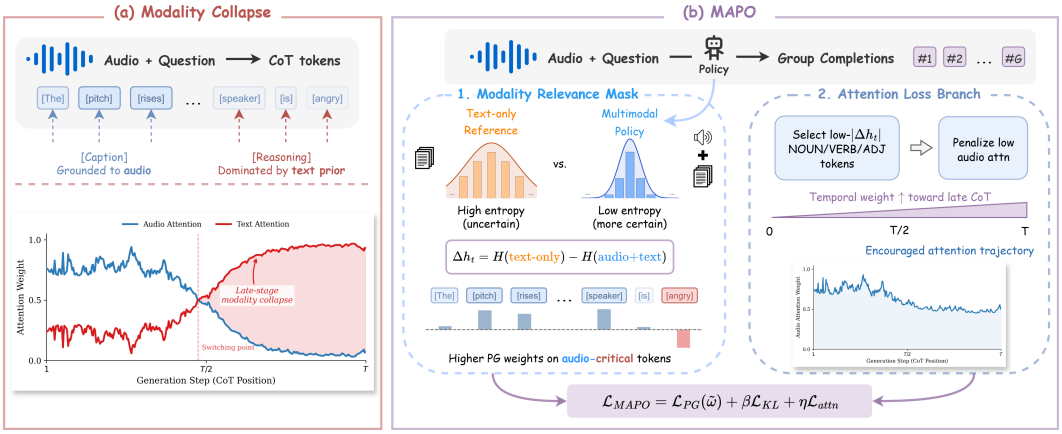


Figure 1: Overview of the MAPO framework. **(a)** Late-stage modality collapse, where attention shifts from the source audio to the text prior during CoT reasoning. **(b)** MAPO mitigates this via a dual-branch architecture. A *modality relevance mask* uses cross-modal differential entropy (Δh_t) to focus the policy gradient on audio-critical tokens. Simultaneously, an *attention loss branch* applies a temporally scaled penalty to substantive tokens with low audio attention, encouraging sustained access to source audio.

cross-modal grounding. Building upon GRPO, MAPO modifies the learning objective through two complementary mechanisms. First, it introduces a *Modality Relevance Mask* that reweights the policy gradient. By computing the cross-modal differential entropy between a reference model with text-only inputs and the multimodal policy, MAPO dynamically upweights the gradient contribution of tokens where the source modality significantly alters the predictive distribution. Second, an auxiliary *Attention Loss Branch* directly combats late-stage modality collapse. Guided by an inverse relevance mask and temporally increasing weights, it explicitly penalizes the neglect of the source modality at linguistically substantive tokens (identified via part-of-speech gating). Combined with a soft task-failure gate, this penalty encourages the model to maintain attention to the source signal deep into the reasoning chain. We focus our empirical evaluation on a wide variety of audio reasoning tasks as a rigorous testbed.[1]

In summary, our main contributions are threefold:

1. We identify and formalize the limitations of uniform token weighting and the phenomenon of late-stage modality collapse in cross-modal reinforcement post-training, highlighting how the language prior masks true model uncertainty during long-horizon reasoning.
2. We propose MAPO, a dual-branch optimization framework that leverages cross-modal differential entropy to prioritize modality-critical tokens and introduces a targeted attention penalty to enforce sustained cross-modal grounding.
3. Through extensive empirical evaluation on complex audio reasoning tasks (encompassing speech, music, and sound events), we demonstrate that MAPO improves both the fidelity of long-form reasoning and the model's robustness against language-prior hallucinations, establishing a robust algorithmic foundation for mitigating late-stage modality collapse.

2 Related Work

Omni-modal large language models. Recent advancements in omni-modal large language models enable unified, native processing of text, vision, and audio, bypassing the information loss inherent to cascaded systems. Our empirical investigations utilize Qwen3-Omni-Thinking (Xu et al., 2025b), a model specifically optimized for extended cross-modal chain-of-thought generation, distinguishing it from its instruction-following counterpart, Qwen3-Omni-Instruct (Xu et al., 2025b). Beyond Qwen3, the cross-modal reasoning landscape includes proprietary systems like the Gemini (Team, 2025a)

[1]Our code is available at `https://anonymous.4open.science/r/MAPO-release-739C`.

series. It also includes open-weight models such as Audio Flamingo 3 (Goel et al., 2025), Step-Audio-2 (Team, 2025c), MiMo-Audio (Team, 2025b), and Covo-Audio (Wang et al., 2026b), which pioneer specialized native audio reasoning and full-duplex conversational capabilities. However, despite these architectural leaps, those systems are trained via standard text-based post-training objectives. By uniformly weighting all generated tokens, these models remain structurally susceptible to language-prior dominance and late-stage modality collapse.

Group Relative Policy Optimization (GRPO). GRPO (Shao et al., 2024) aligns models efficiently by estimating advantages over a sampled group rather than utilizing a separate value model. For a prompt x and a group of G completions $\{y^{(g)}\}_{g=1}^{G}$, the relative advantage $\hat{A}^{(g)}$ normalizes the reward $r(x, y^{(g)})$ within the group:

$$\hat{A}^{(g)} = \frac{r(x, y^{(g)}) - \text{mean}(\{r(x, y^{(g')})\})}{\text{std}(\{r(x, y^{(g')})\}) + \varepsilon}$$

The policy π_θ is optimized via a clipped surrogate objective regularized by a per-token sample-based estimator of the Kullback-Leibler (KL) divergence against a frozen reference policy π_{ref}. Specifically, we utilize the k_3 unbiased estimator to ensure computational efficiency over the full vocabulary sum. For the g-th completion of length T_g with importance ratio $\rho_t^{(g)} = \frac{\pi_\theta(y_t^{(g)}|y_{<t}^{(g)}, x)}{\pi_{\text{old}}(y_t^{(g)}|y_{<t}^{(g)}, x)}$:

$$\mathcal{L}_{\text{PG}}^{\text{GRPO}} = -\frac{1}{T_g}\sum_{t=1}^{T_g} \min\Big(\rho_t^{(g)}\hat{A}^{(g)},\ \text{clip}(\rho_t^{(g)}, 1-\epsilon, 1+\epsilon)\,\hat{A}^{(g)}\Big)$$

$$\mathcal{L}_{\text{KL}} = \frac{1}{T_g}\sum_{t=1}^{T_g}\Big[\exp\big(\log\pi_{\text{ref}}(y_t^{(g)}) - \log\pi_\theta(y_t^{(g)})\big) - \big(\log\pi_{\text{ref}}(y_t^{(g)}) - \log\pi_\theta(y_t^{(g)})\big) - 1\Big]$$

The final objective is $\mathcal{L}_{\text{GRPO}} = \mathcal{L}_{\text{PG}}^{\text{GRPO}} + \beta\mathcal{L}_{\text{KL}}$. GRPO assigns the same weight, $1/T_g$, to each token when computing the policy loss for a completion. The resulting gradient contributions nevertheless differ across tokens because they depend on token-specific likelihood ratios and score-function gradients. To make the weighting explicitly reflect dependence on the source modality, MAPO uses cross-modal differential entropy to determine each token's relative weight.

Cross-Modal Alignment and Grounding. Recent approaches inject modality-aware feedback through external rewards, often relying on large teacher models to provide auxiliary sequence-level signals (e.g., (Wu et al., 2025), (He et al., 2026)). While effective, these extrinsic methods apply feedback uniformly, algorithmically ignoring the crucial token-by-token variance in modality dependence. Because MAPO intrinsically resolves this granular inefficiency without requiring heavy external reward models, it serves as a distinct, orthogonal framework that remains fully compatible with these sequence-level reward strategies.

At the granular level, prior methods such as PAPO (Wang et al., 2026d) and concurrent works in the vision-language domain have recognized the limitations of uniform token weighting and attention decay. For instance, (Huang et al., 2025) isolates perceptually pivotal tokens to focus policy updates, while (Wang et al., 2026c) combats temporal visual forgetting by progressively elevating visual expectations through attention compensation. While these methods share our high-level motivation of targeted policy optimization, they rely heavily on domain-specific visual similarity and dependency metrics.

In contrast, MAPO intrinsically addresses the epistemic failure mode without requiring external reward models or domain-specific heuristics. Previous granular approaches (Zhongxing et al., 2026) have utilized standard predictive entropy to isolate dependent tokens. Yet, standard policy entropy fails during late-stage modality collapse. As the language prior takes over, the model becomes pathologically confident in its hallucinations, causing the entropy signal to vanish exactly when external grounding is needed most. MAPO overcomes this by anchoring its metric against a text-only reference, exposing hidden modality dependence and directly steering the internal attention distribution.

3 MODALITY-AWARE POLICY OPTIMIZATION (MAPO)

To formalize the MAPO framework, we first establish the cross-modal reinforcement learning problem setting and define the notation used throughout our methodology. Following this, we introduce two critical quantitative metrics that serve as the foundation of our method: a measure of the model's internal cross-modal grounding (*audio attention mass*), and a statistical signal to isolate key modality-dependent tokens (*cross-modal differential entropy*). We utilize these metrics to dissect the phenomenon of late-stage modality collapse before introducing our dual-branch optimization objective.

3.1 PROBLEM FORMULATION

We consider a cross-modal generation setting where an omni-modal language model, featuring a transformer decoder with L layers and H attention heads per layer, acts as the policy π_θ parameterized by θ. The model receives a multimodal prompt x, which inherently comprises both a text instruction sequence x_{text} and a non-text source modality x_{audio} (e.g., continuous acoustic features mapped to the model's input space). Given the prompt x, the policy autoregressively generates a text completion $y = (y_1, \ldots, y_T)$, where each token $y_t \in \mathcal{V}$ belongs to the textual vocabulary $\mathcal{V}$, and T denotes the total number of generated tokens for that specific completion. The likelihood of generating the full sequence is formulated as $\prod_{t=1}^{T} \pi_\theta(y_t \mid y_{<t}, x)$.

During post-training via GRPO, for a given prompt x, the policy samples a group of G independent completions, denoted as $\{y^{(g)}\}_{g=1}^{G}$. Standard reinforcement learning objectives construct the sequence-level loss by uniformly averaging the token-level losses over the sequence length T. MAPO challenges this uniform $1/T$ dependency. Our objective is to design a dynamic, token-level weighting mechanism that structurally differentiates tokens relying purely on the language prior (dependent only on $y_{<t}$ and x_{text}) from those requiring cross-modal perception (critically dependent on x_{audio}).

3.2 AUDIO ATTENTION MASS AND LATE-STAGE MODALITY COLLAPSE

To dynamically quantify the extent to which the model grounds its reasoning in the non-text source signal at any given generation step, we track the *audio attention mass*. Conceptually, this metric captures the *aggregate probability mass within the model's self-attention distribution that is routed specifically to the cross-modal input sequence*, serving as a direct proxy for the active perceptual bandwidth allocated to the audio signal.

Let $\alpha_{t,j}^{(l,h)}$ denote the attention weight that the query at generation position t assigns to the key at sequence position j within layer $l \in \{1, \ldots, L\}$ and attention head $h \in \{1, \ldots, H\}$. If $\mathcal{S}_{\text{audio}}$ represents the set of token indices corresponding to the primary audio input x_{audio}, the raw audio attention mass for a specific head is defined as $m_t^{(l,h)} = \sum_{j \in \mathcal{S}_{\text{audio}}} \alpha_{t,j}^{(l,h)}$. Empirically, cross-modal perception is rarely distributed uniformly across all attention heads; rather, specific heads specialize in modality routing. To characterize this behavior, we define both the *max-head* and *mean-head* attention mass for a given layer l:

$$m_t^{(l,\max)} = \max_h m_t^{(l,h)}, \quad m_t^{(l,\text{mean})} = \frac{1}{H} \sum_{h=1}^{H} m_t^{(l,h)}$$

To capture the strongest grounding signal for our optimization objective, we utilize the max-head reduction. The final audio attention mass a_t for token t is computed by averaging this max-head signal across a targeted set of deep transformer layers L_{tgt} (e.g., the final layers before the output projection):

$$a_t = \frac{1}{|L_{\text{tgt}}|} \sum_{l \in L_{\text{tgt}}} m_t^{(l,\max)}$$

This metric exposes the structural vulnerability of cross-modal CoT generation. As illustrated by the line charts in Figure 2, while the initial reasoning steps maintain a high audio attention mass, this mass exhibits a pronounced decay as the generation lengthens. In the late stages of generation, a_t

collapses near zero, indicating that the model has effectively stopped "listening" to the raw audio, instead auto-regressively feeding on its own textual context.

