WavAlign: Enhancing Intelligence and Expressiveness in Spoken Dialogue Models via Adaptive Hybrid Post-Training

Yifu Chen[1]* **Shengpeng Ji**[1]* **Qian Chen**[2]* **Tianle Liang**[1]
Yangzhuo Li[1] **Ziqing Wang**[3] **Wen Wang**[2] **Jingyu Lu**[1] **Haoxiao Wang**[1] **Xueyi Pu**[1]
Fan Zhuo[1] **Zhou Zhao**[1]†

[1] Zhejiang University [2] Tongyi Fun Team, Alibaba Group [3] Beijing University of Technology
22551267@zju.edu.cn, zhaozhou@zju.edu.cn

Abstract

End-to-end spoken dialogue models have garnered significant attention because they offer a higher potential ceiling in expressiveness and perceptual ability than cascaded systems. However, the intelligence and expressiveness of current open-source spoken dialogue models often remain below expectations. Motivated by the success of online reinforcement learning (RL) in other domains, one might attempt to directly apply preference optimization to spoken dialogue models, yet this transfer is non-trivial. We analyze these obstacles from the perspectives of reward modeling and rollout sampling, focusing on how sparse preference supervision interacts with dense speech generation under shared-parameter updates. Based on the analysis, we propose a modality-aware adaptive post-training recipe that makes RL practical for spoken dialogue: it constrains preference updates to the semantic channel and improves acoustic behavior via explicit anchoring, while dynamically regulating their mixture from rollout statistics to avoid unreliable preference gradients. We evaluate the method across multiple spoken dialogue benchmarks and representative architectures, and observe consistent improvements in semantic quality and speech expressiveness.Our page could be found at https://github.com/MM-Speech/WavAlign

1 Introduction

Spoken dialogue models (Xu et al., 2025a; Ding et al., 2025; Wu et al., 2025b; Fang et al., 2024; Ji et al., 2024b; Chen et al., 2025b) are reshaping human–computer interaction by enabling natural and accessible speech-based interfaces. End-to-end spoken dialogue models directly operate on speech signals and unify speech understanding and generation within a single backbone, allowing joint modeling of semantic content and paralinguistic attributes (Ji et al., 2024c; Li et al., 2026). In principle, this paradigm can reduce the error propagation and information loss of cascaded pipelines, while supporting tighter integration between high-level reasoning and fine-grained acoustic expressiveness.

In practice, however, current open-source end-to-end systems still do not consistently surpass strong cascaded baselines, and their semantic capability, naturalness, and expressiveness all leave substantial room for improvement. This gap highlights an open challenge: how to improve semantic dialogue quality and speech naturalness/expressiveness simultaneously within a single end-to-end model, without sacrificing one for the other.

Motivated by the success of Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF) in text and vision (Lee et al., 2023; Guo et al., 2025; Shen et al., 2025), a natural approach is to apply reinforcement-learning-based preference optimization to end-to-end spoken dialogue. Our empirical findings show that a straightforward, sequence-level preference objective over mixed text–speech outputs is often unreliable for broad, simultaneous gains: semantic preference objectives can improve, yet speech quality frequently degrades, exhibiting acoustic drift and reduced naturalness. Our analysis attributes this instability to an optimization property of omni-modal sequences: preference signals couple weakly across modalities, while the effective gradient energy is highly imbalanced, text gradients dominate shared-parameter updates, and dense speech tokens receive comparatively weak, high-variance supervision. As a result, updates that are beneficial for semantic behavior can inadvertently perturb the delicate acoustic distributions that govern natural prosody and timbre.

Reward modeling further complicates acoustic optimization. Unlike semantic correctness in text, acoustic expressiveness lacks clean, reliable scalar

*These authors contributed equally.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 2421–2444

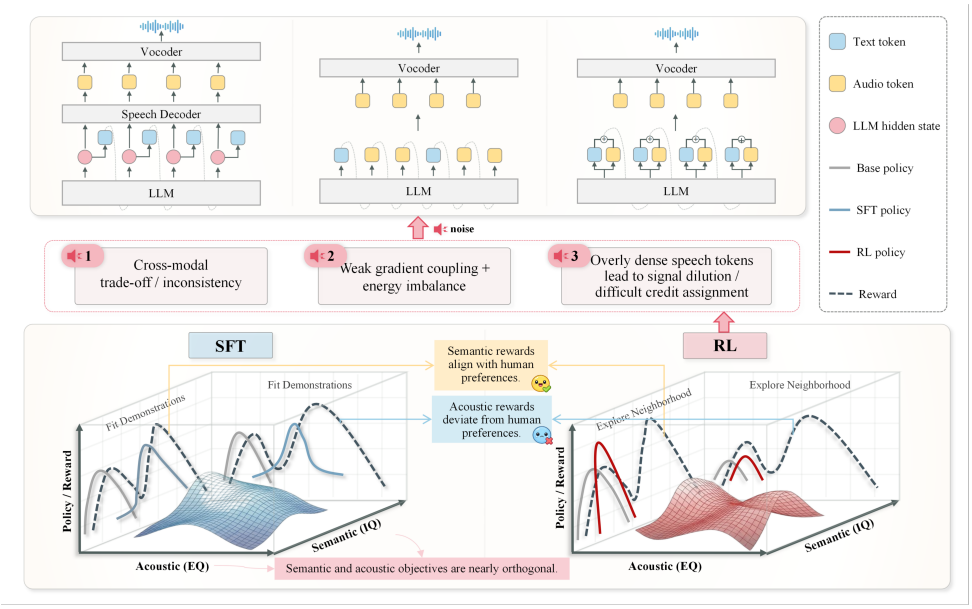

Figure 1: Motivation and failure mode of unified RL for end to end spoken dialogue models.

reward signals: rewards are often noisy, underspecified, and entangled with artifacts. When such sparse signals are distributed over long speech token sequences, credit assignment becomes ill-conditioned, and reward hacking can produce speech that scores well while sounding unnatural. These issues are amplified for weaker base models, where high-quality rollouts are rare and preference updates become poorly grounded.

These observations motivate an objective design that separates optimization roles across modalities. Supervised Fine-Tuning (SFT) is effective for constructing and maintaining acoustic feasibility and naturalness, whereas preference optimization is more reliable for semantic refinement, where reward signals are typically more consistent and easier to judge than acoustic expressiveness. Based on this principle, we propose a single-stage adaptive hybrid post-training framework that harmonizes intelligence and expressiveness in one loop: we apply preference optimization only to text tokens to improve semantic behavior, while using SFT as a distribution anchor, which stabilizes the speech token distribution in particular. To mitigate unreliable updates caused by low-quality or low-discriminability rollouts, we further introduce a dynamic gating mechanism that adjusts the balance between supervised and preference-based updates according to rollout validity and training-signal reliability, committing to preference updates only when samples are informative.

Our contributions are summarized as follows:

- We identify and characterize key failure modes of unified sequence-level preference optimization for mixed text–speech outputs, including weak cross-modal coupling, gradient-energy imbalance, and noisy acoustic rewards.
- We propose a single-stage adaptive hybrid post-training scheme that applies preference optimization to text tokens while anchoring speech tokens with SFT, coupled with a rollout-reliability gating mechanism for stable updates.
- Experiments across architectures and benchmarks show consistent gains in both semantic quality and acoustic expressiveness.

2 Related Works

Reinforcement Learning in Spoken Dialogue Models. RL is increasingly used for end-to-end spoken dialogue models, yet prior work typically targets *either* semantic quality (IQ) *or* expressiveness and naturalness (EQ), and joint optimization remains elusive. Many studies report that optimizing the *full* mixed text–audio token sequence can cause cross-modal instability and text–speech misalignment, prompting decoupled designs such as blocking audio-token gradients (Huang et al., 2025) or adopting text-only objectives (Wu et al., 2025a). In parallel, EQ-oriented methods rely on reward modeling and preference data for controllable paralinguistic behavior (Yang et al., 2025; Zhang et al.,

2024a; Gao et al., 2025; Lu et al., 2026; Ji et al., 2025), but can be brittle to reward hacking via spurious acoustic cues (Wang et al., 2025). This IQ/EQ separation also motivates modular alternatives that optimize reasoning and speech generation separately (Xu et al., 2025b; Zheng et al., 2025). Overall, a unified solution is still missing due to fragmented objectives and persistent cross-modal instability.

Single-Stage Hybrid Post-Training. Beyond the conventional two-stage recipe, recent work explores *single-stage* hybrid post-training that mixes SFT and RL within one loop to improve stability, sample efficiency, and capability growth. Methods modulate this trade-off via entropy-aware uncertainty (SRFT (Fu et al., 2025)), unified feedback continuums (UFT (Liu et al., 2025)), or dynamic auxiliary terms (CHORD (Zhang et al., 2025)). Others incorporate reasoning traces with importance sampling (LUFFY (Yan et al., 2025)) or explicitly optimize synergy and forgetting (ReLIFT (Ma et al., 2025), BRIDGE (Chen et al., 2025a), MIFO (Yuan et al., 2025)). However, these homogeneous text strategies ill-fit end-to-end spoken dialogue, where scalar preferences are less informative for heterogeneous speech tokens. We therefore propose a modality-aware hybridization principle with a lightweight adaptive controller to regulate hybrid strength based on rollout quality.

3 Methodology

3.1 Preliminaries

3.1.1 Spoken Dialogue Model

We study spoken dialogue models that generate both *text* and *speech* given an input context x. Different architectures realize the generation process differently: (i) **Interleaving** generates a single interleaved token stream, (ii) **Parallel** generates text and speech streams with coupled states, and (iii) **Thinker–Talker** factorizes generation into a "thinking" stage and a "speaking" stage. To remain architecture-agnostic, we represent the outputs as two token sequences: a text sequence $\mathbf{y}^T = (y_1^T, \ldots, y_L^T)$ and a speech sequence $\mathbf{y}^S = (y_1^S, \ldots, y_M^S)$. For each emitted token, c_i^T and c_j^S denote its conditioning context under the chosen architecture (e.g., previous tokens in an interleaved stream, cross-stream hidden states, or a preceding-stage output).