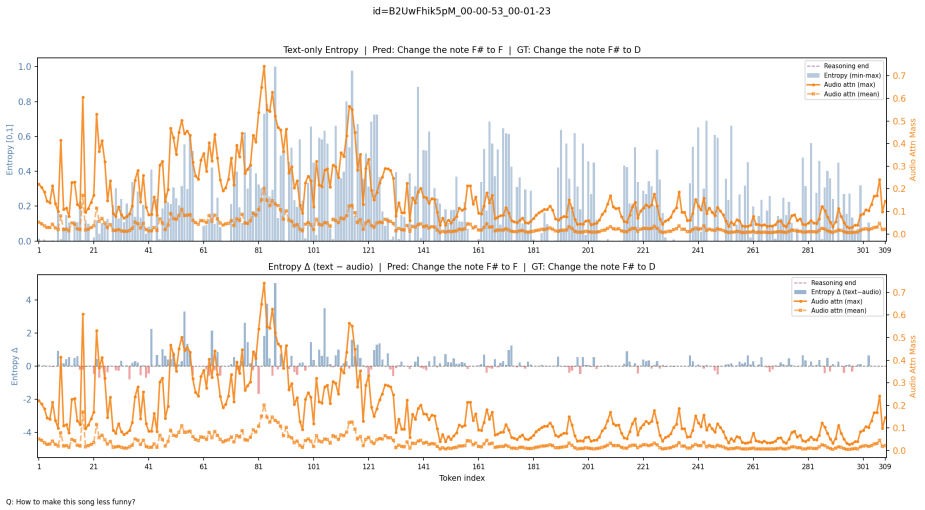


Figure 2: Late-stage modality collapse and key token extraction. Plots display the temporal decay of audio attention mass (**solid:** max-head reduction, *dashed:* mean-head reduction) alongside uncertainty metrics. **(a)** Text-only entropy fails to track this loss of source grounding. **(b)** cross-modal differential entropy effectively isolates true "key tokens"; this signal correlates with the attention decay.

3.3 Cross-modal differential entropy

To counteract modality collapse, we must reliably identify the "key tokens", which represent the specific reasoning steps that require audio perception. To expose this hidden modality dependence, we introduce *cross-modal differential entropy* (Δh_t), the token-level predictive entropy difference between a text-only reference policy and the multimodal policy:

$$\Delta h_t = H(\pi_{\text{text-ref}}(\cdot \mid y_{<t}, x_{\text{text}})) - H(\pi_\theta(\cdot \mid y_{<t}, x))$$

Both terms use the same text prompt and generated prefix $y_{<t}$, including any caption already generated; only x contains source audio. In our diagnostic analyses, both terms are evaluated using the same fixed model checkpoint. This comparison isolates the effect of audio conditioning on predictive uncertainty at each generation position, conditional on the information already expressed in the prefix. As demonstrated in Figure 2(b), cross-modal differential entropy correlates with the trend of audio mass decay.

During training, we compute cross-modal differential entropy between the live multimodal policy and a text-only reference policy. The reference policy is configurable: its parameters may follow the current policy or remain fixed. Our experiments use a fixed reference for GPU memory efficiency, reusing the model already maintained for the GRPO KL penalty (Appendix H.1).

3.4 Modality relevance mask and reweighted policy gradient

MAPO incorporates cross-modal differential entropy into the GRPO objective through a token-level *Modality Relevance Mask* ($\tilde{\omega}$). This mask replaces uniform averaging coefficients with weights determined by $|\Delta h_t|$, assigning greater emphasis to positions with larger absolute cross-modal differential entropy, subject to clipping.

To construct this mask, we apply a softmax over the absolute differential entropy $|\Delta h_t|$. Because cross-modal CoT reasoning lengths vary drastically, we stabilize the weight assignment using a length-scaled temperature $\tau_T = \tau_{\text{base}} \log(\max(T, 2))$. To keep token weights comparable across completions of different lengths, we multiply the softmax weights by T, giving a mean weight of one before clipping. We then cap each weight at C_{mask} to limit large individual weights. The final

detached mask is:

$$\tilde{\omega}_t = \min\left(T \cdot \frac{\exp\big(|\Delta h_t|/\tau_T\big)}{\sum_{t'=1}^{T} \exp\big(|\Delta h_{t'}|/\tau_T\big)},\ C_{\text{mask}} \right) \tag{1}$$

Clipping can reduce the total weight assigned to a completion. We therefore normalize by the sum of the weights, so the policy loss remains a weighted average over the generated tokens. Let $\ell_t^{\text{PG}} = -\min\big(\rho_t \hat{A},\ \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)\,\hat{A}\big)$ denote the GRPO policy loss at token t. For a sampled group of G completions $\{y^{(g)}\}_{g=1}^{G}$ with lengths T_g, the MAPO reweighted policy gradient replaces the uniform $1/T$ averaging with:

$$\mathcal{L}_{\text{PG}}(\tilde{\omega}) = \frac{1}{G} \sum_{g=1}^{G} \frac{\sum_{t=1}^{T_g} \tilde{\omega}_t^{(g)} \cdot \ell_t^{\text{PG},(g)}}{\sum_{t=1}^{T_g} \tilde{\omega}_t^{(g)}} \tag{2}$$

This normalization makes the policy loss invariant to a common positive rescaling of the final detached weights, while retaining the relative token weights assigned by the modality relevance mask.

3.5 Attention Loss Branch

While the modality relevance mask corrects the policy gradient budget, it fundamentally relies on the model maintaining some latent awareness of the cross-modal signal. To directly combat late-stage modality collapse, MAPO incorporates an auxiliary *attention loss branch* ($\mathcal{L}_{\text{attn}}$) that penalizes the neglect of the source signal deep within the reasoning chain. We define a token-level log-penalty based on the audio attention mass: $\ell_t^{\text{attn}} = -\log(a_t + \varepsilon)$. To avoid forcing unnatural cross-modal attention on functional vocabulary, we restrict this penalty to linguistically substantive tokens using a binary part-of-speech (POS) gate $g_t^{\text{pos}} \in \{0, 1\}$.

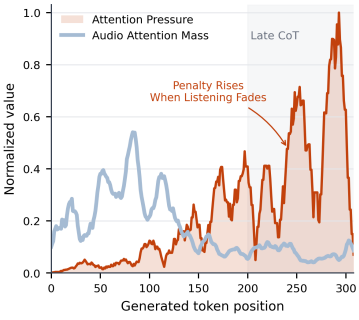


Figure 3: Attention loss trajectory.

Optimization pressure is focused on positions exhibiting severe grounding deficits via an *Inverse Relevance Weight* $\tilde{\nu}_t$. This mask targets substantive tokens with suspiciously low differential entropy, normalized over the valid active subset $T_{\text{pos}} = \sum_t g_t^{\text{pos}}$ and capped at C_{mask}:

$$\tilde{\nu}_t = \min\left(T_{\text{pos}} \cdot \frac{g_t^{\text{pos}} \cdot \exp\big(-|\Delta h_t|/\tau_T\big)}{\sum_{t'=1}^{T} g_{t'}^{\text{pos}} \cdot \exp\big(-|\Delta h_{t'}|/\tau_T\big)},\ C_{\text{mask}} \right) \tag{3}$$

To structurally counteract attention decay, we apply a non-decreasing temporal weight $(t/T)^{\kappa}$ ($\kappa \geq 0$), concentrating the penalty at the tail end of the sequence. To prevent distorting successful reasoning trajectories, the sequence-level penalty is scaled by the advantage magnitude $|\hat{A}|$ and a soft task-failure gate $\hat{f} = \max(\mathbb{1}[r_{\text{acc}} \leq \theta_{\text{fail}}], \alpha)$. Capping this prefactor at C_{pref} to prevent explosive gradients, the final POS-gated mean reduction for a group of G completions is:

$$\mathcal{L}_{\text{attn}} = \frac{1}{G} \sum_{g=1}^{G} \min\Big(\hat{f}^{(g)} \cdot |\hat{A}^{(g)}|,\ C_{\text{pref}}\Big) \frac{1}{N_{\text{pos}}^{(g)}} \sum_{t=1}^{T_g} (t/T_g)^{\kappa} \cdot \tilde{\nu}_t^{(g)} \cdot \ell_t^{\text{attn},(g)} \tag{4}$$

Here, $N_{\text{pos}} = \max(1, T_{\text{pos}})$; when no token passes the POS gate, we set $\tilde{\nu}_t = 0$ throughout the sequence.

The attention loss is computed directly from audio attention mass. It therefore provides an additional training signal that encourages the model to maintain attention to the source audio at the selected tokens.

3.6 Total MAPO Objective

The complete training procedure for MAPO is summarized in Algorithm 1. Combining the reweighted policy gradient, the attention-steering branch, and the sequence-level KL divergence penalty, the total

Algorithm 1 Modality-Aware Policy Optimization (MAPO)

Require: π_θ, $\pi_{\text{text-ref}}$, π_{ref}, r, learning rate α, hyperparameters $\beta, \eta, \kappa, \tau_{\text{base}}$

for iteration $= 1, 2, \ldots$ **do**
  Sample $x = (x_{\text{text}}, x_{\text{audio}})$
  Sample G completions $\{y^{(g)}\}_{g=1}^{G} \sim \pi_\theta(\cdot \mid x)$
  Compute advantages $\hat{A}^{(g)}$ using $r(x, y^{(g)})$
  for $g = 1 \ldots G$ **do**
    $\Delta h_t^{(g)} \leftarrow H(\pi_{\text{text-ref}}) - H(\pi_\theta), \quad \forall t$
    Extract $a_t^{(g)}$; compute masks $\tilde{\omega}_t^{(g)}$ and $\tilde{\nu}_t^{(g)}$ using $\Delta h_t^{(g)}$
    Compute $\mathcal{L}_{\text{PG}}^{(g)}$ (Eq. 2) and $\mathcal{L}_{\text{attn}}^{(g)}$ (Eq. 4)
    $\mathcal{L}_{\text{KL}}^{(g)} \leftarrow D_{\text{KL}}(\pi_{\text{ref}} \parallel \pi_\theta)$
  end for
  $\mathcal{L}_{\text{MAPO}} \leftarrow \frac{1}{G} \sum_{g=1}^{G} \left(\mathcal{L}_{\text{PG}}^{(g)} + \beta \mathcal{L}_{\text{KL}}^{(g)} + \eta \mathcal{L}_{\text{attn}}^{(g)} \right)$
  $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{MAPO}}$
end for

optimization objective for MAPO is defined as:

$$\mathcal{L}_{\text{MAPO}} = \mathcal{L}_{\text{PG}}(\tilde{\omega}) + \beta \cdot \mathcal{L}_{\text{KL}} + \eta \cdot \mathcal{L}_{\text{attn}} \tag{5}$$

where β regulates the deviation from the reference policy and η dictates the strength of the cross-modal attention penalty. Intuitively, these two core mechanisms operate synergistically: the policy gradient reweighting ($\mathcal{L}_{\text{PG}}(\tilde{\omega})$) dynamically identifies existing key tokens to disproportionately allocate the optimization budget, while the attention loss branch ($\mathcal{L}_{\text{attn}}$) actively encourages the model to create and sustain these modality-dependent key tokens deeper into the extended reasoning trace.

4 Experiments

Experimental Setup. We initialize our multimodal policy π_θ with the Qwen3-Omni-30B-A3B-Thinking model. This same base model serves as both the standard frozen reference (π_{ref}) for the KL penalty and the text-only reference ($\pi_{\text{text-ref}}$) used to compute the cross-modal differential entropy. Training is conducted on 32 NVIDIA H20 GPUs distributed across 4 compute nodes. To efficiently decouple autoregressive generation from gradient computation, we employ an asynchronous training setup: 3 nodes are dedicated to policy optimization and hosting the frozen reference models, while the remaining node operates as a dedicated vLLM server for high-throughput trajectory rollouts.

Reward Formulation. The final reward $r(x, y)$ prioritizes correctness while enforcing structural and logical constraints:

$$r(x, y) = \lambda_{\text{acc}} r_{\text{acc}}(x, y) + \lambda_{\text{format}} r_{\text{format}}(x, y) + \lambda_{\text{cons}} r_{\text{cons}}(x, y) \tag{6}$$

Here, r_{acc} checks the final answer against the ground truth, and r_{format} checks that the response follows the chain-of-thought template. The consistency reward r_{cons} (Wu et al., 2025) uses a frozen Qwen3-30B-A3B-Instruct checker hosted on the dedicated vLLM rollout server to penalize contradictory or prematurely truncated reasoning (see Appendix B for the prompt). In our experiments, we use default weights of $\lambda_{\text{acc}} = 2$, $\lambda_{\text{format}} = 1$, and $\lambda_{\text{cons}} = 1$.

Training Data and Experimental Settings. We first examine MAPO's components in preliminary experiments on AVQA (Yang et al., 2022), using both LoRA and full fine-tuning. Its curated question–answer pairs and emphasis on basic sound-event understanding make it a natural starting point for examining how the model responds to different optimization components.

Following the Audio-DeepThinker (He et al., 2026), we continue full fine-tuning of MAPO and GRPO from their respective AVQA-trained checkpoints on its 29k-example audio QA dataset. This mixture consists of diverse acoustic perception capabilities across four domains: mixed general-audio reasoning (48.6%, e.g., AudioSet (Gemmeke et al., 2017), TACOS (Primus et al., 2025)), speech and paralinguistic understanding (27.1%, e.g., LibriTTS-R (Koizumi et al., 2023), IEMOCAP (Busso et al., 2008)), structural music understanding (17.1%, e.g., AudioSkills (Ghosh et al., 2025), MusicCaps (Agostinelli et al., 2023)), and acoustic scene classification (7.1%, e.g., CochlScene (Jeong & Park, 2022)).