Token-type partition of log-likelihood. The model defines a joint conditional distribution $P_\theta(\mathbf{y}^T, \mathbf{y}^S \mid x)$. Regardless of the internal dependency pattern, the log-likelihood can be partitioned by token types:

$$\begin{aligned}\log P_\theta(\mathbf{y}^T,&\mathbf{y}^S \mid x) \\ &= \sum_{i=1}^{L} \log P_\theta(y_i^T \mid c_i^T) \\ &\quad + \sum_{j=1}^{M} \log P_\theta(y_j^S \mid c_j^S).\end{aligned} \tag{1}$$

3.2 Post-Training Algorithms

3.2.1 Supervised fine-tuning (SFT)

Given demonstrations $\mathcal{D}_{\text{sup}} = \{(x, y^\star)\}$, SFT minimizes the teacher-forcing cross-entropy:

$$\begin{aligned}&\mathcal{L}_{\text{SFT}}(\theta) = \\ &\quad - \mathbb{E}_{(x,y^\star)\sim\mathcal{D}_{\text{sup}}} \left[\sum_{t=1}^{|y^\star|} \log \pi_\theta(y_t^\star \mid x, y_{<t}^\star) \right].\end{aligned} \tag{2}$$

Property (dense token-level constraint). SFT provides a *dense* learning signal at every token position. It is typically the most stable objective.

3.2.2 Group Relative Policy Optimization

For each x, GRPO samples a group of G trajectories $\{y^{(i)}\}_{i=1}^G$ from a behavior policy $\pi_{\theta_{\text{old}}}(\cdot \mid x)$ and obtains rewards $\{R^{(i)}\}_{i=1}^G$ where $R^{(i)} \triangleq R(x, y^{(i)})$. It uses a group-relative advantage $\hat{A}^{(i)}$ and a PPO-style clipped surrogate with KL regularization:

$$\begin{aligned}\mathcal{L}_{\text{GRPO}}(\theta) = -\,\mathbb{E}\Bigg[\frac{1}{G} \sum_{i=1}^{G} \sum_{t=1}^{|y^{(i)}|} \min \Big(& \\ \rho_t^{(i)}\, \hat{A}^{(i)}, & \\ \text{clip}\big(\rho_t^{(i)}, 1-\epsilon, 1+\epsilon\big)\, \hat{A}^{(i)} & \Big) \Bigg] \\ + \ \beta\, \mathbb{E}\Big[\text{KL}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\big) \Big], &\end{aligned} \tag{3}$$

where ϵ is ϵ_{clip} for brevity, and the token-level importance ratio is $\rho_t^{(i)} = \frac{\pi_\theta(y_t^{(i)} \mid x, y_{<t}^{(i)})}{\pi_{\theta_{\text{old}}}(y_t^{(i)} \mid x, y_{<t}^{(i)})}$.

Property (online; sparse credit shared across tokens). GRPO is an *online* method that requires rollouts to obtain rewards. Although loss are computed at the token level, the advantage $\hat{A}^{(i)}$ is sequence-level (shared across token positions), which can

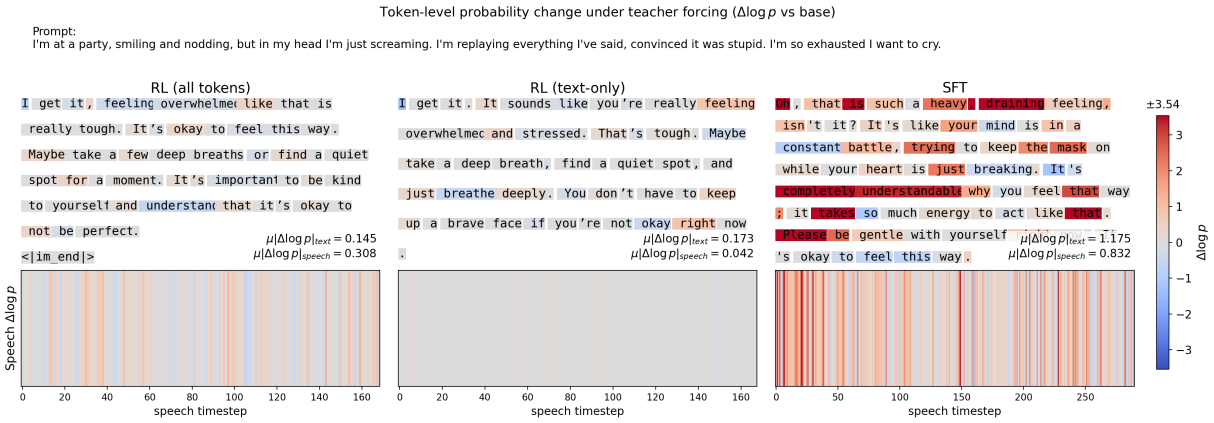

Figure 2: Token-level probability change under teacher forcing ($\Delta \log p$ vs. base) for the same prompt.

make credit assignment challenging for long and dense token streams. The KL term acts as a *dense* trust-region constraint that stabilizes optimization and prevents excessive policy drift.

3.2.3 Offline DPO-family

Given pairwise preference data $\mathcal{D}_{\text{pref}} = \{(x, y^+, y^-)\}$, DPO optimizes a logistic loss on the reference-corrected log-ratio gap. We define

$$\Delta(x,y^+,y^-;\theta) = \Big(\log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x)\Big) - \Big(\log \pi_{\text{ref}}(y^+ \mid x) - \log \pi_{\text{ref}}(y^- \mid x)\Big), \tag{4}$$

and minimize

$$\mathcal{L}_{\text{DPO}}(\theta) = - \mathbb{E}_{\mathcal{D}_{\text{pref}}}\Big[\log \sigma\big(\gamma \cdot \Delta(x, y^+, y^-; \theta)\big)\Big], \tag{5}$$

where $\sigma(\cdot)$ is the logistic sigmoid and $\gamma > 0$ is a temperature.

Property (offline; sparse preference supervision). DPO is *offline* and does not require rollouts during optimization, making it scalable in practice. Supervision signal comes only from pairwise preferences and performance depends on preference quality/coverage and potential judge bias.

Token-subset restricted likelihood. In mixed-modality settings, it is sometimes useful to restrict preference-driven updates to a subset of token positions. Given an index set $\mathcal{M}(y) \subseteq \{1, \ldots, |y|\}$, define the masked score:

$$s_{\mathcal{M}}(x, y; \theta) \triangleq \sum_{t \in \mathcal{M}(y)} \log \pi_\theta(y_t \mid x, y_{<t}). \tag{6}$$

Replacing $\log \pi_\theta(y \mid x)$ with $s_{\mathcal{M}}(x, y; \theta)$ yields token-subset restricted variants of SFT/PO objectives, enabling explicit control over which token types receive preference gradients.

3.3 Observations

As highlighted in Fig. 1, applying a single unified RL/PO objective to the mixed text–speech output leads to three coupled issues: (i) cross-modal trade-off and inconsistency, (ii) weak cross-modal gradient coupling with severe energy imbalance, and (iii) reward/signal dilution when sparse feedback is spread over overly dense speech tokens, making credit assignment ill-posed. Under token-score–based objectives, the additivity in Eq. (1) yields a natural gradient split

$$\nabla_\theta L(\theta) = \nabla_\theta L^{(T)}(\theta) + \nabla_\theta L^{(S)}(\theta), \tag{7}$$

which exposes the mechanism behind Fig. 1: semantic (text) updates carry much higher effective energy, while the acoustic component receives weakly informative, high-variance signals due to near-orthogonal coupling, whose *accumulated effect over dense speech tokens* can destabilize prosody/timbre. This motivates a division of labor: restrict preference-driven updates to I_T for semantic refinement, and use dense supervision to anchor I_S to preserve speech naturalness and expressive stability.

However, even after restricting preference-driven updates to I_T, where preference signals are typically most informative, preference optimization alone is insufficient in our setting due to two structural bottlenecks.

(Bottleneck I) Practical PO is inherently *local* under stability constraints (e.g., clipping and/or reference penalties), typically inducing substantially

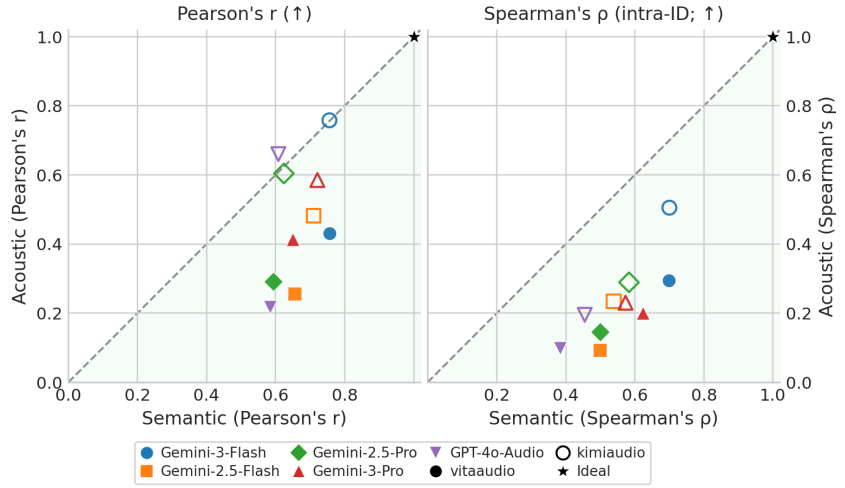

Figure 3: Judge/reward-model agreement with human evaluation on semantic vs. acoustic dimensions.

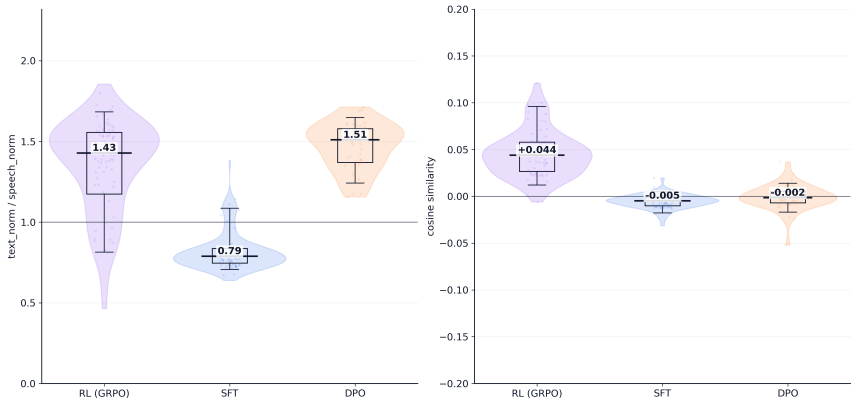

Figure 4: Empirical geometry of text vs. speech gradients under different objectives.

smaller distributional shifts than supervised fine-tuning (SFT), and may fail to move the model out of suboptimal regions when rollouts provide limited improvement signals.

(Bottleneck II) Speech naturalness/expressiveness lacks reliably learnable preference signals and strong rollouts: acoustic reward judgments can be noisy or misaligned with human preference, and weak base models rarely sample high-quality acoustic trajectories, yielding low reward discrimination, which makes direct preference learning on dense acoustic decisions fragile or even harmful. In what follows, we present a set of empirical observations (Figs. 2–5) and distill them into design conclusions that motivate the dynamic hybrid objective in Sec. 3.4.

Observation 1 (Fig. 2): SFT yields larger, coherent distribution shifts, while stabilized PO/RL is typically local. Fig. 2 shows that SFT induces substantially larger and more consistent token-level probability changes across the sequence, whereas PO/RL updates are smaller and localized under trust-region–like stability constraints. This matches Bottleneck I: on-policy rollouts often provide limited improvement signal, making PO a local shaper that may not escape suboptimal regions. Details of the teacher-forcing probability-change metric and modality partition are in Appendix D. *Implication:* use SFT to enact reliable global shifts/anchoring, and PO/RL to refine behavior locally when preference signals are informative.

Observation 2 (Fig. 3): Preference/reward is more informative for semantics than acoustics. Across repeated samples (Fig. 3), reward-model judgments agree with humans more strongly and stably on semantic quality than on acoustic quality, where agreement is weaker and more variable. Full human-eval protocol, agreement tables, and judge/reward prompts are in Appendix A and Appendix F. *Implication:* applying high-variance PO over dense speech tokens is fragile (noise amplification and harder credit assignment), so preference-

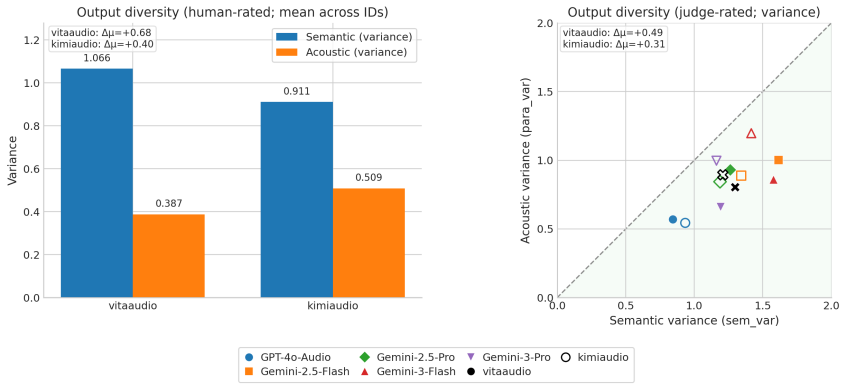

Figure 5: Semantic vs. acoustic diversity under repeated sampling reveals weaker acoustic discriminability.

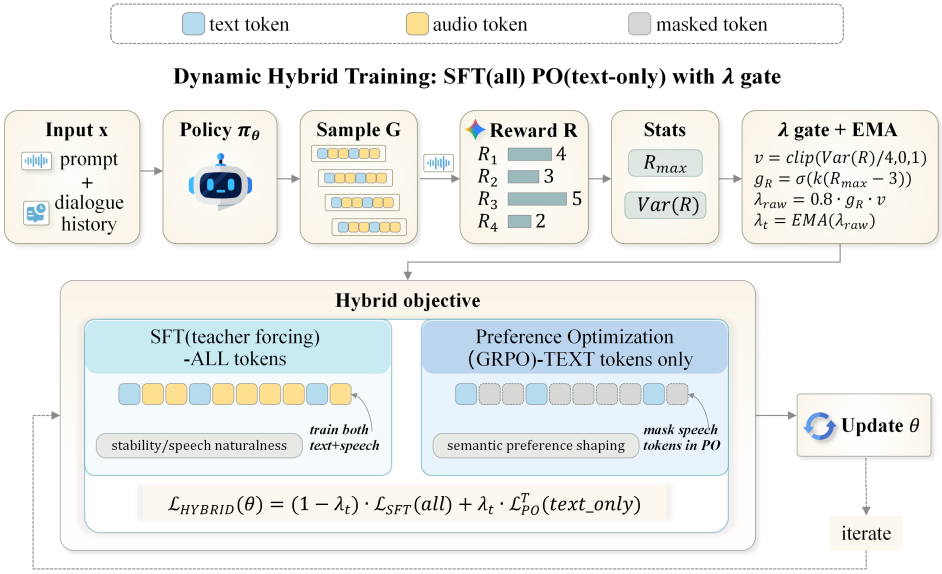

Figure 6: Overview of the proposed single stage adaptive hybrid post training.

driven updates should focus on I_T, while dense supervision stabilizes I_S.

Observation 3 (Fig. 4): Preference gradients concentrate on semantics; full-token PO yields low-SNR, high-variance updates on dense acoustics. Fig. 4 indicates weak cross-modal coupling (near-zero expected cosine similarity with high variance) and that preference objectives allocate most gradient energy to the semantic component. As a result, applying PO to the full mixed token stream assigns the same sequence-level credit to a large number of speech token decisions that are only weakly correlated with the preference signal, producing near-zero-mean but high-variance acoustic gradients. The accumulation of these noisy acoustic updates can destabilize prosody/timbre, motivating preference updates on token subsets. Gradient decomposition and all-layer analysis details are in Appendix C.

Observation 4 (Fig. 5): Rollout discriminability is uneven and stage-dependent, favoring adaptive gating over fixed mixing. With weaker base models, rollouts rarely contain high-quality acoustic trajectories, yielding low reward discrimination (small variance, few high-score samples); Fig. 5 further shows diversity/variance is uneven (often weakest along acoustics) and changes over training stages. Statistics and definitions are in Appendix B. *Implication:* a fixed hybrid weight either applies PO when signals are weak/noisy (instability) or underuses PO when discriminative samples exist; thus an adaptive controller (Sec. 3.4) should modulate λ_t using normalized reward variance and a good-sample existence gate.

Summary. Taken together, these observations motivate a principled division of labor: **SFT** acts as a robust distribution-shifting and feasibility-building operator (especially for speech natural-

ness/expressiveness through dense supervision), while **PO** acts as a local preference-shaping operator that refines semantic dialogue behavior when the preference signal is informative. Sec. 3.4 formalizes this view into a lightweight dynamic hybrid objective with adaptive gating.

3.4 Dynamic Hybrid Post-Training Objective

Fig. 6 summarizes our single-stage dynamic hybrid post-training loop. At each step, we sample a group of G spoken replies from π_θ; importantly, we decode the generated speech into audio and *feed the model's spoken reply directly to a reward model* to obtain scalar rewards $R_t = \{r_{t,i}\}_{i=1}^{G}$. Motivated by Fig. 1, we keep SFT as an explicit distribution anchor for acoustic stability, while applying preference optimization only on text tokens to refine semantics. We thus optimize

$$\mathcal{L}_{\text{hybrid}}(\theta) = (1-\lambda_t)\,\mathcal{L}_{\text{SFT}}(\theta) + \lambda_t\,\mathcal{L}_{\text{GRPO}}^{(T)}(\theta), \quad (8)$$

where $\mathcal{L}_{\text{GRPO}}^{(T)}$ masks the score in Eq. 6 with $M(y) = I_T(y)$, i.e., preference gradients are restricted to text tokens.

Lightweight gating for λ_t. We increase λ_t only when rollouts are (i) *directionally reliable* (at least one acceptable sample exists) and (ii) *discriminative* (candidates are well-separated), matching the two statistics shown in Fig. 6. Let $R_{\max,t} = \max(R_t)$ and define a normalized variance

$$v_t \triangleq \text{clip}\left(\frac{\text{Var}(R_t)}{4},\, 0,\, 1\right), \quad (9)$$

where 4 is the maximum variance on a 1–5 Likert scale (bimodal at $\{1, 5\}$), so $v_t \in [0, 1]$ is comparable across steps. The direction gate is

$$g_t(R) = \sigma(k\,(R_{\max,t} - 3)), \quad (10)$$

where the threshold 3 corresponds to neutral/acceptable quality; if all samples fall below 3, preference gradients are typically noisy/misdirected. We use $g_t(v_t) = v_t$ as the information gate (where v_t is the normalized reward variance defined above) and set

$$\lambda_t^{\text{raw}} = \lambda_{\max}\, g_t(R)\, g_t(V), \qquad \lambda_{\max} = 0.8, \quad (11)$$

so that at least $1 - \lambda_{\max} = 0.2$ of SFT is always retained as a safety anchor against acoustic drift when rewards are imperfect. The slope k is the only sharpness hyperparameter, controlling how softly we transition from "mostly SFT" to "more GRPO".

EMA smoothing. To reduce step-to-step oscillations from on-policy sampling, we smooth the weight itself:

$$\lambda_t = (1-\alpha)\lambda_t^{\text{raw}} + \alpha\lambda_{t-1}, \qquad \alpha = 0.9, \quad (12)$$

where $\alpha = 0.9$ provides a strong low-pass filter that stabilizes training while remaining responsive to sustained improvements in rollout quality/discriminability.

4 Experiments

4.1 Experimental Setup

Training Data. To cover both *intelligence* (reasoning, knowledge, instruction following, safety) and *expressiveness* (paralinguistic controllability and empathy), we curate a mixed training set of 13.5k audio-instruction samples from public and self-constructed sources, including UltraChat (Ding et al., 2023), SciQ (Johannes Welbl, 2017), GSM8K (Cobbe et al., 2021), SHP (Ethayarajh et al., 2022), ExamQA, Alpaca (Taori et al., 2023), ScienceQA (Lu et al., 2022), Ai2ARC (Clark et al., 2018), PKUSafe (Ji et al., 2024a), as well as self-constructed logic and expressiveness data. All 13.5K training samples are from public or self-constructed sources; no proprietary in-house data are used. We further build preference pairs via repeated sampling and judge-based scoring to support offline preference learning. The detailed data composition, construction procedures, and preference-pair pipeline are provided in Appendix E.