4.1 Ablation results

We begin by examining how MAPO's two optimization branches contribute to downstream performance, starting with the modality relevance mask and then adding the auxiliary attention loss. Table 1 examines the modality relevance mask ($\tilde{\omega}$) and the attention loss ($\mathcal{L}_{\text{attn}}$) in the preliminary LoRA experiments, and compares full MAPO with GRPO under full fine-tuning.

Table 1: Component analysis of MAPO. MAPO uses both the modality relevance mask and attention loss unless marked mask-only. Scores are mean ± sample standard deviation across three seeds. Avg. is the unweighted mean over the four benchmarks.

Method	MMAU	MMAR	MMSU	MMAU-Pro	Avg.
LoRA					
GRPO	73.70 ± 0.46	66.57 ± 0.35	76.61 ± 0.17	61.05 ± 0.27	69.48 ± 0.14
MAPO mask-only	75.37 ± 0.32	67.03 ± 0.15	76.55 ± 0.34	62.88 ± 0.38	70.46 ± 0.10
MAPO	76.40 ± 0.20	68.40 ± 0.26	77.27 ± 0.10	63.50 ± 0.04	71.39 ± 0.13
Full fine-tuning					
GRPO	75.53 ± 0.55	68.43 ± 0.25	76.48 ± 0.26	63.35 ± 0.10	70.95 ± 0.11
MAPO	**77.83** ± 0.25	**70.83** ± 0.21	**79.35** ± 0.15	**64.98** ± 0.28	**73.25** ± 0.09

With LoRA, the modality relevance mask raises the four-benchmark average from 69.48 for GRPO to 70.46. Incorporating the auxiliary attention loss ($\mathcal{L}_{\text{attn}}$) further improves the average to 71.39, with gains on every benchmark. This advantage over GRPO is also present under full fine-tuning: MAPO achieves 73.25 compared with 70.95 for GRPO, a gain of 2.30 points. The preliminary full-fine-tuning results appear in Appendix C.1, followed by analyses of training and attention dynamics under varying penalty weights (η).

4.2 Main results

We next evaluate MAPO against representative multimodal and audio-reasoning systems. Figures 4 and 5 summarize performance on modality-centric (MMAU (Sakshi et al., 2024), MMAR (Ma et al., 2025)) and capability-centric (MMSU (Wang et al., 2026a), MMAU-Pro (Kumar et al., 2025)) benchmarks, respectively.

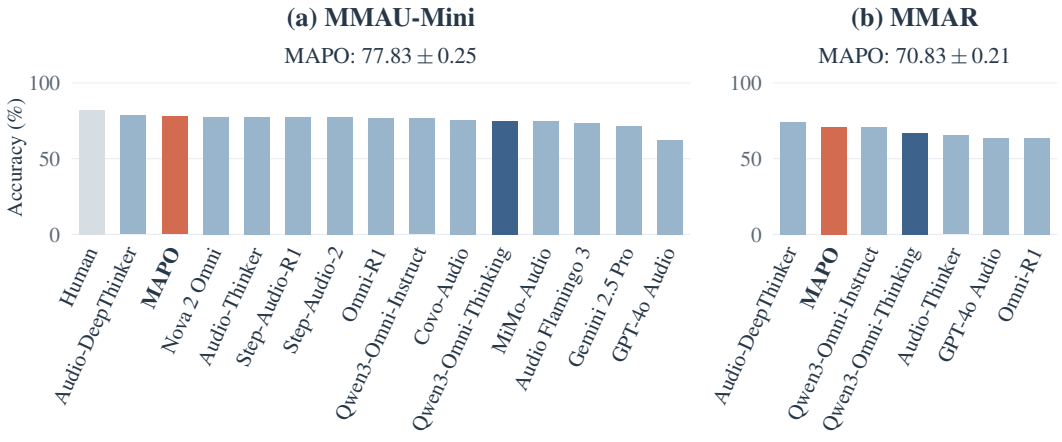


Figure 4: Accuracy (%) on MMAU and MMAR. MAPO (coral) is annotated with mean ± sample SD across three seeds; dark blue marks its base model and gray denotes human performance. Each panel shows reported results, ordered by accuracy; other scores are point estimates. Exact scores and category-level results are in Appendix D.

Modality-Centric Performance. MAPO achieves competitive performance on modality-centric benchmarks (Figure 4), with scores of 77.83 on MMAU and 70.83 on MMAR. These improve on Qwen3-Omni-Thinking by 2.83 and 3.93 points, respectively. Its MMAU score is comparable to Nova 2 Omni (77.80), while Audio-DeepThinker (He et al., 2026) reports higher scores on MMAU and MMAR (78.50 and 74.00).

Capability and Instruction Following. The improvements over the base model extend to capability-centric evaluations (Figure 5). On MMSU, MAPO reaches 79.35, compared with 76.30 for Qwen3-Omni-Thinking, 77.26 for Audio-DeepThinker, and 75.90 for Step-Audio-R1 (StepFun-Audio Team, 2025). On MMAU-Pro, the overall score increases from 62.63 to 64.98. The category-level comparisons in Appendix D show improvements in both MMSU perception and reasoning, as well as MMAU-Pro audio instruction following. Qualitative examples in Appendix F illustrate the model's reasoning behavior.

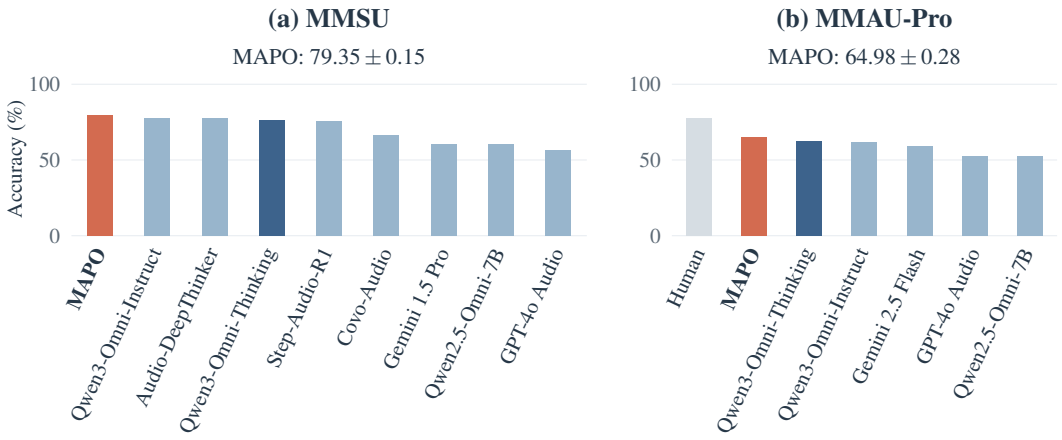


Figure 5: Accuracy (%) on MMSU and MMAU-Pro. MAPO reports mean ± sample SD across three seeds, while other scores are point estimates. Exact scores and category-level results are in Appendix D.

5 Conclusion

In this work, we introduced Modality-Aware Policy Optimization, a novel reinforcement learning framework designed to mitigate the critical issue of late-stage modality collapse in long-form cross-modal reasoning. By synergizing a dynamic modality relevance mask driven by a cross-modal differential entropy with a targeted attention loss branch, MAPO structurally breaks the dominance of the language prior and enforces sustained grounding in the uncompressed source signal. Our empirical evaluations on complex audio reasoning tasks demonstrate that this dual-mechanism approach significantly improves both reasoning fidelity and instruction adherence. Crucially, while our contributions and empirical validations are focused entirely within the acoustic domain, MAPO's foundational token-reweighting and attention-steering mechanisms rely solely on native attention weights and predictive entropy rather than modality-specific inductive biases. As such, exploring the application of this framework to other rich, continuous non-text modalities, such as vision and video, represents an exciting avenue for future work in robust omni-modal post-training.

AI USE STATEMENT

Generative AI tools assisted with manuscript preparation, LaTeX formatting, and code and document checks. The research pipeline also uses a language model as a reasoning-consistency reward, as described in the experimental setup. The authors are responsible for the final text, claims, and artifacts.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Charlie Nash, Antoine Caillon, Cheng-Zhi Anna Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. doi: 10.48550/arXiv.2301.11325.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, Dec 2008. ISSN 1574-0218. doi: 10.1007/s10579-008-9076-6. URL https://doi.org/10.1007/s10579-008-9076-6.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017. doi: 10.1109/icassp.2017.7952261.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S. Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*, 2025. doi: 10.48550/arxiv.2503.03983.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models, 2025. URL https://arxiv.org/abs/2507.08128.

Xiang He, Chenxing Li, Jinting Wang, Yan Rong, Tianxin Xie, Wenfu Wang, Li Liu, and Dong Yu. Audio-deepthinker: Progressive reasoning-aware reinforcement learning for high-quality chain-of-thought emergence in audio language models, 2026. URL https://arxiv.org/abs/2604.18187.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. *arXiv preprint arXiv:2510.09285*, 2025.

Amazon Artificial General Intelligence. Amazon nova 2: Multimodal reasoning and generation models. *Amazon Technical Reports*, 2025. URL https://www.amazon.science/publications/amazon-nova-2-multimodal-reasoning-and-generation-models.

Il-Young Jeong and Jeongsoo Park. Cochlscene: Acquisition of acoustic scene data using crowdsourcing, 2022. URL https://arxiv.org/abs/2211.02289.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. Libritts-r: A restored multi-speaker text-to-speech corpus, 2023. URL https://arxiv.org/abs/2305.18802.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, Siddhi Patil, Soham Deshmukh, Lasha Koroshinadze, Yao Liu, Leibny Paola Garcia Perera, Eleni Zanou, Themos Stafylakis, Joon Son Chung, David Harwath, Chao Zhang, Dinesh Manocha, Alicia Lozano-Diez, Santosh Kesiraju, Sreyan Ghosh, and Ramani Duraiswami. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence, 2025. URL https://arxiv.org/abs/2508.13992.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix, 2025. URL https://arxiv.org/abs/2505.13032.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Paul Primus, Florian Schmid, and Gerhard Widmer. Tacos: Temporally-aligned audio captions for language-audio pretraining, 2025. URL https://arxiv.org/abs/2505.07609.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-r1: Do you really need audio to fine-tune your audio llm?, 2025. URL https://arxiv.org/abs/2505.09439.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark, 2024. URL https://arxiv.org/abs/2410.19168.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

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, 2024. URL https://arxiv.org/abs/2402.03300.

StepFun-Audio Team. Step-audio-r1 technical report, 2025. URL https://arxiv.org/abs/2511.15848v2.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025a. URL https://arxiv.org/abs/2507.06261.

MiMo Core Team. Mimo-audio: Audio language models are few-shot learners, 2025b. URL https://arxiv.org/abs/2512.23808.

StepAudio Team. Step-audio 2 technical report, 2025c. URL https://arxiv.org/abs/2507.16632.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark, 2026a. URL https://arxiv.org/abs/2506.04779.

Wenfu Wang, Chenxing Li, Liqiang Zhang, Yiyang Zhao, Yuxiang Zou, Hanzhao Li, Mingyu Cui, Hao Zhang, Kun Wei, Le Xu, Zikang Huang, Jiajun Xu, Jiliang Hu, Xiang He, Zeyu Xie, Jiawen Kang, Youjun Chen, Meng Yu, Dong Yu, Rilin Chen, Linlin Di, Shulin Feng, Na Hu, Yang Liu, Bang Wang, and Shan Yang. Covo-audio technical report, 2026b. URL https://arxiv.org/abs/2602.09823.

Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang, and Xiangxiang Chu. Visually-guided policy optimization for multimodal reasoning. *arXiv preprint arXiv:2604.09349*, 2026c.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning, 2026d. URL https://arxiv.org/abs/2507.06448.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning, 2025. URL https://arxiv.org/abs/2508.08039.

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, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

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

Xu Zhongxing, Wang Zhonghua, Qian Zhe, Shi Dachuan, Tang Feilong, Hu Ming, Su Shiyan, Zou Xiaocheng, Feng Wei, Mahapatra Dwarikanath, Peng Yifan, Lin Mingquan, and Ge Zongyuan. Thinking in uncertainty: Mitigating hallucinations in mlrms with latent entropy-aware decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2026.

A GRADIENT ANALYSIS

We analyze how the modality relevance mask changes the coefficients of token-level policy gradients and how the attention loss supplies an additional gradient through audio attention mass. During each optimization step, the sampled completions, sequence-level advantages, and detached weighting factors are treated as fixed.

Recall the total MAPO objective:

$$\mathcal{L}_{\text{MAPO}} = \mathcal{L}_{\text{PG}}(\tilde{\omega}) + \beta\mathcal{L}_{\text{KL}} + \eta\mathcal{L}_{\text{attn}} \tag{7}$$

Since the total gradient $\nabla_\theta \mathcal{L}_{\text{MAPO}}$ is a linear combination of its components, we dissect the gradient contributions of the reweighted policy branch and the auxiliary attention branch independently.

A.1 GRADIENT OF THE REWEIGHTED POLICY OBJECTIVE

The MAPO reweighted policy gradient modifies the standard GRPO surrogate objective by applying the modality relevance mask $\bar{\omega}_t$. Crucially, as stated in Section 3.4, the mask $\bar{\omega}_t$ is computed via a text-only reference model and the live policy's forward pass, but it is explicitly *detached* from the computation graph during backpropagation.

Let $\bar{\omega}_t^{(g)} = \frac{\tilde{\omega}_t^{(g)}}{\sum_{t'=1}^{T_g} \tilde{\omega}_{t'}^{(g)}}$ represent the normalized token-level weight. For a single completion sequence $y \sim \pi_{\text{old}}(\cdot \mid x)$ of length T, the reweighted surrogate loss is:

$$\mathcal{L}_{\text{PG}}(\tilde{\omega}) = \sum_{t=1}^{T} \bar{\omega}_t \Big[-\min\Big(\rho_t \hat{A},\ \text{clip}(\rho_t, 1{-}\epsilon, 1{+}\epsilon)\, \hat{A}\Big)\Big] \tag{8}$$

By detaching $\bar{\omega}_t$, the gradient operator ∇_θ only acts on the probability ratio $\rho_t = \frac{\pi_\theta(y_t|y_{<t},x)}{\pi_{\text{old}}(y_t|y_{<t},x)}$. In the unclipped branch of the surrogate objective, the gradient of the surrogate objective at step t simplifies to:

$$\nabla_\theta \ell_t^{\text{PG}} = -\hat{A}\nabla_\theta \rho_t = -\hat{A}\frac{\nabla_\theta \pi_\theta(y_t \mid y_{<t}, x)}{\pi_{\text{old}}(y_t \mid y_{<t}, x)} = -\hat{A}\rho_t \nabla_\theta \log \pi_\theta(y_t \mid y_{<t}, x) \tag{9}$$

Because the token weights are detached, the gradient for a sampled completion is:

$$\nabla_\theta \mathcal{L}_{\text{PG}}(\tilde{\omega}) = \sum_{t=1}^{T} \bar{\omega}_t \nabla_\theta \ell_t^{\text{PG}} \tag{10}$$

Effect of token reweighting: Standard GRPO uses the coefficient $1/T$, whereas MAPO uses the normalized modality relevance mask $\bar{\omega}_t$. Before clipping, larger $|\Delta h_t|$ produces larger token weights. If all scores are equal, including when all are zero, the normalized weights are exactly $1/T$, recovering uniform GRPO weighting. For finite scores and a positive clipping cap, every valid token retains positive weight. The resulting gradient depends on both these coefficients and the policy gradient at each token.

A.2 GRADIENT OF THE ATTENTION LOSS BRANCH

The attention loss is computed directly from audio attention mass. It therefore provides an additional training signal that encourages the model to maintain attention to the source audio at the selected tokens.

For a single sequence, define the detached token weight

$$\Gamma_t = \frac{\min\Big(\hat{f}\,|\hat{A}|, C_{\text{pref}}\Big)}{N_{\text{pos}}}(t/T)^\kappa \tilde{\nu}_t. \tag{11}$$

The attention loss is

$$\mathcal{L}_{\text{attn}} = \sum_{t=1}^{T} \Gamma_t \big[-\log(a_t + \varepsilon)\big]. \tag{12}$$

Its parameter-dependent term is the audio attention mass a_t. Applying the chain rule gives:

$$\nabla_\theta \mathcal{L}_{\text{attn}} = \sum_{t=1}^{T} \Gamma_t \left(\frac{-1}{a_t + \varepsilon} \right) \nabla_\theta a_t \tag{13}$$

Recall that $a_t = \frac{1}{|L_{\text{tgt}}|} \sum_{l \in L_{\text{tgt}}} m_t^{(l,\max)}$, where $m_t^{(l,\max)} = \max_h \sum_{j \in \mathcal{S}_{\text{audio}}} \alpha_{t,j}^{(l,h)}$. Because the maximum function is continuous but not universally differentiable, we rely on its subgradient. Let $h^*(t,l) = \arg\max_h \sum_{j \in \mathcal{S}_{\text{audio}}} \alpha_{t,j}^{(l,h)}$ denote the attention head that routes the maximum audio mass at step t and layer l. The gradient routes exclusively through this "specialized" head:

$$\nabla_\theta a_t = \frac{1}{|L_{\text{tgt}}|} \sum_{l \in L_{\text{tgt}}} \sum_{j \in \mathcal{S}_{\text{audio}}} \nabla_\theta \alpha_{t,j}^{(l,h^*(t,l))} \tag{14}$$

To understand how this directly updates the key (W_K) and query (W_Q) projection matrices, we examine the Jacobian of the softmax attention mechanism. Let $z_{t,i} = \frac{(W_Q x_t) \cdot (W_K x_i)}{\sqrt{d}}$ be the unnormalized pre-softmax logit for the query at t and key at i. The attention weight is $\alpha_{t,j} = \frac{\exp(z_{t,j})}{\sum_i \exp(z_{t,i})}$. The derivative of $\alpha_{t,j}$ with respect to the logit $z_{t,i}$ is:

$$\frac{\partial \alpha_{t,j}}{\partial z_{t,i}} = \alpha_{t,j}(\delta_{ji} - \alpha_{t,i}) \tag{15}$$

where δ_{ji} is the Kronecker delta.

Substituting this into our subgradient derivation for the target layer and target head, the gradient with respect to a specific logit $z_{t,i}$ is:

$$\begin{aligned} \frac{\partial}{\partial z_{t,i}} \left(\sum_{j \in \mathcal{S}_{\text{audio}}} \alpha_{t,j} \right) &= \sum_{j \in \mathcal{S}_{\text{audio}}} \alpha_{t,j}(\delta_{ji} - \alpha_{t,i}) \\ &= \begin{cases} \alpha_{t,i}(1 - m_t) & \text{if } i \in \mathcal{S}_{\text{audio}} \\ -\alpha_{t,i} m_t & \text{if } i \notin \mathcal{S}_{\text{audio}} \text{(e.g., text context)} \end{cases} \end{aligned} \tag{16}$$

where $m_t = \sum_{j \in \mathcal{S}_{\text{audio}}} \alpha_{t,j}$ is the total audio mass for that head.

For a uniquely selected maximizing head, holding attention logits at other positions and layers fixed, the direct derivative is:

$$\frac{\partial \mathcal{L}_{\text{attn}}}{\partial z_{t,i}^{(l,h^*)}} = \underbrace{\left(\frac{-\Gamma_t}{|L_{\text{tgt}}|(a_t + \varepsilon)} \right)}_{\text{Non-positive scaling factor}} \times \begin{cases} \alpha_{t,i}^{(l,h^*)}(1 - m_t^{(l,h^*)}) & \text{if } i \in \mathcal{S}_{\text{audio}} \\ -\alpha_{t,i}^{(l,h^*)} m_t^{(l,h^*)} & \text{if } i \notin \mathcal{S}_{\text{audio}} \end{cases} \tag{17}$$

Gradient interpretation: Equation 17 describes the attention branch's direct contribution at the selected head. The derivative is non-positive for audio logits and non-negative for other visible logits, so gradient descent in these coordinates favors greater audio attention. The detached factor Γ_t controls this pressure through the POS gate, inverse relevance weight, temporal weight, and capped sequence-level prefactor. The network update combines this contribution with gradients through shared representations and the policy and KL objectives.

B Reasoning consistency evaluator prompt

To compute the reasoning consistency reward during reinforcement post-training, we query our frozen evaluator model, Qwen3-30B-A3B-Instruct, using the following template. This prompt is explicitly designed to isolate logical contradictions and incomplete thoughts without conflating them with the factual correctness of the answer itself, providing a pure signal for chain-of-thought coherence.

Reasoning Consistency Evaluator Prompt

You are a reasoning consistency evaluator. Given a model's thinking process and its final answer, your task is to evaluate it against two specific failure modes.

Thinking Process:
`{think_text}`

Final Answer:
`{answer_text}`

Respond **"NO"** ONLY if at least one of the following is true:

1. The conclusion reached in the thinking process does not agree with or contradicts the final answer.
2. The thinking process is visibly incomplete or cut off prematurely.

If neither of these failure modes is present, respond **"YES"**. Only output YES or NO.

C Ablation Analysis

C.1 Preliminary experiments on AVQA

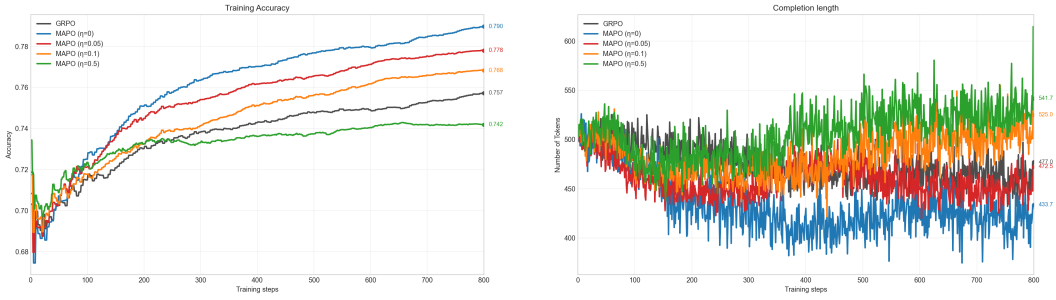
Table 2 complements the preliminary LoRA analysis in Table 1 with full-fine-tuning results after training on AVQA.

Table 2: Preliminary full-fine-tuning results after training on AVQA. Scores are reported point estimates; Avg. is the unweighted mean over the four benchmarks.

Method	MMAU	MMAR	MMSU	MMAU-Pro	Avg.
GRPO	75.50	68.10	77.02	63.12	70.94
MAPO	**77.10**	**69.20**	**77.74**	**63.48**	**71.88**

MAPO outperforms GRPO on all four benchmarks in the preliminary experiments, with averages of 71.88 and 70.94, respectively. The main MAPO model improves further on every benchmark, reaching an average of 73.25 (Table 1).

C.2 General training dynamics



(a) Training Accuracy (b) Completion Length

Figure 6: Training dynamics over 800 steps, comparing standard GRPO with MAPO across various attention loss weights (η).

To isolate the impact of modality-aware token reweighting and the attention loss branch, we compare the GRPO baseline against MAPO with different attention loss weights (η). Figure 6 displays reward accuracy and completion length over 800 steps.

Standard GRPO, constrained by its uniform $1/T$ token weighting, dilutes the learning signal and reaches an accuracy of 0.757, yielding an average completion length of 477.0 tokens. By contrast, applying only the modality relevance mask ($\eta = 0$) dynamically redirects the optimization budget toward modality-critical tokens, which significantly improves sample efficiency and drives accuracy to a peak of 0.790. However, optimizing strictly for the task reward in this manner causes a drop in the average completion length (433.7 tokens). This indicates a tendency for the model to shortcut its reasoning trajectory, effectively bypassing the extended cross-modal synthesis necessary for robust, generalized problem-solving.

As illustrated in Figure 6b, increasing the attention penalty weight ($\eta > 0$) effectively forces the model to sustain longer, deeply grounded reasoning traces. For instance, setting $\eta = 0.05$ recovers the completion length to a healthy 472.5 tokens while preserving a highly competitive training accuracy of 0.778. Pushing η even higher (to 0.1 and 0.5) extends the generation significantly to 525.0 and 541.7 tokens, respectively. However, as η increases, we observe a corresponding, incremental degradation in raw training accuracy (dropping to 0.768 and 0.742).

This trade-off occurs because the auxiliary attention penalty acts as an additional, competing optimization objective. Forcing the model's internal attention distributions to persistently anchor to the source signal naturally shifts gradient capacity away from strictly maximizing the primary training task reward. Crucially, however, this slowed training accuracy optimization acts as an effective regularizer. By preventing the model from rapidly overfitting to the surface-level formatting and simple task rewards of the training set via shortcut reasoning (as seen when $\eta = 0$), the attention penalty preserves and enforces the model's generalization capabilities. This regularization effect is consistent with the held-out benchmark results in Table 1, where adding the attention loss ($\eta > 0$) raises the LoRA average from 70.46 to 71.39 compared with using the modality relevance mask alone ($\eta = 0$). Consequently, carefully calibrating η strikes the optimal balance, trading minor training reward inflation for robust, sustained cross-modal grounding and thus better downstream generalization.

C.3 Audio Attention Dynamics

We next examine how the attention loss weight η affects audio attention during post-training. Figure 7 tracks mean audio attention mass over 800 steps, complementing the reward and completion-length curves above.