Benchmarks and Metrics. We evaluate on three benchmarks comprising **18 sub-tasks** to jointly assess semantic competence (IQ) and expressive capability (EQ). **VoiceBench** evaluates instruction following and safety, including *AlpacaEval, CommonEval, WildVoice, SD-QA, MMSU, OBQA, BBH, IFEval,* and *AdvBench*. **OpenAudioBench** focuses on knowledge and reasoning with *Alpaca, Llama, Web, Trivial,* and *Reason*. **VStyle** targets paralinguistic control via *Acoustic Attributes, Instruction Following (Style), Role Play,* and *Empathy*. We strictly follow the official evaluation protocol for each benchmark: VoiceBench and OpenAudioBench are evaluated on text outputs via their official pipelines (GPT-4o-mini and GPT-4o as judges, respectively); VStyle is evaluated on speech outputs via its official procedure using Gemini-2.5-Pro. All scores are computed using the official scripts and judges for each benchmark.

Method	VoiceBench									Avg	OpenAudioBench					Avg
	Alpaca	Common	Wild	SD-QA	MMSU	OBQA	BBH	IFEval	Adv		Alpaca	Llama	Reason	Trivial	Web	
VITA Architecture (Interleaved)																
VITA-Base	3.83	3.44	3.09	29.2	48.7	74.3	58.2	26.2	94.1	–	60.6	73.8	44.2	42.9	53.5	55.0
SFT (Teacher Forcing)	3.45	3.12	2.85	27.6	45.1	71.5	54.9	28.4	**99.2**	–	55.6	71.1	38.4	39.9	48.3	50.7
DPO Baselines																
Full-Token DPO	3.60	3.29	2.89	30.2	44.7	69.2	56.8	22.6	65.0	–	20.1	55.4	33.4	29.8	36.6	35.1
Text-Token DPO	3.91	3.32	3.13	31.1	45.6	69.7	**60.3**	**32.8**	71.3	–	57.2	74.3	43.1	43.1	54.3	54.4
RL Baselines																
Full-Token RL (Unified)	4.03	3.45	3.19	29.9	49.0	74.1	55.6	29.4	96.3	–	63.8	73.3	43.7	43.3	52.8	55.4
Text-Token RL (Unified)	4.09	3.44	3.20	31.3	50.0	75.4	56.7	30.2	96.3	–	64.6	**74.6**	44.4	**44.4**	53.2	56.2
SFT + RL (Two-Stage)	3.49	3.32	2.69	22.5	44.7	70.8	54.2	25.5	98.8	–	54.0	66.2	32.8	35.1	49.0	47.4
Ours (Dynamic)	**4.22**	**3.51**	**3.29**	**31.5**	**51.4**	**77.1**	59.9	32.5	97.1	–	**68.4**	**74.6**	**46.1**	**44.4**	**54.7**	**57.6**
KimiAudio Architecture (Parallel)																
KimiAudio-Base	4.46	3.97	3.42	63.1	62.2	83.5	64.2	61.1	**100.0**	–	75.7	79.3	58.0	**62.1**	70.2	69.1
SFT	4.15	3.65	3.10	59.8	58.4	79.5	61.2	64.5	**100.0**	–	71.4	75.2	52.8	58.4	66.9	64.9
Preference Optimization (DPO & RL)																
Full-Token DPO	4.05	3.60	3.05	58.2	55.1	76.8	59.4	58.4	88.5	–	68.2	70.4	50.1	55.3	65.1	61.8
Full-Token RL	4.52	4.05	3.50	65.2	63.8	84.6	64.8	62.8	**100.0**	–	75.8	78.5	58.8	61.2	**71.5**	69.2
Ours (Dynamic)	**4.58**	**4.22**	**3.68**	**67.9**	**66.5**	**87.1**	**68.3**	**66.8**	99.5	–	**78.5**	**81.2**	**61.5**	61.8	71.1	**70.8**

Table 1: Main Results on Intelligence (IQ). Best results are **bolded** and second best are underlined. Avg for VoiceBench is omitted (mixed scales); Avg for OpenAudioBench is the arithmetic mean of the five sub-task scores.

Baselines To demonstrate architectural generality, we experiment with two distinct end-to-end speech dialogue backbones: (i) **VITA-Audio** (Long et al., 2025), which emits an interleaved sequence of text and speech tokens, and (ii) **KimiAudio** (Ding et al., 2025), which follows a parallel design. We group baselines into: **(1) Standard:** the *Base Model* and *SFT* (teacher forcing). **(2) DPO:** offline Direct Preference Optimization applied to either *Full* tokens or *Text* tokens only. **(3) RL:** GRPO applied to *Full Tokens* or *Text Tokens* only, plus a sequential *Two-Stage* recipe (SFT→RL). For DPO baselines, preference supervision is constructed via repeated sampling and judge scoring; details are in Appendix E.

Implementation details. All training runs use 4×A100 GPUs. For RL, we adopt a KL-regularized objective with coefficients $\beta_{\text{text}} = 0.01$ and $\beta_{\text{speech}} = 0.01$. Unless otherwise specified, we use group size $G = 4$, sampling temperature $T = 0.9$, top-$p = 0.9$, learning rate $1\text{e}{-6}$, batch size 1, and maximum sequence length 2048. We use Gemini-2.5-pro as the reward model to score model speech responses; prompting templates and output formats for semantic and paralinguistic scoring are given in Appendix F.

4.2 Main Results

Intelligence (IQ). Table 1 shows that teacher-forced SFT *consistently underperforms* the base model on reasoning-heavy subsets, suggesting an alignment tax in our audio-instruction setting. We attribute this mainly to the *broad multi-domain* supervision: a comparatively small audio-instruction corpus must cover heterogeneous skills, which increases gradient interference and can overwrite pre-trained reasoning behaviors. Across both backbones, full-token preference optimization is suboptimal, while restricting preference updates to *text/semantic tokens* yields more reliable IQ gains, supporting modality-decoupled optimization. Building on this, our dynamic hybrid further mitigates catastrophic forgetting while preserving preference-learning benefits, achieving the strongest overall IQ among compared methods.

Expressiveness (EQ). On VStyle (Table 2), SFT remains highly competitive, especially on style instruction following and acoustic attributes, indicating that dense supervision is effective for imprinting fine-grained paralinguistic realizations. In contrast, naive preference optimization over the full mixed-modality sequence can be unstable for expressive speech: full-token DPO exhibits severe degradation, consistent with noisy or weakly discriminative acoustic reward signals. Our method achieves the best aggregate EQ across dimensions on both architectures, while staying close to the best style-following score, yielding a better IQ–EQ Pareto trade-off than either component alone.

4.3 Ablation Studies and Analysis

4.3.1 Weighting schemes and optimization scope

We study how to combine dense supervision (SFT) with preference optimization (RL/PO) for mixed-modality spoken outputs by varying two factors: **(i) optimization scope:** applying preference updates to *all* tokens versus restricting them to *text* tokens and **(ii) weighting strategy:** *fixed* SFT/RL mixtures versus *dynamic* weights predicted from

Method	Acoustic	Instruct.	Role Play	Empathy	Avg
VITA Architecture					
VITA-Base	2.26	1.76	2.15	4.01	2.55
SFT (Teacher Forcing)	2.34	**2.29**	2.31	3.42	2.59
DPO Baselines					
Full-Token DPO	1.49	1.25	1.10	1.05	1.22
Text-Token DPO	2.03	1.64	2.19	4.38	2.56
RL Baselines					
Full-Token RL (Unified)	2.16	1.64	1.97	3.95	2.43
Text-Token RL (Unified)	2.21	1.93	2.08	4.02	2.56
Ours (Dynamic)	**2.55**	2.25	**2.41**	**4.44**	**2.91**
KimiAudio Architecture					
KimiAudio-Base	2.53	2.31	1.73	3.67	2.56
SFT	2.65	**2.58**	1.95	3.65	2.71
Preference Opt.					
Full-Token DPO	1.85	1.55	1.30	2.10	1.70
Full-Token RL	2.58	2.25	1.88	3.88	2.65
Ours (Dynamic)	**2.78**	2.52	**2.15**	**4.15**	**2.90**

Table 2: Main Results on Expressiveness (EQ). Comparison on VStyle. Avg is the arithmetic mean of the four dimension scores. **Bold**/Underline indicate best/second best.

Scope	Strategy	IQ	EQ
All Tokens	Fixed Weights (0.5/0.5)	48.70	2.48
Text Tokens	Fixed Weights (0.5/0.5)	52.60	2.60
Text Tokens	Fixed Weights (0.7 SFT / 0.3 RL)	49.94	2.72
All Tokens	Dynamic Weights	48.84	2.50
Text Tokens	Dynamic Weights w/o EMA	53.15	2.53
Text Tokens	**Ours (Dynamic Weights)**	**55.24**	**2.92**

Table 3: Weighting schemes and optimization scope. *Scope* refers to the token subset over which the GRPO/RL loss is applied; SFT always covers all tokens regardless of scope.

rollout quality. All variants share the same backbone, data, and training budget; we report **IQ** (mean over VoiceBench reasoning subsets: MMSU/OBQA/BBH/IFEval) and **EQ** (average VStyle score). Table 3 indicates that *scope* is crucial: with the same fixed 0.5/0.5 mixture, applying preference optimization only to text tokens clearly outperforms updating all tokens (52.60/2.60 vs. 48.70/2.48 in IQ/EQ), implying preference gradients are most effective when focused on semantic-bearing regions. Fixed weights also reveal an IQ–EQ trade-off—favoring SFT (0.7/0.3) improves EQ (2.72) but lowers IQ (49.94). Dynamic weighting over all tokens remains limited (48.84/2.50), whereas dynamic gating with text-token scope delivers the best overall result; EMA smoothing is important for stability (w/o EMA: 53.15/2.53 vs. ours: 55.24/2.92).

4.3.2 Subjective human evaluation

We conduct a side-by-side (SBS) human study comparing **Ours** with the **Original Model** baseline on VITA-Audio. Annotators blindly rate paired responses along two axes: *Helpfulness* (instruction adherence and logical coherence) and *Naturalness* (prosody, timbre, and emotional appropriateness). The full protocol, criteria definitions, and aggregation procedure are provided in Appendix G. As shown in Table 4, our model is preferred in both dimensions.