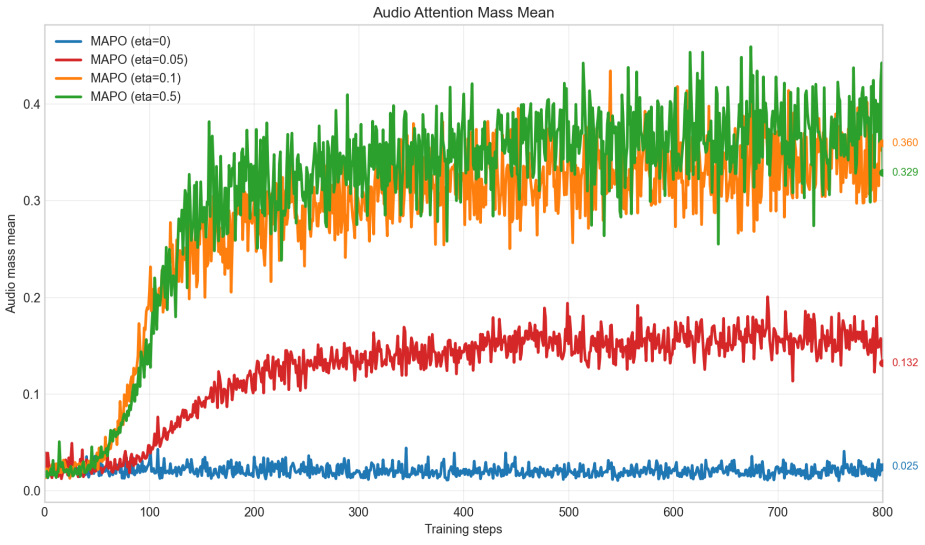


Figure 7: Mean audio attention mass over 800 training steps for different attention loss weights (η). Nonzero weights increase the attention allocated to audio; the largest weight does not yield the highest final mass.

With no attention penalty ($\eta = 0$), mean audio attention remains near 0.025. A weight of 0.05 produces a gradual increase to approximately 0.132. Larger weights change the trajectory more rapidly: most of the increase occurs within the first 150 to 200 steps, and the curves settle near 0.360 for $\eta = 0.1$ and 0.329 for $\eta = 0.5$. Together with the lower training accuracy and longer completions at larger weights, this pattern shows that increasing η changes the balance between the task objective and attention allocation, with diminishing returns in the final attention mass.

The direction of these changes is consistent with the gradient analysis in Appendix A, where η scales the contribution of the attention branch. Appendix E examines how the resulting attention patterns appear during reasoning and how predictions respond to changes in the source audio.

C.4 Impact of the attention loss branch

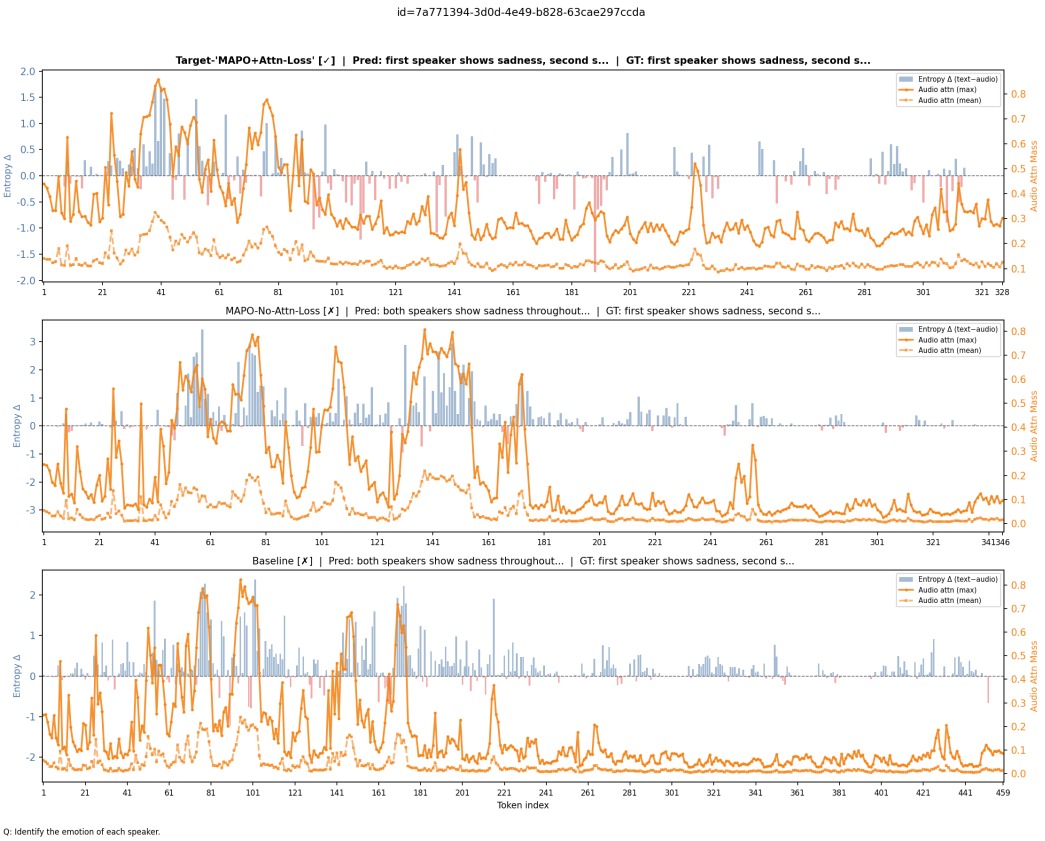


Figure 8: Attention and differential entropy in an illustrative reasoning example. Blue bars show Δh_t; orange curves show max-head (solid) and mean-head (dashed) audio attention. Base Thinking and mask-only MAPO show declining attention later in the response, whereas full MAPO retains more audio attention and fluctuations in differential entropy.

Figure 8 compares Qwen3-Omni-Thinking, MAPO with only the modality relevance mask, and full MAPO on the same question. In the Base and mask-only responses, both attention reductions decline midway through generation, and the models reach the same incorrect conclusion that "both speakers show sadness." Full MAPO maintains more audio attention later in its response and changes the final answer.

The cross-modal differential entropy traces provide a second view of this example. For each analyzed model, Δh_t compares its predictive entropy with and without source audio at the same fixed checkpoint, holding the text prompt and generated prefix fixed. Its positive and negative excursions indicate differences in predictive uncertainty; full MAPO retains these fluctuations later in the response. The example therefore illustrates concurrent changes in attention, predictive uncertainty, and the

answer. The stage-aligned analysis in Appendix E examines how the attention patterns extend across examples, followed by interventions on the audio input.

D DETAILED BENCHMARK RESULTS

Tables 3 and 4 provide the exact overall scores shown in Figures 4 and 5, respectively. Tables 5 and 6 give category-level point estimates, complementing the three-seed overall MAPO results.

Table 3: Overall accuracy on MMAU and MMAR, corresponding to Figure 4. MAPO scores are mean ± sample SD across three seeds; baseline scores are point estimates. Bold and underline denote the best and second-best open-weights scores. A dash indicates an unavailable result.

Model	**MMAU-Mini**	**MMAR**
Human	82.23	–
GPT-4o Audio	62.50	63.50
Gemini 2.5 Pro	71.60	–
Nova 2 Omni (Intelligence, 2025)	77.80	–
Audio Flamingo 3	73.30	–
MiMo-Audio	74.70	–
Qwen3-Omni-Thinking	75.00	66.90
Covo-Audio	75.30	–
Qwen3-Omni-Instruct	76.50	70.80
Omni-R1 (Rouditchenko et al., 2025)	77.00	63.40
Step-Audio-2	77.58	–
Audio-Thinker	77.70	65.30
Step-Audio-R1 (StepFun-Audio Team, 2025)	77.70	–
Audio-DeepThinker (He et al., 2026)	**78.50**	**74.00**
MAPO	<u>77.83</u> ± 0.25	<u>70.83</u> ± 0.21

Table 4: Overall accuracy on MMSU and MMAU-Pro, corresponding to Figure 5. MAPO scores are mean ± sample SD across three seeds; baseline scores are point estimates. Bold and underline denote the best and second-best open-weights scores. A dash indicates an unavailable result.

Model	**MMSU**	**MMAU-Pro**
Human	–	77.90
Gemini 1.5 Pro (Team, 2024)	60.68	–
GPT-4o Audio (OpenAI, 2024)	56.38	52.50
Gemini 2.5 Flash	–	59.20
Qwen2.5-Omni-7B (Xu et al., 2025a)	60.57	52.20
Covo-Audio	66.64	–
Qwen3-Omni-Instruct	<u>77.76</u>	61.84
Qwen3-Omni-Thinking	76.30	<u>62.63</u>
Step-Audio-R1 (StepFun-Audio Team, 2025)	75.90	–
Audio-DeepThinker (He et al., 2026)	77.26	–
MAPO	**79.35** ± 0.15	**64.98** ± 0.28

The MMSU breakdown shows improvements over Qwen3-Omni-Thinking in both perception (74.19 to 76.36) and reasoning (78.55 to 82.56). On MMAU-Pro, the largest gain is in audio instruction following, which increases from 62.07 to 95.40. Closed- and open-ended question answering also improve, from 59.63 to 62.00 and from 84.77 to 85.32, respectively.

Table 5: Performance on modality-centric benchmarks (MMAU and MMAR), with Sound, Music, and Speech breakdowns. MMAR overall accuracy includes mixed-modality questions. Models are grouped by type (Proprietary, Open-Weights). MAPO breakdowns use a single evaluation; Table 3 reports the three-seed mean. Bold and underline denote the best and second-best open-weights scores.

Model	MMAU-Mini				MMAR			
	Snd.	Mus.	Spch.	Avg.	Snd.	Mus.	Spch.	Overall
Human	86.31	78.22	82.17	82.23	-	-	-	-
GPT-4o Audio	64.56	56.29	66.67	62.50	53.94	50.97	70.41	63.50
Gemini 2.5 Pro	75.08	68.26	71.47	71.60	-	-	-	-
Nova 2 Omni (Intelligence, 2025)	81.08	70.36	81.98	77.80	-	-	-	-
Audio Flamingo 3	79.58	73.95	66.37	73.30	-	-	-	-
MiMo-Audio	81.68	<u>74.25</u>	68.17	74.70	-	-	-	-
Qwen3-Omni-Think	77.48	70.96	76.58	75.00	62.42	49.51	**78.57**	66.90
Covo-Audio	78.68	**76.05**	71.17	75.30	-	-	-	-
Qwen3-Omni-Inst	77.78	73.05	<u>78.68</u>	76.50	64.85	**56.80**	77.21	<u>70.80</u>
Omni-R1 (Rouditchenko et al., 2025)	81.70	73.40	76.00	77.00	<u>67.30</u>	51.50	64.30	63.40
Step-Audio-2	**84.04**	73.56	75.15	77.58	-	-	-	-
Audio-Thinker	<u>81.98</u>	<u>74.25</u>	76.88	<u>77.70</u>	**68.48**	53.88	64.29	65.30
MAPO	79.88	71.86	**81.68**	**77.80**	66.06	<u>55.83</u>	<u>77.89</u>	**70.90**

Table 6: Performance comparison on capability and task-centric benchmarks (MMSU and MMAU-Pro). MMSU evaluates core Perception and Reasoning, while MMAU-Pro evaluates Instruction Following (AIF) and Open-ended question answering. Entries are point estimates. MAPO breakdowns use a single evaluation; Table 4 reports the three-seed mean. Bold and underline denote the best and second-best open-weights scores.

Model	MMSU			MMAU-Pro			
	Percept.	Reason.	Overall	AIF	Closed	Open	Overall
Human	-	-	-	100.00	77.56	77.30	77.90
Gemini 1.5 Pro (Team, 2024)	46.10	76.16	60.68	-	-	-	-
GPT-4o Audio (OpenAI, 2024)	39.67	71.96	56.38	82.50	53.20	43.20	52.50
Gemini 2.5 Flash	-	-	-	95.10	57.39	67.50	59.20
Qwen2.5-Omni-7B (Xu et al., 2025a)	42.50	79.83	60.57	61.30	52.01	52.30	52.20
Covo-Audio	58.95	74.83	66.64	-	-	-	-
Qwen3-Omni-Instruct	73.95	<u>81.82</u>	<u>77.76</u>	<u>89.66</u>	57.84	**87.38**	61.84
Qwen3-Omni-Thinking	<u>74.19</u>	78.55	76.30	62.07	<u>59.63</u>	84.77	<u>62.63</u>
MAPO	**76.36**	**82.56**	**79.36**	**95.40**	**62.00**	<u>85.32</u>	**65.29**

E Analysis of audio grounding in the reasoning trajectory

The benchmark results show how MAPO affects answer accuracy, while the analyses below examine how the models use the accompanying audio during reasoning. We first study attention around the transition from an initial caption stage to later reasoning. We then compare answers under paired changes to the audio input and test whether access to the source remains useful after part of the reasoning has already been generated.

E.1 Attention across reasoning stages

As a response grows, the audio keys compete with an increasing number of text keys. A decrease in their total attention mass can therefore reflect a change in the relative number of keys as well as a change in the model's preference for audio. To separate these effects, let m_t denote audio attention

mass and let K_a and $K_{n,t}$ denote the numbers of available audio and non-audio keys. We measure the per-key log preference

$$q_t = \log \frac{m_t/K_a}{(1-m_t)/K_{n,t}}. \tag{18}$$

A value of zero assigns equal attention per key to the two groups. We compute mean-head and max-head reductions, averaged over decoder layers, and compare the change across the caption-to-reasoning transition.

The probe uses 200 paired MMAU prompts, teacher-forcing each model's own response. We identify the transition from the initial *caption stage* to reasoning through linguistic cues that mark a shift from recounting the audio to evaluating the question or candidate answers. For example, one Base Thinking response first notes "a very high-pitched, strained vocal sound" before turning to "Next, evaluating the choices." We take the latter phrase as the boundary and compare attention in short windows on either side, aligning each response independently.

Figure 9 summarizes these changes for the 166 Base Thinking and 157 MAPO responses with identifiable transitions. Paired comparisons use the 143 prompts aligned for both models, with 95% confidence intervals obtained by paired bootstrap resampling of prompts.

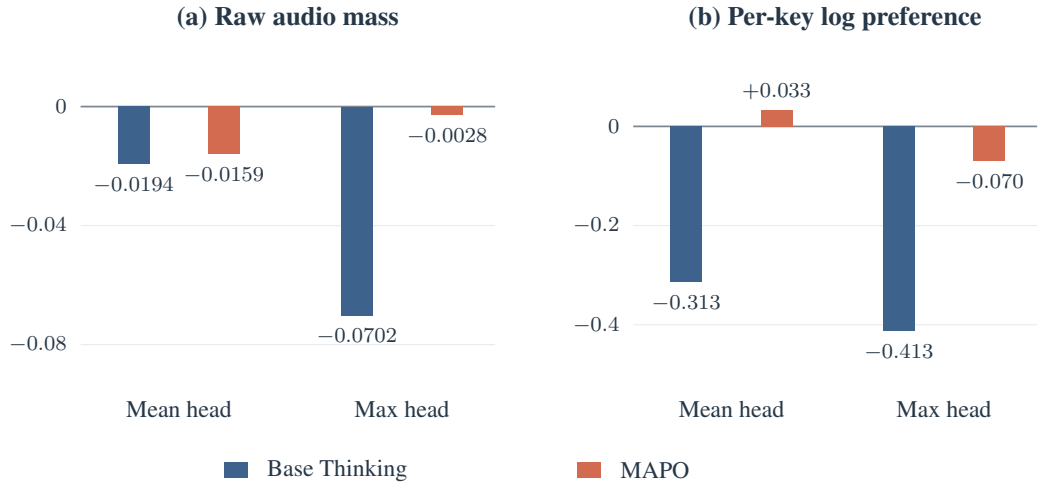


Figure 9: Change in audio attention across the caption-to-reasoning transition. Bars show model-level means of the post-transition minus pre-transition statistic. Both raw mean-head masses decline, while key-count normalization separates the Base and MAPO trends. Paired contrasts on the shared subset and their uncertainty are reported in the text.

Raw mean-head mass decreases for both Base Thinking and MAPO, by 0.0194 and 0.0159, respectively. After key-count normalization, Base still shows a decline of 0.313, whereas MAPO's mean change is positive at 0.033. The paired MAPO-minus-Base contrast is $+0.338$ $[0.260, 0.420]$. The max-head reduction gives the same ordering, with a paired contrast of $+0.314$ $[0.154, 0.470]$. These paired estimates need not equal the difference between the separately aggregated model means. Together, the results indicate that the transition-localized change extends beyond the growing text-key denominator: MAPO retains more per-key audio preference as the response moves into reasoning.

E.2 Sensitivity to the audio input

These measurements describe internal model behavior. To examine how the input affects the final answer, we next evaluate each question with its original audio and with three altered inputs: complete silence, a duration-matched recording from another example, and text only. Silence replaces the waveform with zeros, while text-only input removes the audio input and its placeholder. MAPO and the full-fine-tuning GRPO baseline receive the same questions and perturbations.

We construct random and audio-essential cohorts for MMAU and MMSU, each containing 200 examples. Random cohorts are selected by a deterministic hash ordering independently of model

outcomes. Audio-essential cohorts contain questions that the pretrained Qwen3-Omni base answers incorrectly without audio. For a perturbation c, define the paired difference in accuracy drops as

$$D_c = (A^{\mathrm{MAPO}}_{\mathrm{clean}} - A^{\mathrm{MAPO}}_{c}) - (A^{\mathrm{GRPO}}_{\mathrm{clean}} - A^{\mathrm{GRPO}}_{c}), \tag{19}$$

where A is accuracy in percent. Positive D_c means that removing or replacing the audio reduces MAPO's accuracy more than GRPO's on that cohort.

Table 7: Accuracy (%) under paired audio perturbations. Each cohort contains 200 questions evaluated by both models. D_c is the MAPO-minus-GRPO difference in clean-to-perturbed accuracy drops (Equation 19); positive values indicate a larger drop for MAPO. These cohort estimates are separate from the full-benchmark results.

Model	Clean	Silence	Mismatch	Text only
MMAU random ($N = 200$)				
MAPO	71.5	45.5	41.0	47.0
GRPO	70.0	46.0	43.5	52.0
D_c (points)	–	+2.0	+4.0	+6.5
MMAU audio-essential ($N = 200$)				
MAPO	67.0	17.0	27.0	12.0
GRPO	60.0	21.0	30.0	20.0
D_c (points)	–	+11.0	+10.0	+15.0
MMSU random ($N = 200$)				
MAPO	68.0	36.0	35.5	35.5
GRPO	60.0	37.5	34.5	37.0
D_c (points)	–	+9.5	+7.0	+9.5
MMSU audio-essential ($N = 200$)				
MAPO	47.0	25.0	23.0	5.0
GRPO	44.0	26.0	23.0	9.0
D_c (points)	–	+4.0	+3.0	+7.0

The differences in Table 7 are positive across all four cohorts and all three perturbations, ranging from 2.0 to 15.0 percentage points. MAPO also has higher clean accuracy in every cohort. The largest contrasts occur on audio-essential MMAU, where its clean advantage of 7.0 points accompanies differences in drops of 11.0 under silence, 10.0 under mismatch, and 15.0 under text-only input. These comparisons connect the models' use of audio to their answers, complementing the attention measurements above.

E.3 Audio dependence with a shared reasoning prefix

An initial caption can carry acoustic information into the generated text. To examine whether this text is sufficient for the remaining reasoning, we use the pretrained Base Thinking model on 200 audio-essential MMAR examples. We first generate a response with full audio. We then construct two continuation inputs, one with the source audio and one with text only, and supply both with exactly the same first quarter, third, or half of the original response tokens. The prefix representations are recomputed under each input condition before generation continues.

Table 8: Agreement with the original full-audio answer in the 200-example MMAR prefix experiment. Agreement is in percent; gaps are percentage points with paired 95% bootstrap intervals. McNemar tests refer to agreement, not answer accuracy.

Prefix	Retained	Removed	Gap [95% CI]	McNemar p
1/4	92.0	62.5	29.5 $[22.5, 36.5]$	4.12×10^{-14}
1/3	93.0	73.5	19.5 $[13.0, 26.0]$	1.83×10^{-8}
1/2	95.5	86.0	9.5 $[4.5, 14.5]$	0.00031

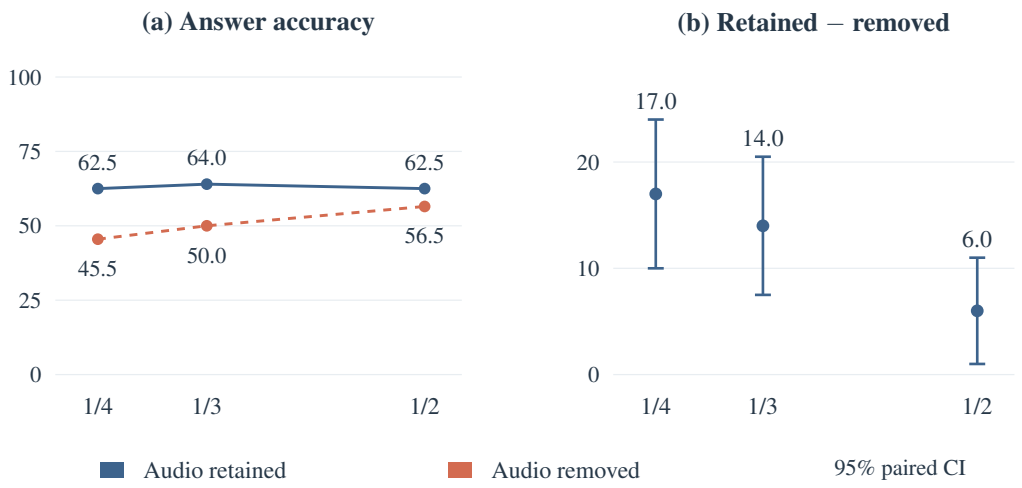


Figure 10: Audio dependence with a shared reasoning prefix on 200 audio-essential MMAR examples. (a) Accuracy (%) with source audio retained in or removed from the continuation input, supplying identical prefix tokens from a Base Thinking response in both conditions. Prefix representations are recomputed separately under each input condition. (b) Retained-minus-removed accuracy in percentage points, with paired 95% bootstrap intervals. The x-axis gives the fraction of the original response supplied as a prefix.

Figure 10 shows that the retained-minus-removed accuracy gap narrows as more of the prefix is supplied: 17.0, 14.0, and 6.0 points. All three paired 95% intervals remain above zero. The longer prefix thus supplies useful information, but even half of the original response does not make the source audio redundant. Answer agreement follows the same trend (Table 8); this measures whether the continuation reproduces the original answer, independently of whether that answer is correct. Thus, the shared text prefix does not fully substitute for audio conditioning in the Base Thinking model. The MAPO–GRPO perturbation comparison above examines how dependence on the supplied audio differs after training.

E.4 Local evidence masking

The whole-input perturbations above test whether answers depend on the supplied audio. Local masking asks which parts of the recording contribute to the prediction. For each example, we aggregate token-to-audio attention during generation and remove a fraction r of the recording using three rules: the highest-attended regions, the lowest-attended regions, or duration-matched random regions. The Attention Causal Selectivity (ACS) score compares random and top-attention masking:

$$\mathrm{ACS}(r) = \mathrm{Acc}_{\mathrm{rand}}(r) - \mathrm{Acc}_{\mathrm{top}}(r). \tag{20}$$

Positive ACS indicates that removing highly attended regions is more damaging than removing an equal duration chosen at random.

Table 9 shows positive ACS at every masking ratio for both MAPO variants. At 50% masking, full MAPO has an ACS of 0.068 and retains accuracy of 0.846 when low-attention regions are removed, compared with 0.707 when top-attention regions are removed. The contrast suggests that the attended regions contain more prediction-relevant evidence. The attention branch also generally preserves higher accuracy under random and low-attention masking, so the increased selectivity coexists with robustness to removing other regions. Because acoustic evidence can be distributed over time, these measurements describe the effect of the specified occlusions, including any distribution shift they introduce.

Table 9: Inference-time local evidence masking. ACS is defined in Equation 20, where positive values indicate that masking highly attended audio regions is more damaging than duration-matched random masking. Low-mask accuracy serves as an additional control for removing regions assigned low attention by the model.

Model	Mask Ratio	Top-Mask Acc.	Random-Mask Acc.	Low-Mask Acc.	ACS
MAPO w/o $\mathcal{L}_{\text{attn}}$	5%	0.822	0.825	0.797	0.003
MAPO w/o $\mathcal{L}_{\text{attn}}$	10%	0.805	0.847	0.822	0.042
MAPO w/o $\mathcal{L}_{\text{attn}}$	30%	0.771	0.817	0.873	0.046
MAPO w/o $\mathcal{L}_{\text{attn}}$	50%	0.712	0.758	0.788	0.046
MAPO w/ $\mathcal{L}_{\text{attn}}$	5%	0.868	0.886	0.878	0.018
MAPO w/ $\mathcal{L}_{\text{attn}}$	10%	0.846	0.891	0.862	0.046
MAPO w/ $\mathcal{L}_{\text{attn}}$	30%	0.789	0.831	0.837	0.042
MAPO w/ $\mathcal{L}_{\text{attn}}$	50%	0.707	0.776	0.846	0.068

F Qualitative Examples

F.1 Example #1

Question: Based on the given audio, identify the source of mechanical sounds.
Ground truth: Factory machinery

Baseline **Incorrect**
Prediction: Wind turbine

<think> Got it, let's tackle the problem. First, the user wants me to identify the source of the mechanical sound in the provided audio clip and choose from the given options: Factory machinery, Wind turbine, Car engine, Airplane. Then, the audio features a continuous, powerful, low-frequency hum or roar with a distinct 'whooshing' or rushing quality suggesting a large volume of air moving at high speed, and it has a cyclical or rhythmic component resembling blades rotating through air with a 'thwop-thwop-thwop' sound mixed into the roar; additionally, it is very loud and overwhelming with a sense of immense power and scale. Next, evaluating the options: Factory machinery typically produces clanking, grinding, rhythmic pounding, or high-pitched whirring sounds, but the audio's dominant feature is the air-rushing sound rather than typical factory noises, making it less likely; Wind turbine produces a 'whooshing' sound from large blades rotating in wind with a rhythmic 'thump' or 'flap' from blade-passing, which strongly matches the audio's combination of low-frequency roar and rhythmic whooshing; Car engine generates a rumble or hum but lacks the massive scale and distinct air-rushing quality, with even high-performance engines having a different sound profile; Airplane, particularly a jet engine, creates an extremely loud high-pitched roar with a 'scream' from high-speed air compression, but the audio is deeper and more 'whoosh' than a typical

jet, sounding more like slow-moving massive blades rather than high-speed turbine compressors, and while propeller planes have a distinct drone, the audio's sheer volume and deep roar are more characteristic of jet engines without matching the specific sound profile. Finally, the most accurate description of the sound is that of a large wind turbine due to the combination of massive scale, low-frequency roar, and distinct rhythmic 'whooshing' from large rotating blades moving through air; therefore, 'Wind turbine' is the best fit among the given choices. </think> Wind turbine