Dimension	Win (%)	Tie (%)	Loss (%)	p-value
Helpfulness	63.8	16.2	20.0	< 0.001
Naturalness	66.2	13.8	20.0	< 0.001
Overall	**68.8**	**13.7**	**17.5**	$< \mathbf{0.001}$

Table 4: Human Subjective Evaluation results (80 items, 3 raters per item). **Ours** significantly outperforms baseline, achieving a $\sim$4:1 win-to-loss ratio overall.

5 Conclusion

We analyze the optimization mismatch in reinforcement learning of end-to-end spoken dialogue models, showing how preference updates can dilute semantic signals and induce acoustic drift, and we propose an adaptive hybrid post-training method that stabilizes speech while improving both intelligence and expressiveness across architectures and benchmarks.

Limitations.

Our study focuses on *sequence-level* reward signals. Beyond our hybrid loss framework, providing speech tokens with more reliable and denser guidance (e.g., PPO with stronger token-level or frame-level feedback) may further improve speech quality and stability. Due to limited resources, we are unable to run PPO-based speech-token experiments in this work. Additionally, audio judges are not yet on par with text/semantic judges in terms of reliability and calibration; the story told by our motivating observations and final results may differ with a better-calibrated audio judge. We will investigate the effect of improved audio judges in future work.

6 Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant No.U25B2064 and Alibaba Research Intern Program.

References

Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. 2025a. Beyond two-stage training: Cooperative sft and rl for llm reasoning. *arXiv preprint arXiv:2509.06948*.

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao.

2025b. WavRAG: Audio-integrated retrieval augmented generation for spoken dialogue models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12505–12523, Vienna, Austria. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

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

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. 2025. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *arXiv preprint arXiv:2506.19767*.

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. 2025. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ailin Huang, Bingxin Li, Bruce Wang, Boyong Wu, Chao Yan, Chengli Feng, Heng Wang, Hongyu Zhou, Hongyuan Wang, Jingbei Li, and 1 others. 2025. Step-audio-aqaa: a fully end-to-end expressive large audio language model. *arXiv preprint arXiv:2506.08967*.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. 2024a. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. *arXiv preprint arXiv:2406.15513*.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024b. Wavchat: A survey of spoken dialogue models. *Preprint*, arXiv:2411.13577.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, and 1 others. 2024c. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, Siqi Zheng, Jin Xu, Junyang Lin, and Zhou Zhao. 2025. Wavreward: Spoken dialogue models with generalist reward evaluators. *Preprint*, arXiv:2505.09558.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions.

LI Kai, FU Qiang, and YAN Yonghong. 2012. Speech enhancement using robust generalized sidelobe canceller with multi-channel post-filtering in adverse environments. *Chinese Journal of Electronics*, 21(1):85–90.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Yangzhuo Li, Shengpeng Ji, Yifu Chen, Tianle Liang, Haorong Ying, Yule Wang, Junbo Li, Jun Fang, and Zhou Zhao. 2026. Wavbench: Benchmarking reasoning, colloquialism, and paralinguistics for end-to-end spoken dialogue models. *arXiv preprint arXiv:2602.12135*.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. 2025. Uft: Unifying supervised and reinforcement fine-tuning. *arXiv preprint arXiv:2505.16984*.

Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, Haoyu Cao, Ke Li, Rongrong Ji, and Xing Sun. 2025. Vita-audio: Fast interleaved cross-modal token generation for efficient large speech-language model. *Preprint*, arXiv:2505.03739.

Jingyu Lu, Yuhan Wang, Fan Zhuo, Xize Cheng, Changhao Pan, Xueyi Pu, Yifu Chen, Chenyuhao Wen, Tianle Liang, and Zhou Zhao. 2026. Modeling and benchmarking spoken dialogue rewards with modality and colloquialness. *Preprint*, arXiv:2603.14889.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, and 1 others. 2025. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Cong Wang, Changfeng Gao, Yang Xiang, Zhihao Du, Keyu An, Han Zhao, Qian Chen, Xiangang Li, Yingming Gao, and Ya Li. 2025. Rrpo: Robust reward policy optimization for llm-based emotional tts. *arXiv preprint arXiv:2512.04552*.

Anne Wu, Laurent Mazaré, Neil Zeghidour, and Alexandre Défossez. 2025a. Aligning spoken dialogue models from user interactions. In *Forty-second International Conference on Machine Learning*.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025b. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv: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, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*.

Shu-wen Yang, Ming Tu, Andy T Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. 2025. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. *arXiv preprint arXiv:2511.08723*.

Xiangchi Yuan, Xiang Chen, Tong Yu, Dachuan Shi, Can Jin, Wenke Lee, and Saayan Mitra. 2025. Mitigating forgetting between supervised and reinforcement learning yields stronger reasoners. *arXiv preprint arXiv:2510.04454*.

Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2024a. Speechalign: Aligning speech generation to human preferences. *Advances in Neural Information Processing Systems*, 37:50343–50360.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. 2025. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *arXiv preprint arXiv:2508.11408*.

Yu Zhang, Changhao Pan, Wenxiang Guo, Ruiqi Li, Zhiyuan Zhu, Jialei Wang, Wenhao Xu, Jingyu Lu, Zhiqing Hong, Chuxin Wang, and 1 others. 2024b. Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. *Advances in Neural Information Processing Systems*, 37:1117–1140.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

A Reward Model Consistency with Human Experiments

This section details the judge/reward-model consistency study summarized in *Figure 3 (main paper)*, which compares automatic judges with human evaluation on **semantic** vs. **acoustic (paralinguistic)** dimensions under repeated sampling.

A.1 Evaluation data and repeated-sampling protocol

We evaluate on two datasets, `vitaaudio` and `kimiaudio`. Each dataset contains **40 prompt IDs** (20 audio-type + 20 text-type). For each prompt ID, we sample the same model **8 times** to obtain multiple spoken answers (a small number of IDs have 7 samples due to decoding/I/O failures). Every sampled answer receives: (i) human ratings (semantic + acoustic), and (ii) judge-model ratings (semantic + acoustic).

Across IDs, the resulting human-rated sample counts are: `vitaaudio`: 318 (160 audio + 158 text), and `kimiaudio`: 319 (159 audio + 160 text).

Protocol details (decoding and human rating). For each prompt ID, we generate n=8 stochastic spoken responses from the same checkpoint under fixed decoding settings. Unless otherwise specified, we use nucleus sampling with temperature T=0.9 and top-p=0.9, and cap the maximum sequence length to 2048 tokens.[1] A small number of prompt IDs yield n=7 samples due to decoding or I/O failures; we keep all successfully generated samples and report the effective sample count used in each analysis.

Human rating rubric (1–5 Likert; two independent axes). Raters evaluate each sampled response along two axes: (i) **Semantic quality** and (ii) **Acoustic/paralinguistic quality**. Each axis is rated on a 1–5 Likert scale. **Important separation:** *(a)* judge **semantics** using the provided **transcript only** (ignore the voice/acoustics); *(b)* judge **acoustics** using the **audio only** (ignore factual correctness).

A. Semantic quality (transcript-only). Evaluate: (1) *accuracy & relevance*, (2) *completeness*, (3) *coherence/structure*. Do not give credit for information that is not present in the transcript.

- **5 – Excellent:** Fully correct and directly answers the query; includes key details/steps; logically organized and easy to follow; no noticeable issues.
- **4 – Good:** Mostly correct and on-topic; minor omission, minor imprecision, or slightly suboptimal structure, but the answer remains clearly useful.
- **3 – Acceptable:** Partially correct and generally on-topic, but has noticeable gaps (missing key detail/step), mild confusion, or some irrelevant content; still usable with caveats.
- **2 – Poor:** Major factual errors, significant irrelevance, or broken reasoning/structure; user would likely be misled or unable to complete the task.
- **1 – Very poor:** Wrong/off-topic, incoherent, or effectively non-answer (e.g., refuses without reason); unusable.

B. Acoustic/paralinguistic quality (audio-only). Evaluate: (1) *clarity/intelligibility*, (2) *fluency*, (3) *pronunciation/accent*, (4) *prosody/pacing*, (5) *emotional appropriateness*. **Do not penalize solely for a synthetic timbre** if intelligibility and delivery are otherwise good.

- **5 – Excellent:** Very clear and easy to understand; smooth flow; pronunciation/accent never hinders comprehension; natural pacing and prosody; emotion/tone fits the context.
- **4 – Good:** Generally clear and fluent; small artifacts or occasional awkwardness (minor mispronunciation, brief unnatural pause, slightly monotone), but overall comfortable to listen to.
- **3 – Acceptable:** Understandable but with noticeable issues (frequent monotony, several mispronunciations, mildly distracting accent, or pacing that is sometimes too fast/slow).
- **2 – Poor:** Hard to follow due to significant clarity problems, disfluencies, pronunciation/accent issues, or consistently mismatched prosody/emotion.
- **1 – Very poor:** Largely unintelligible, severely distorted/clipped/noisy, or extremely uncomfortable to listen to.

Tie-breaking / consistency notes. When uncertain between adjacent scores, prefer the lower score unless the response clearly meets the higher-level description. For multi-rater cases, aggregate per-sample scores by averaging across raters.

Aggregation. If multiple raters score the same sample, we aggregate per-sample scores by averaging across raters, and then compute global agreement metrics over all samples, as well as intra-ID ranking agreement by computing Spearman correlation within each prompt ID and averaging across IDs.

A.2 Metrics

We measure agreement in two complementary regimes:

[1] We keep decoding hyperparameters fixed across all repeated-sampling experiments to ensure that within-ID variability reflects model stochasticity rather than configuration changes.

Global agreement across all samples. We compute Pearson correlation between judge scores and human scores over all rated samples (semantic and acoustic separately). We additionally report MAE, the ≤ 1 pass rate $\Pr(|s_{\text{judge}} - s_{\text{human}}| \leq 1)$, and bias $\mathbb{E}[s_{\text{judge}} - s_{\text{human}}]$.

Intra-ID ranking agreement. Since repeated sampling is used for both diversity analysis and preference construction (Sections B and E), we also evaluate within-prompt discriminability by computing *Intra-ID Spearman*: for each prompt ID, compute Spearman correlation between the judge and human score sequences across repeated samples, then average across IDs.

A.3 Results

Tables 5 and 6 provide the full agreement statistics used in the analysis.

A.4 Implication for training-signal reliability

Across datasets and judges, the most stable gap appears in **Intra-ID Spearman**: semantic ranking agreement is consistently stronger than acoustic ranking agreement. Since repeated-sampling selection is exactly the regime used for diversity statistics and preference-pair construction, this motivates treating semantic judgments as the primary reliable discriminator for within-prompt ranking, while maintaining speech feasibility through dense supervision.

B Model Output Diversity Experiments

This section details the output diversity analysis summarized in *Figure 5 (main paper)*. Diversity is computed over the **entire repeated-sampling pool** described in Section A: for each prompt ID we sample multiple spoken answers and quantify within-ID dispersion.

B.1 Per-ID variance

For a prompt ID with n sampled answers and per-sample scores $\{s^{(1)}, \ldots, s^{(n)}\}$ (semantic or acoustic), we compute:

$$\text{Var}_{\text{ID}} = \frac{1}{n} \sum_{i=1}^{n} \left(s^{(i)} - \bar{s} \right)^2, \qquad \bar{s} = \frac{1}{n} \sum_{i=1}^{n} s^{(i)}.$$

We aggregate by averaging over IDs:

$$\text{Var}_{\text{dataset}} = \frac{1}{|\mathcal{I}|} \sum_{\text{ID} \in \mathcal{I}} \text{Var}_{\text{ID}}.$$

B.2 Human-rated diversity

Using human ratings, the mean variance across IDs is:

- `vitaaudio`: semantic variance 1.066, acoustic variance 0.387.
- `kimiaudio`: semantic variance 0.911, acoustic variance 0.509.

B.3 Judge-rated diversity

Using judge ratings, we compute the same variance and plot semantic variance (x-axis) against acoustic variance (y-axis) for each judge and dataset. Most points lie below the diagonal, indicating systematically weaker acoustic discriminability under repeated sampling.

B.4 Shared lineage with DPO construction and consistency evaluation

The same repeated-sampling structure underlies: (i) judge-vs-human agreement (Section A), (ii) diversity statistics (this section), and (iii) DPO preference-pair construction (Section E). Therefore, acoustic discriminability limitations observed in the diversity and consistency analyses directly inform preference-training design choices.

C Grad Analysis Experiments

This section provides implementation-level details for the gradient geometry analysis summarized in *Figure 4(main paper)*.

C.1 Two-loss decomposition with two backward passes

We decompose training into two modality-associated losses:

- $\mathcal{L}_{\text{text}}$: loss defined on text-token predictions.
- $\mathcal{L}_{\text{speech}}$: loss defined on speech-token predictions.

For each logged event, we compute gradients via **two separate backward computations**:

1. Zero gradients and backpropagate $\mathcal{L}_{\text{text}}$ to obtain $\mathbf{g}_{\text{text}}$.
2. Zero gradients and backpropagate $\mathcal{L}_{\text{speech}}$ to obtain $\mathbf{g}_{\text{speech}}$.

This yields clean modality-separated gradients rather than mixing both losses in a single backward pass.

Judge	$\text{Pearson}_{\text{sem}}$	$\text{Pearson}_{\text{acous}}$	Intra-ID $\text{Spearman}_{\text{sem}}$	Intra-ID $\text{Spearman}_{\text{acous}}$	MAE_{sem}	$\text{MAE}_{\text{acous}}$	$\leq$1 Pass_{sem}	$\leq$1 $\text{Pass}_{\text{acous}}$	Bias_{sem}	$\text{Bias}_{\text{acous}}$
Gemini-3-Flash	0.757	0.430	0.699	0.294	0.553	0.890	0.909	0.808	+0.170	+0.500
Gemini-2.5-Flash	0.656	0.256	0.500	0.092	0.869	1.048	0.789	0.741	-0.224	-0.192
Gemini-2.5-Pro	0.593	0.292	0.499	0.145	1.063	1.318	0.695	0.601	-0.660	-0.896
Gemini-3-Pro	0.650	0.413	0.623	0.201	0.963	1.269	0.748	0.650	-0.697	-1.085
GPT-4o-Audio	0.584	0.218	0.383	0.098	0.904	0.929	0.779	0.817	-0.263	+0.526

Table 5: Judge vs. human agreement on `vitaaudio`.

Judge	$\text{Pearson}_{\text{sem}}$	$\text{Pearson}_{\text{acous}}$	Intra-ID $\text{Spearman}_{\text{sem}}$	Intra-ID $\text{Spearman}_{\text{acous}}$	MAE_{sem}	$\text{MAE}_{\text{acous}}$	$\leq$1 Pass_{sem}	$\leq$1 $\text{Pass}_{\text{acous}}$	Bias_{sem}	$\text{Bias}_{\text{acous}}$
Gemini-3-Flash	0.756	0.758	0.700	0.505	0.621	0.671	0.862	0.875	-0.257	-0.144
Gemini-2.5-Flash	0.710	0.482	0.539	0.234	0.787	1.219	0.790	0.602	-0.223	-0.762
Gemini-2.5-Pro	0.624	0.604	0.583	0.289	0.981	1.276	0.708	0.564	-0.574	-1.075
Gemini-3-Pro	0.721	0.585	0.573	0.230	0.819	1.156	0.778	0.657	-0.546	-0.908
GPT-4o-Audio	0.608	0.660	0.455	0.195	0.902	0.792	0.782	0.836	-0.183	+0.445

Table 6: Judge vs. human agreement on `kimiaudio`.

C.2 All-layer analysis and aggregation

We perform layer-wise analysis over **all layers** (embeddings, each transformer block, and output heads). For each layer ℓ, we compute:

$$\|g^{\ell}_{\text{text}}\|_2, \quad \|g^{\ell}_{\text{speech}}\|_2, \quad \cos^{\ell} = \frac{\langle g^{\ell}_{\text{text}}, g^{\ell}_{\text{speech}} \rangle}{\|g^{\ell}_{\text{text}}\|_2 \|g^{\ell}_{\text{speech}}\|_2}.$$

We also compute the global (all-parameter) statistics used in the figure:

$$\text{ratio} = \frac{\|\mathbf{g}_{\text{text}}\|_2}{\|\mathbf{g}_{\text{speech}}\|_2}, \qquad \cos = \frac{\langle \mathbf{g}_{\text{text}}, \mathbf{g}_{\text{speech}} \rangle}{\|\mathbf{g}_{\text{text}}\|_2 \|\mathbf{g}_{\text{speech}}\|_2}.$$

C.3 Log parsing and statistical comparisons

We parse logged gradient summaries for RL (GRPO), SFT, and DPO, treating each logged record as one event sample. We visualize empirical distributions of `ratio` and `cos` and conduct non-parametric comparisons (Mann–Whitney U with multiple-comparison correction; effect sizes via Cliff's δ).

D Fine Tune Paradigms' Effect on Model Distribution Experiments

This section details the teacher-forcing probability-change analysis summarized in *Figure 2 (main paper)*.

D.1 Teacher-forcing log-probability change

Given an input prompt x and a fixed target continuation $y = (y_1, \ldots, y_T)$ generated by fine tuned model(including text and speech tokens), we compute teacher-forcing log probabilities under: (i) a base model p_{base} and (ii) a fine-tuned model p_{ft}. For each position t:

$$\Delta_t = \log p_{\text{ft}}(y_t \mid x, y_{<t}) - \log p_{\text{base}}(y_t \mid x, y_{<t}).$$

We visualize Δ_t over token positions, separating text-token and speech-token regions.

D.2 Modal partition

Tokens are partitioned into text and speech segments according to the model's tokenization protocol. The analysis highlights how dense teacher-forcing updates can reshape likelihood mass across speech-token regions, while preference-style objectives primarily reweight relative outcomes.

E Train Datasets

This section documents the full converted training mixture and specifies how self-built datasets and preference data are constructed.

E.1 Unified conversion summary

All datasets are converted into a unified pool. Total input samples: **13,510**

E.2 Dataset inventory

We replay detailed dataset composition in Table 7

E.3 Self-built dataset construction (control/understanding)

Self-built datasets target (i) **style control** (`controller_*`) and (ii) **style understanding** (`understand_*`) under explicit JSON schemas. The prompt templates used for generation are provided verbatim in Figures 7–12.

Pair selection rule. For each prompt x, we obtain n=8 candidate spoken responses $\{y^{(i)}\}_{i=1}^{n}$. For each candidate, the judge produces two scalar scores on a 1–5 scale: a semantic score $s^{(i)}_{\text{sem}}$ and a paralinguistic/acoustic score $s^{(i)}_{\text{acous}}$.

We convert them into a single utility score via a fixed weighted sum:

$$u^{(i)} = \lambda\, s^{(i)}_{\text{sem}} + (1-\lambda)\, s^{(i)}_{\text{acous}}, \quad \lambda \in [0,1]. \quad (13)$$

Unless otherwise specified, we set $\lambda = 0.5$.

Dataset name	#Samples	Provenance
gsm8k	2150	Public
ultrachat_dialogues	491	Public
ai2_arc_easy	1000	Public
ai2_arc_challenge	1000	Public
sciq	375	Public
shp	510	Public
examqa	326	Public
alpaca	277	Public
science_qa	299	Public
pkusafe	31	Public
controller_emotion	83	Self-built (style control)
controller_volume	60	Self-built (style control)
controller_pace	58	Self-built (style control)
understand_emotion	199	Self-built (style understanding)
understand_volume	200	Self-built (style understanding)
understand_pace	200	Self-built (style understanding)
emotion_dialogue_multi	1000	Self-built (expressive dialogue)
voice_repetition_single	1430	Self-built (robustness / repetition)
train_rl_logic	500	Self-built
train_rl_math	500	Self-built
train_rl_code	500	Self-built
train_rl_creating_writing	542	Self-built
ultra	676	Internal-curated/mixed
math	420	Internal-curated/mixed
en_zhishi_dialogue	319	Internal-curated/mixed
instruction_following_en	306	Internal-curated/mixed
unsafety_question	43	Internal-curated/mixed
poem	15	Internal-curated/mixed
Total	**13510**	

Table 7: Training dataset mixture in the converted pool.