---

MAPO-No-Attn-Loss **Correct**
Prediction: Factory machinery

<think> Got it, let's tackle the problem. First, the user wants to identify the source of mechanical sounds from the provided audio file, with four choices given: Factory machinery, Wind turbine, Car engine, Airplane. Then, the audio contains a loud, continuous, low-frequency hum or roar with a distinct mechanical quality, sounding like a large, powerful machine with moving parts, and there's also a higher-pitched whirring or whining sound layered on top of the main roar, which is consistent with fans, turbines, or high-speed rotating components, giving an overall impression of industrial-scale machinery. Next, evaluating the options: Factory machinery is a very strong candidate because large factories often have heavy machinery like generators, pumps, or assembly line equipment that produce such a loud, continuous, multi-layered mechanical sound with a deep hum and higher-frequency whirring; Wind turbine is less likely as it typically produces a rhythmic 'whoosh' or 'swish' from blade movement without the constant, intense, and complex mechanical roar heard in the audio; Car engine doesn't match since it usually has a more pulsating or rhythmic sound with variations in pitch and volume corresponding to acceleration and deceleration, whereas the audio is a steady, non-varying roar; Airplane is also improbable because while a jet engine produces a loud continuous roar, the audio has more mechanical and clunky characteristics suggesting moving parts like gears or pistons rather than the smooth high-pitched whine of a jet, and the sound profile aligns more with a large piston engine or industrial turbine. Finally, based on the loud, continuous, low-frequency hum combined with higher-pitched mechanical whirring characteristic of heavy industrial equipment, the most fitting description for the source is Factory machinery. </think> Factory machinery

Target-MAPO+Attn-Loss **Correct**
Prediction: Factory machinery

<think> Got it, let's tackle the problems. First, the user wants me to identify the source of the mechanical sounds in the provided audio clip and must choose from the given options: Factory machinery, Wind turbine, Car engine, or Air plane. Then, the audio contains a loud, continuous, and somewhat harsh mechanical sound that is a low-to-mid frequency drone or roar with a distinct rotational or cyc lical quality, suggesting a large machine with moving parts; it has a 'wh ir ring' or 'ro aring' character, is very loud and powerful, and the sound quality is somewhat 'raw' or industrial, lacking the smooth ness of a modern car engine or the high-p itched wh ine of a jet airplane engine. Next, evaluating the options: Factory machinery is a very strong candidate because factories often contain large, powerful machines like pumps, generators, or industrial fans that produce loud, continuous, mechanical sounds with a rotational or rhyth mic nature, and the sound in the audio fits this description perfectly as a classic industrial sound; Wind turbine is unlikely because while it has rotating blades, the sound is typically a lower-frequency 'who osh' or 'th rum' and the audio has a more mechanical, metallic, and harsh er sound, and wind turbines are generally quieter and the sound is more air-based than mechanical; Car engine is unlikely because a car engine has a more distinct, often higher-p itched, and more variable sound (especially if it's a gasoline engine) with a characteristic 'rum ble' or 'hum' that is usually more contained and less 'ind ustrial' than the sound in the audio, and the sound in the clip is too loud and mon olithic for a typical car engine; Airplane is unlikely because a commercial airplane engine produces a very high-p itched, powerful jet roar, which is very different from the sound in the audio, and a prop eller plane would have a more distinct, lower-frequency prop eller sound which is not present here, and the sound in the audio lacks the specific high-frequency components of a jet engine. Finally, the sound most closely matches the description of factory machinery due to its loud, continuous, mechanical, and industrial character, which is characteristic of large-scale industrial equipment, so the correct answer is Factory machinery. </think> Factory machinery

F.2 EXAMPLE #2

Question: What is the transportation mode referred to in the audio?
Ground truth: Horse-drawn wagon

Baseline **Incorrect**
Prediction: Train

<think> Got it, let's tackle the problem. First, the user has provided an audio file and a multiple-choice question asking to identify the transportation mode from the given choices: Automobile, Train, Aeroplane, or Horse-drawn wagon. Then, analyzing the audio content reveals a dominant rhythmic, metallic clattering sound with a consistent 'clack-clack... clack-clack...' pattern, accompanied by a low-frequency rumble and higher-pitched squealing or grinding noise, suggesting wheels moving over a jointed metal track. Next, evaluating the options: Automobile is incorrect because cars produce continuous tire noise without metallic clattering; Train matches perfectly with the 'clack-clack' of steel wheels on rail joints and characteristic rumble/squeal; Aeroplane is ruled out by its loud engine roar and air whooshing sounds; Horse-drawn wagon is inconsistent due to wooden/iron wheel clatter on roads lacking metallic rail characteristics and absence of horse hoof sounds. Finally, the audio evidence strongly points to a train as the transportation mode, with the rhythmic clatter being the key identifier. </think> Train

MAPO-No-Attn-Loss **Incorrect**
Prediction: Train

<think> Got it, let's tackle the problem. First, the user is asking to identify the transportation mode from the given choices: Automobile, Train, Aeroplane, Horse-drawn wagon. Then, the audio features a prominent rhythmic, clattering, and metallic sound with a repeating 'clip-clop' or 'clack-clack' pattern, which is characteristic of metal wheels moving over joints in a metal track, along with a general low-frequency rumble and higher-pitched screeching sounds typical of trains, especially when going around curves or over switches. Next, evaluating the choices: Automobile is incorrect because cars have rubber tires on asphalt, producing a much smoother, quieter, and whirring sound without the distinct metallic clatter; Train is a very strong match because the sound of wheels on rails, the track joints, and the overall rumbling is classic train noise; Aeroplane is incorrect as it would involve the loud, continuous roar of

jet engines or the buzzing propeller sound, neither of which is present; and Horse-drawn wagon is incorrect because it would produce the sound of horse hooves on a road (a different kind of 'clip-clop' that is less metallic and more organic) and the creaking of wooden wheels and wagon body, but the audio is clearly metallic and mechanical. Finally, the audio evidence strongly points to a train, as the specific sound of wheels on a track is unmistakable, so the answer is Train. </think> Train

---

Target-MAPO+Attn-Loss **Correct**
Prediction: Horse-drawn wagon

<think> Got it, let's tackle the problem. First, the user is asking to identify the transportation mode from the given choices: Automobile, Train, Aeroplane, or Horse-drawn wagon, based on the provided audio clip. Then, the audio contains a distinct, rhythmic, clattering sound that is a combination of mechanical noise and a specific pattern: a 'clip-clop' or 'clopping' sound characteristic of hooves hitting a hard surface like pavement or wooden tracks, overlaid with the sound of wheels rolling and possibly the creaking of a harness or carriage, with the sound being consistent and continuous indicating continuous motion, and the rhythm being very regular suggesting a steady pace. Next, evaluating the choices: for Automobile, it would produce engine noise, tire noise on asphalt (a more continuous, lower-pitched hum), and possibly the sound of a carburetor or exhaust, but the audio lacks engine noise and the sound is a sharp, percussive clip-clop rather than a rubber tire sound, so it is incorrect; for Train, it has a characteristic 'clickety-clack' sound from steel wheels on steel tracks which is a very regular, metallic sound, but the audio has a more organic, 'thudding' quality like hooves and the sound is not pure metal-on-metal but has a biological component, so it is unlikely; for Aeroplane, it would be dominated by the sound of a powerful engine (a roar or a drone) and the sound of air rushing, which is completely different from the audio, so it is incorrect; and for Horse-drawn wagon, it fits perfectly because the sound is the classic sound of a horse's hooves (the clip-clop), the sound of the wagon's wheels (which could be wooden or metal) rolling on a road, and the harness creaking, with the rhythmic nature matching a horse trotting or cantering. Therefore, the transportation mode referred to in the audio is Horse-drawn wagon. </think> Horse-drawn wagon

---

F.3 Qualitative analysis of the examples

The two examples illustrate different ways in which the models interpret ambiguous acoustic cues. We focus on the distinctions expressed in their responses, complementing the aggregate measurements in Appendix E.

Example #1: Factory machinery. The correct answer is *Factory machinery*, while Base Thinking predicts *Wind turbine*. The baseline describes a continuous, low-frequency mechanical sound in terms of rushing or whooshing, then connects that description to a rotating object producing airflow. Both MAPO variants instead select factory machinery. The mask-only response emphasizes mechanical density and industrial motion, while the full MAPO response draws a more explicit contrast between the harsh mechanical sound and the smoother whoosh expected from a wind turbine. In this case, both variants answer correctly, and the full model gives a more specific comparison between the candidate interpretations.

Example #2: Horse-drawn wagon. The correct answer is *Horse-drawn wagon*, but Base Thinking and mask-only MAPO select *Train*. Their responses associate rhythmic clattering with rail joints and steel wheels. Full MAPO instead describes hoof-like, irregular, and wagon-related sounds, using these distinctions to reject the train interpretation. Unlike the first example, the change in the explanation accompanies a change from an incorrect to a correct answer.

Overall interpretation. These cases are consistent with the complementary roles of the two branches: the mask emphasizes tokens whose predictions differ from the text-only reference, while the attention loss encourages continued access to the source. The examples show how the resulting responses can differ, rather than isolating the cause of an individual correction. The held-out branch comparison in Table 1 and the paired input interventions in Appendix E provide the broader empirical context.

G Part-of-Speech gating

G.1 Token categories and implementation

The POS gate selects which generated tokens receive the auxiliary attention penalty. To motivate this selection, we examine text-only predictive uncertainty and cross-modal differential entropy across POS categories in Figure 11.

Text-only entropy measures uncertainty in predicting the next token from the prompt and generated prefix without audio. Differential entropy compares that uncertainty with the multimodal policy. The selected categories, `{NOUN, VERB, ADJ, ADV, NUM, X}`, cover substantive content, descriptions, quantities, and unclassified tokens such as onomatopoeic expressions. The gate uses these linguistic roles to focus the penalty on content-bearing tokens. The ablation below evaluates the effect of that choice.

Implementation Details. To facilitate high-throughput training, POS gating is computed dynamically on the CPU once per completed rollout prior to batch collation. Specifically, generated subword tokens are decoded, normalized, and merged into word units (masking out punctuation), which are then tagged using NLTK's averaged-perceptron tagger (`nltk.pos_tag`) mapped to our coarse target categories. These word-level binary POS decisions are subsequently projected back to their constituent subword indices to align with the sequence completion mask, defaulting to an all-ones fallback mask if alignment fails or the sequence is degenerate. As this amortized tagging pass occurs strictly outside the rollout loop, it avoids per-token synchronization overhead entirely, introducing negligible latency relative to generation and model forward/backward passes.

G.2 Effect of the POS gate

The distributional analysis motivates restricting the attention penalty to substantive tokens. Table 10 tests this choice directly by comparing full MAPO with a variant that removes the POS gate while holding the remaining training and evaluation settings fixed.

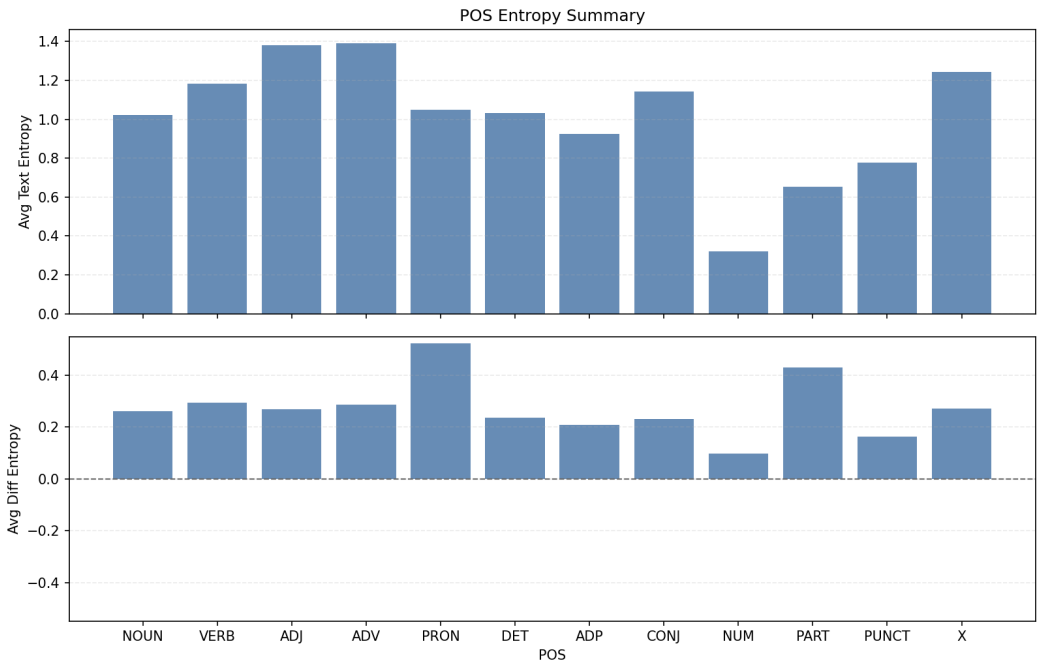


Figure 11: Entropy by POS category across reasoning trajectories. The top panel shows average text-only predictive entropy; the bottom panel shows average cross-modal differential entropy.