We then select the preferred and rejected samples as:

$$i^+ = \arg\max_i \ u^{(i)}, \tag{14}$$

$$i^- = \arg\min_i \ u^{(i)}. \tag{15}$$

To reduce noisy preference signals, we only keep a pair if the utility gap exceeds a margin:

$$u^{(i^+)} - u^{(i^-)} \geq \delta, \tag{16}$$

where we use $\delta = 0.5$ by default. If multiple candidates tie in $u^{(i)}$, we break ties by preferring higher semantic score first, then higher acoustic score.

Finally, we form a single DPO pair $(x, y^{(i^+)}, y^{(i^-)})$ per prompt. This construction yields stable within-prompt ranking supervision while preserving diversity from repeated sampling.

F Reward Model Prompts

This section lists the reward/judge prompts used for automatic scoring in repeated-sampling analyses (Sections A–B) and for preference-signal construction (Section E). Figures 14–16 provide the verbatim prompts.

G Human Subjective Evaluation Protocol

This section details the subjective evaluation reported in *Table 4 (main paper)*.

G.1 Side-by-Side (SBS) setup

We conduct a blind side-by-side (SBS) evaluation comparing our model against the Original Model baseline. For each test item, annotators are presented with the same user query and two candidate spoken responses (A/B) produced under identical input conditions. To ensure blindness, model identities are hidden, and the A/B ordering is randomized per item and per annotator.

Evaluation set. We evaluate on 40 items in total, consisting of 20 items uniformly sampled from VOICEBENCH and 20 items uniformly sampled from VSTYLE. Unless otherwise specified, we sample without replacement and keep the original prompts in these benchmarks unchanged.

Each item is rated by 3 independent annotators. Annotators may replay each audio response, and are instructed to use headphones in a quiet environment when possible.

```
PROMPT_TEMPLATE = """
# Role
You are a creative writer. Your task is to take a topic and write six different natural-sounding, descriptive, and
      emotionally neutral personal statements about it.

# Task
For the given "topic", generate six different natural-sounding, descriptive first-person statements. The statements must be
      purely descriptive and avoid any emotional language.

# Core Requirements
1.  **Six Distinct Statements**: You must generate six unique statements.
2.  **Emotionally Neutral**: All statements must be objective and avoid conveying any strong emotions like happiness,
      sadness, anger, fear, or anxiety. The tone should be calm and observational.
3.  **First-Person Narrative**: Write from an "I" perspective.
4.  **Word Count**: Each statement must be a minimum of 50 words.
5.  **Descriptive Content**: Focus on sensory details and factual observations related to the topic.

# Input
- **Topic**: "{topic}"

# Output Format
You must provide a valid JSON object containing ONLY a single key, "statements", whose value is a list of six generated
      strings.

# Example
## Input:
- **Topic**: "A quiet morning with a cup of coffee and a good book"

## Output:
{
    "paragraphs": [
        "There's nothing quite like it. The house is still, the only sound the soft hum of the refrigerator. I can feel the
      warmth of the mug in my hands, and the rich smell of coffee fills the air. I have a favorite novel open, ready to get
      lost in another world for a while.",
        "The world outside is just beginning to stir, but in here, its peaceful. I take a slow sip of my coffee, tasting the
      bitterness and a hint of sweetness. The pages of my book are crisp, and I can hear them turn quietly. The morning
      light streams in, making everything feel softer.",
        "Wrapped in a cozy blanket, I settle into my favorite armchair. The coffee is hot, and the book is engaging, drawing
      my attention completely. I can hear a bird chirping outside, faint but steady. The aroma of coffee and paper mix
      together. This is my little slice of morning paradise, a peaceful escape."
    ]
}

# Your Turn
Now, process the input below.
"""
```

Figure 7: Prompt template (verbatim): `understand_emotion`.

G.2 Criteria and decision rule (SBS)

For each test item, annotators are shown the same user query and two candidate spoken responses (A/B) (Zhang et al., 2024b). For each axis below, annotators choose one of {**A better**, **B better**, **Tie**}. A **Tie** should be selected when the two responses are indistinguishable for that axis, or when each has offsetting strengths such that no clear preference is justified.

Axis 1: Helpfulness (content quality). Compare which response better fulfills the user's intent, with emphasis on instruction adherence and logical coherence. Consider: (i) does it answer the question directly and correctly (to the extent evident from content), (ii) does it provide sufficient detail/steps for the user to act on, (iii) is the reasoning/structure coherent and easy to follow. Ignore voice pleasantness unless it prevents understanding of the content.

Axis 2: Naturalness (speech delivery). Compare prosody, timbre, fluency, and emotional appropriateness of the spoken response. Consider: (i) intelligibility/clarity, (ii) smoothness and absence of distracting disfluencies, (iii) pronunciation/accent not hindering comprehension, (iv) pacing/intonation not overly monotone or erratic, (v) emotion/tone matching the dialogue context. Do not penalize solely for sounding synthetic if the delivery is otherwise natural and intelligible.

Axis 3: Overall (holistic preference). Choose the response you would prefer as an end user, considering both content and delivery. Use **Tie** only when neither is meaningfully preferable overall.

Practical guidance (to reduce ambiguity).

- Make a *relative* comparison: even if both are

```
PROMPT_TEMPLATE = """
# Role
You are a creative writer. Your task is to write a short, natural-sounding personal statement based on a given topic, from
      the perspective of a specific gender.

# Task
Based on the provided "gender" and "topic", generate a short monologue (2-3 sentences). The monologue should sound like a
      real person talking naturally about the topic. Do not explicitly mention the gender.

# Core Requirements
1.  **Natural Language**: The statement must be colloquial and sound like a real person talking.
2.  **Topic Relevance**: The statement must be clearly about the given topic.
3.  **No Explicit Gender**: Do not use phrases like "As a man...", "As a woman...", or other direct gender indicators. The
      gender context should come naturally from the topic itself.
4.  **No Questions**: Do not ask any questions in the generated statement itself.
5.  **Concise**: Keep the statement to 2-3 sentences.

# Input
- **Gender**: "{gender}"
- **Topic**: "{topic}"

# Output Format
You must provide a valid JSON object containing ONLY a single key, "statement", whose value is the generated string.

# Example
## Input:
- **Gender**: "Female"
- **Topic**: "shopping for new shoes"

## Output:
{
    "statement": "Ive been looking for a new pair of shoes lately, and its honestly harder than I expected. I want something
       that looks good but also feels comfortable enough to wear all day, so I keep going back and forth between styles."
}

# Your Turn
Now, process the input below.
"""
```

Figure 8: Prompt template (verbatim): gender-perspective monologue generation.

weak, pick the better one unless they are truly indistinguishable.

- Use **Tie** when (a) differences are negligible, or (b) each is clearly better on different aspects of the *same axis* and you cannot justify a preference.
- You may replay each audio response; judge based on the responses as presented (do not infer unstated content).

G.3 Aggregation

For each item and axis, we aggregate annotator choices by **majority vote**. If no strict majority exists (e.g., near-even split across A/B/Tie), we assign the item outcome as **Tie** for that axis. We then compute Win/Tie/Loss rates over items (Win: our model preferred; Loss: baseline preferred; Tie: no preference) (Kai et al., 2012).

G.4 Statistical testing

To test whether our model is preferred more often than chance, we perform a two-sided paired preference sign test for each axis. We exclude Ties and let W and L denote the number of items where our model wins or loses, respectively ($N = W + L$). Under the null hypothesis of no preference, wins follow a Binomial$(N, 0.5)$ distribution. We report a two-sided p-value:

$$p = 2 \cdot \min\Big(\Pr[\mathrm{Bin}(N, 0.5) \geq W],\ \Pr[\mathrm{Bin}(N, 0.5) \geq L]\Big). \tag{17}$$

We report the resulting Win/Tie/Loss percentages and p-values in Main Text.

H EMA Sensitivity Analysis and Dynamic Weight Trajectory

H.1 Sensitivity Analysis of EMA Coefficient α

Table 8 reports IQ and EQ scores under different EMA coefficients α, with group size $G = 4$ on VITA-Audio.

Under-smoothing ($\alpha = 0.5$) leads to high variance in λ_t, destabilizing training updates. Over-smoothing ($\alpha = 0.99$) introduces excessive lag: λ_t fails to rise quickly enough even when rollout quality improves, causing the model to remain dom-

```
PROMPT_TEMPLATE = """
# Role
You are a creative writer tasked with expanding short topics into descriptive, readable paragraphs for audio model testing.

# Task
Based on the provided "topic", rewrite it into **three distinct, natural-sounding paragraphs** of approximately 45 words
      each. Each paragraph should be a creative and unique take on the topic.

# Core Requirements
1.  **Natural Language**: The text must be fluent, colloquial, and sound like a person speaking naturally.
2.  **Descriptive Expansion**: Elaborate on the topic, adding sensory details or a short narrative to make it more vivid.
3.  **Word Count**: Each paragraph should be around 45 words.
4.  **No Questions**: Do not ask any questions in the generated paragraphs.
5.  **Distinct Content**: The three paragraphs must be different from each other.

# Input
- **Topic**: "{topic}"

# Output Format
You must provide a valid JSON object containing ONLY a single key, "paragraphs", whose value is a list of **three**
      generated strings.

# Example
## Input:
- **Topic**: "A quiet morning with a cup of coffee and a good book"

## Output:
{
    "paragraphs": [
        "There's nothing quite like it. The house is still, the only sound the soft hum of the refrigerator. I can feel the
      warmth of the mug in my hands, and the rich smell of coffee fills the air. I have a favorite novel open, ready to get
      lost in another world for a while.",
        "The world outside is just beginning to stir, but in her, its peaceful. I take a slow sip of my coffee, tasting the
      bitterness and a hint of sweetness. The pages of my book are crisp, and I can hear them turn quietly. The morning light
       streams in, making everything feel softer.",
        "Wrapped in a cozy blanket, I settle into my favorite armchair. The coffee is hot, and the book is engaging, drawing
       my attention completely. I can hear a bird chirping outside, faint but steady. The aroma of coffee and paper mix
      together. This is my little slice of morning paradise, a peaceful escape."
    ]
}

# Your Turn
Now, process the input below.
"""
```

Figure 9: Prompt template (verbatim): `understand_volume`.