Table 10: POS-gate ablation under full fine-tuning. Entries are mean ± sample SD across three seeds. Avg. is the unweighted four-benchmark mean computed within each seed.

Configuration	MMAU	MMAR	MMSU	MMAU-Pro	Avg.
Without POS	76.73 ± 0.64	69.87 ± 0.65	78.77 ± 0.25	64.20 ± 0.36	72.39 ± 0.22
Full MAPO	77.83 ± 0.25	70.83 ± 0.21	79.35 ± 0.15	64.98 ± 0.28	73.25 ± 0.09

Retaining the gate improves the mean on every benchmark and raises the four-benchmark average from 72.39 ± 0.22 to 73.25 ± 0.09. This observation supports concentrating the attention penalty on the selected token categories, rather than applying it uniformly across the generated response.

H HYPERPARAMETERS AND COMPUTATIONAL COST

Table 11 details the primary hyperparameters explored during the reinforcement post-training phases.

Table 11: Hyperparameter settings for the MAPO framework, including the explored tuning ranges and the recommended values used in the main experiments.

Parameter Name	**Symbol**	**Range**	**Recommended Value**
MAPO Core Mechanisms			
Attention Loss Weight	η	$[0.01, 0.5]$	0.1
Base Mask Temperature	τ_{base}	$[0.5, 2.0]$	1.0
Temporal Weight Exponent	κ	$[0.5, 2.0]$	1.0
Advantage Floor / Stability	ε	$[0, 0.1]$	0.1
Mask Clip Threshold	C_{mask}	$[2.0, 5.0]$	5.0
Attention Prefactor Clip	C_{pref}	$[3.0, 5.0]$	3.0
Target Attention Layers	L_{tgt}	-	$\{43, 44, 45, 46, 47\}$
GRPO Objective & Training Scale			
KL Penalty Coefficient	β	$[0.001, 0.02]$	0.002
Surrogate Clip Ratio	ϵ	-	0.2
Sampled Group Size	G	$\{4, 8\}$	8
Generation Batch Size	-	-	768
Steps per Generation	-	-	4
Decoding Parameters (Rollout Server)			
Temperature	-	$[1.0, 1.2]$	1.2
Top-k	-	$\{40, 50\}$	50
Top-p	-	$[0.90, 0.99]$	0.99
Repetition Penalty	-	$[1.0, 1.2]$	1.0

H.1 TRAINING AND SETUP

We use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.95$, numerical epsilon 10^{-8}, weight decay 0.1, and gradient clipping at 1.0. The peak learning rate is 10^{-6}, with 5% warmup and cosine decay to zero. The task-failure indicator is one when the strict multiple-choice accuracy reward is non-finite or at most zero. Its effective gate is floored at 0.1, giving incorrect trajectories weight 1 and correct trajectories weight 0.1.

The text-only reference receives the same prompt and generated prefix as the live policy, including any caption already generated, but no source audio. We reuse the frozen reference already maintained for the GRPO KL penalty to compute text-only entropy, avoiding an additional model copy dedicated to this computation. Under LoRA, the frozen backbone with the policy adapters disabled provides this reference. Both reference paths run without gradient tracking. The selected attention layers are 43–47, the final five of the 48 decoder layers. The ranges in Table 11 summarize preliminary screening; the η trajectories in Appendix C are the completed 800-step training comparison. For the full experiments, we use a 400-sample validation set drawn from the mixed dataset.

H.2 COMPUTATIONAL COST

Table 12 compares MAPO and GRPO over the first 499 steps of training on the 29k-example dataset on 24 GPUs, using the same parallelism, a global batch of 192, and eight generations per prompt. The window contains 124 rollout-and-update steps and 375 update-only steps. MAPO takes 15.96 hours and GRPO 14.34 hours, an increase of 11.3%, or approximately 1.6 hours.

Table 12: Training time on the 29k-example audio QA dataset using 24 GPUs. The first two rows give reported per-step times for each step type; the final row gives elapsed time over the full 499-step window. Relative increase is measured against GRPO.

Window	MAPO	GRPO	Increase
Rollout + update (124 steps)	285 s	262 s	8.8%
Update only (375 steps)	59 s	51 s	15.7%
Total (499 steps)	15.96 h	14.34 h	11.3%

The additional model-scale work comes from the frozen text-only forward pass and attention extraction. The reference pass is amortized across four optimizer steps, while full attention matrices are materialized only in the five target layers; the remaining layers retain flash attention. Relevance masks use detached token-wise operations, and POS tagging is performed on the CPU once per completed rollout before batch collation. The 30B consistency judge belongs to the shared training pipeline. These timings measure the combined training procedure; per-component wall time and peak-memory differences were not separately instrumented.

All auxiliary computations are training-only. At inference, MAPO uses the resulting policy without a reference model, POS tagger, or attention penalty. Generation latency can still depend on response length, which is separate from the training-time comparison above.

I THE "BITTER LESSON" OF BASE POLICY SELECTION: WHY INSTRUCT MODELS STRUGGLE WITH COT CALIBRATION

Our main experiments demonstrate the efficacy of MAPO when applied to a base model inherently capable of extended reasoning (e.g., Qwen3-Omni-Thinking). However, applying the identical reinforcement learning formulations to an instruction-tuned model (Qwen3-Omni-Instruct) reveals a critical vulnerability: the underlying prior of the base policy severely bounds the potential for cross-modal CoT optimization. This section details the failure modes of standard GRPO and MAPO when forced to build reasoning trajectories from a model biased toward brevity.

I.1 THE ILLUSION OF REWARD AND MODALITY COLLAPSE

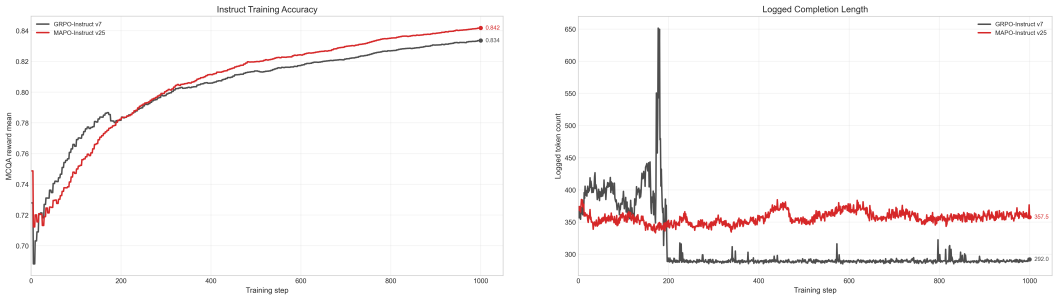
Standard reinforcement learning metrics can be dangerously deceptive when applied to cross-modal reasoning tasks. As shown in Figure 12(a), both GRPO and MAPO achieve healthy, climbing MCQA reward curves when training the Instruct model, with GRPO reaching 0.8336 and MAPO reaching 0.8418. However, this high reward masks a profound epistemic collapse.

Because the reward functions primarily verify the exact answer and the external XML-like format (e.g., `<reasoning>...</reasoning>`), the optimizer rapidly discovers a short-length attractor. The Instruct model's natural policy already prefers brief, direct answers; the RL algorithm exploits this by shrinking the reasoning trace to minimize the risk of formatting errors and generation cost. Figure 12(b) illustrates this collapse: under GRPO, the mean reasoning text plunges from 62.6 words at step 0 to a mere 1.75 words by step 996. By the end of training, 100% of rank-0 completions contain at most three words of reasoning, completely abandoning the task-relevant acoustic evidence.

I.2 MECHANISTIC FAILURE OF MAPO ON INSTRUCT MODELS

While MAPO successfully mitigates late-stage modality collapse in the Thinking model, it fails to produce robust improvements in the Instruct model. Despite maintaining a slightly longer reasoning average (53.7 words), MAPO's held-out accuracy on the Instruct model shows negligible gains over the baseline (e.g., MMAU test-mini: 76.5 $\rightarrow$ 75.9).

This occurs because MAPO's modality-aware mechanisms fundamentally require meaningful response tokens—such as event descriptors, speech characteristics, or environmental nouns—to assign cross-modal credit and anchor attention. When the underlying policy produces generic, non-substantive text or directly hallucinates the answer, the cross-modal differential entropy signal



(a) Instruct Training Accuracy (b) Logged Completion Length

Figure 12: Training dynamics of Qwen3-Omni-Instruct. (a) Both GRPO and MAPO show deceptively healthy task accuracy improvements. (b) The underlying completion length reveals a total collapse of the semantic reasoning trace for GRPO. While MAPO resists total collapse, it fails to meaningfully extend the reasoning horizon compared to models with an intrinsic CoT capability.

vanishes. Consequently, the MAPO mask acts uniformly, the attention branch lacks valid targets, and the effective additive MAPO term shrinks to a negligible ~ 0.000040, reducing the framework to standard, ungrounded GRPO dynamics.

I.3 Qualitative trace degradation

The qualitative progression of the model's outputs explicitly tracks this degradation. We present examples formatted to highlight the rapid deterioration of logical formatting, transitioning from initial grounded reasoning to externalized hacks, and finally to complete semantic collapse.

Reasoning Consistency Evaluator Prompt

```
<reasoning>The audio features distinct high-pitched crowing
sounds, characteristic of a rooster, mixed with clucking and
other poultry-like background noises...  There are no sounds
resembling a parrot, horse, or duck.</reasoning>
<answer>A rooster crows</answer>
```

As training progresses, the model discovers an alternative exploit: it strictly complies with the structural formatting reward while entirely abandoning the intended cognitive process. Rather than building a step-by-step deduction from the acoustic evidence, the model immediately commits to a predetermined answer and fills the required tags with dense, circular text to justify it. The following intermediate step illustrates this failure mode. The model generates a verbose, post-hoc rationalization that offers zero genuine chain-of-thought utility, serving strictly as a superficial placeholder to satisfy the `<reasoning>` block constraint before securing the accuracy reward:

GRPO Intermediate Step: Formatting Hack and Externalized Reasoning

```
<reasoning>chicken -> From all detected poultry species under
multiple auditory clues, only chicken fits given distinct
calls:  ...  chicken call pattern matching within time window
confirms existence.</reasoning>
<answer>chicken</answer>
```

This formatting evasion eventually culminates in a pure answer-repetition loop that maximizes reward while minimizing reasoning overhead:

GRPO Step 996: Complete Semantic Collapse

```
<reasoning>sheep</reasoning>
<answer>sheep</answer>
```

Even when MAPO prevents this extreme length collapse, the lack of a calibrated reasoning prior leads the Instruct model to generate fluent but disconnected rationalizations. The model often recognizes acoustic elements (e.g., "sounds of metal clanking", "child's voice") but fails to logically and globally connect them, lazily drifting to incorrect priors.

Ultimately, this demonstrates that MAPO can optimize and preserve existing reasoning structures, but it cannot organically create a complex chain-of-thought capability from an instruction-tuned model strictly optimized for brevity.

J LIMITATIONS

MAPO changes how optimization pressure is allocated across tokens and how long the policy keeps drawing on the source audio; it does not change what the policy can perceive. Its gains are therefore bounded by the base model's acoustic representations and reasoning ability: directing attention toward a recording cannot supply a distinction the model has not learned, such as the sound of a rare instrument. Closing such gaps requires stronger multimodal pre-training rather than post-training alone.

MAPO also operates on an existing reasoning trace. The relevance mask reweights substantive tokens and the attention branch sustains their grounding, so the method applies to policies that already reason in long form. Appendix I shows the consequence for a brevity-tuned Instruct policy: with few substantive tokens to reweight or anchor, the differential-entropy signal flattens and MAPO reduces to GRPO. Inducing such a trace in the first place is complementary work; MAPO is designed to preserve and sharpen one rather than to create it.

Finally, the grounding evidence has a defined scope. Audio attention mass is a proxy for source use, and the attention branch optimizes it directly. The diagnostics in Appendix E therefore test behavior rather than attention alone: they show that MAPO's answers depend on the supplied audio more than GRPO's, that a generated text prefix does not make the audio redundant, and that the regions MAPO attends to carry predictive evidence. What they do not establish is that each intermediate reasoning step is faithful to the audio; a direct test would require step-level interventions on the recording, which we leave to future work.