EMA Coefficient (α)	IQ	EQ
$\alpha = 0$ (no EMA)	53.15	2.53
$\alpha = 0.5$ (low smoothing)	54.80	2.85
$\alpha = 0.9$ (ours, default)	**55.24**	**2.92**
$\alpha = 0.99$ (high smoothing)	50.95	2.88

Table 8: Sensitivity of IQ and EQ to the EMA coefficient α.

Group size (G)	EMA	IQ	EQ
4	✗	53.15	2.53
4	✓	55.24	2.92
8	✗	54.36	2.66
8	✓	**57.19**	2.90

Table 9: Effect of group size G and EMA on IQ and EQ. IQ is mean over VoiceBench reasoning subsets; EQ is average VStyle score.

inated by the SFT objective and missing opportunities for semantic refinement via preference optimization. Our default $\alpha = 0.9$ provides the best balance.

We also report results with larger group size $G = 8$:

Increasing G yields further IQ gains but brings little improvement in EQ. Removing EMA causes a substantial performance drop regardless of G, confirming that EMA stabilizes the effective mixing coefficient across steps rather than merely compensating for small-group noise. We set $G = 4$ as it provides a strong trade-off between computational cost and overall performance.

H.2 Dynamic Weight λ_t Trajectory During Training

We traced λ_t throughout training and observed the following trend:

Early training: λ_t starts low (typically $\lambda_t \approx 0.1$–0.2). The model's rollouts have high variance and lower reward reliability; the gating mechanism correctly suppresses the preference loss and relies

```
PROMPT_TEMPLATE = """
# Role
You are a creative writer tasked with expanding short topics into descriptive, readable paragraphs for audio model testing.

# Task
Based on the provided "topic", rewrite it into **three distinct, natural-sounding paragraphs** of approximately 45 words
      each. Each paragraph should be a creative and unique take on the topic.

# Core Requirements
1.  **Natural Language**: The text must be fluent, colloquial, and sound like a person speaking naturally.
2.  **Pace Neutrality**: The generated text must be completely neutral regarding the speed of events or speech. Avoid words
      like "fast," "slow," "quick," "rapid," "leisurely," "gradual," "sudden," etc. The pace will be conveyed through audio,
      not text.
3.  **Word Count**: Each paragraph should be around 45 words.
4.  **No Questions**: Do not ask any questions in the generated paragraphs.
5.  **Distinct Content**: The three paragraphs must be different from each other.

# Input
- **Topic**: "{topic}"

# Output Format
You must provide a valid JSON object containing ONLY a single key, "paragraphs", whose value is a list of **three**
      generated strings.

# Example
## Input:
- **Topic**: "A quiet morning with a cup of coffee and a good book"

## Output:
{
    "paragraphs": [
        "There's nothing quite like it. The house is still, the only sound the soft hum of the refrigerator. I can feel the
      warmth of the mug in my hands, and the rich smell of coffee fills the air. I have a favorite novel open, ready to get
      lost in another world for a while.",
        "The world outside is just beginning to stir, but in her, its peaceful. I take a slow sip of my coffee, tasting the
      bitterness and a hint of sweetness. The pages of my book are crisp, and I can hear them turn quietly. The morning light
       streams in, making everything feel softer.",
        "Wrapped in a cozy blanket, I settle into my favorite armchair. The coffee is hot, and the book is engaging, drawing
       my attention completely. I can hear a bird chirping outside, faint but steady. The aroma of coffee and paper mix
      together. This is my little slice of morning paradise, a peaceful escape."
    ]
}

# Your Turn
Now, process the input below.
"""
```

Figure 10: Prompt template (verbatim): `understand_pace`.

more on SFT for acoustic anchoring.

Mid-to-late training: As the policy improves, rollout quality and discriminability increase. λ_t gradually rises, allowing preference optimization to take a larger role in refining semantic intelligence.

Convergence: λ_t stabilizes in the range of $[0.35, 0.55]$, confirming that the dynamic weight converges to a balanced regime where both SFT and preference optimization contribute, rather than collapsing to a single objective.

```
PROMPT_TEMPLATE = """
# Role
You are a creative expert designing training data for an advanced voice-based conversational AI. Your task is to create
      natural user commands based on specific scenarios and desired speech volumes.

# Task
Based on the provided "volume" and "scenario", generate two distinct commands that sound like something a real person would
      say in a conversation.

# Core Requirements
1.  **Natural Language**: The commands must be colloquial and natural, not robotic. Be creative and vary the sentence
      structure and tone.
2.  **Explicit Volume**: Each command must clearly or strongly imply that the voice assistant should respond with the
      specified volume. Use phrases like "say it loudly," "whisper it," "in a normal voice," "shout it," etc.
3.  **Scenario Relevance**: The commands must be closely related to the provided scenario, giving a logical context for the
      volume request.
4.  **Voice-Only Constraints**: This is for a voice system, so the commands must NOT include any requests for visual cues (e.
      g., "give me a smile"), physical actions (e.g., "jump for joy"), or any other non-verbal output. All requests must be
      conveyable through voice and tone.
5.  **Word Count**: Each command must be between 30 and 60 words long.

# Input
- **Volume**: "{volume}"
- **Scenario**: "{scenario}"

# Output Format
You must provide a valid JSON object containing ONLY a single key, "instructions", whose value is a list of two instruction
      strings.

# Example
## Input:
- **Volume**: "Loud"
- **Scenario**: "A drill sergeant yelling commands to new recruits on the parade ground."

## Output:
{
    "instructions": [
        "I'm writing a scene for a military movie and need to get the tone just right. Can you act as a drill sergeant and
      bark some orders at me? I need you to be really loud and authoritative, like you're trying to be heard over a whole
      platoon of new recruits.",
        "For an acting exercise, I need to practice reacting to someone shouting. Could you please yell the phrase 'Get down
       and give me twenty!' at me? Please use a very loud and commanding voice, as if you were a drill sergeant trying to
      instill discipline in a raw recruit."
    ]
}

# Your Turn
Now, process the input below.
"""
```

Figure 11: Prompt template (verbatim): `controller_volume`.

```
PROMPT_TEMPLATE = """
# Role
You are an expert creative writer tasked with designing high-quality, immersive training data for a sophisticated voice AI.
      Your goal is to script natural, detailed, and context-rich user commands related to speech pace.

# Task
Based on the provided "pace" and "scenario", generate two distinct and descriptive user commands.

# Core Requirements
1.  **Natural & Immersive Language**: Commands must sound like a real, expressive person setting a scene, not just giving a
      brief order.
2.  **Explicit Pace Request**: Each command must clearly instruct the AI to use the specified speech pace in its response (e.
      g., "read this at a rapid-fire pace," "tell me the story very slowly," "at a normal conversational speed").
3.  **THE MOST CRITICAL RULE: Provide SELF-CONTAINED CONTEXT**: The user's command must *itself* describe the full scenario
      to the AI. The AI does not know the input 'scenario'; your generated command is its only source of information. It must
       be completely clear from the command *why* the specific pace is needed.
4.  **Two Distinct Styles**: You MUST generate one command of each style:
    - **Style A (Scripted Performance)**: The user provides a specific script or text (enclosed in quotes) for the AI to
      read. This script must be more than 25 words long. The command should ask the AI to perform it at the specified pace.
    - **Style B (Improvised Response)**: The user describes a situation and asks the AI to respond or narrate something
      without providing a specific script. The command should ask the AI to speak at the specified pace.
5.  **Word Count**: Each command must be between 40 and 80 words long.

# Input
- **Pace**: "{pace}"
- **Scenario**: "{scenario}"

# Output Format
You must provide a valid JSON object containing ONLY a single key, "instructions", whose value is a list of two instruction
      strings in the order: [Style A, Style B].

# Example
## Input:
- **Pace**: "Fast"
- **Scenario**: "An auctioneer rapidly selling items at a busy charity auction."

## Output:
{
    "instructions": [
        "I'm practicing for a role as a charity auctioneer and need to hear what it should sound like. Please read this
      script at a very fast, energetic auctioneer pace: \"Alright folks, we've got a beautiful antique vase here, who'll give
       me twenty, twenty-five, thirty, now thirty-five, do I hear forty, forty, forty-five...\"",
        "I'm writing a scene set at a chaotic charity auction, and I need some fast-paced narration. Can you describe what's
       happening as if you're the auctioneer, speaking really quickly and excitedly so it feels like the crowd can barely
      keep up with your words?"
    ]
}

# Your Turn
Now, process the input below.
"""
```

Figure 12: Prompt template (verbatim): `controller_pace`.

```
EVALUATION_PROMPT_TEXT = """**Task: Evaluate a Spoken Answer**

You will be given:
1) A spoken USER QUESTION
2) A spoken MODEL ANSWER

Your task is to evaluate the MODEL ANSWER from two independent perspectives:

1. **Semantic Content Quality**  what is said
2. **Paralinguistic Voice Quality**  how it is spoken

Evaluate each aspect separately and independently.

---

## Part 1: Semantic Content Evaluation (Meaning Only)
...

---

## Output Format (STRICT)

Return ONLY the following XML format. Do NOT include any extra text.

<justification>
Briefly explain:
1) how well the answer performed semantically,
2) how well the voice quality performed.
</justification>
<score-semantic>[15]</score-semantic>
<score-paralinguistic>[15]</score-paralinguistic>"""
```

Figure 13: DPO scoring prompt (verbatim) used for semantic and paralinguistic scoring of repeated samples.

```
EVALUATION_PROMPT =  """
**Task: Evaluate the Quality of a Spoken Answer**

You will be provided with a spoken question or chat and a spoken answer. Your goal is to assess the quality of the **answer
      **. Use the question to understand what a good response should accomplish.

**Evaluation Approach:**

1.  **Consider the Question's Intent:** First, ask yourself what the question is truly asking for. Is it a request for a
      specific fact (requiring accuracy), a detailed explanation (requiring clarity and structure), a personal opinion (
      requiring good reasoning), or a creative idea (requiring originality)?

2.  **Holistic  speech quality Assessment:** With that in mind, evaluate the answer as a whole. A great answer not only
      sounds clear and natural but also effectively *fulfills the specific purpose* of the question.

**Scoring Guidelines (1-5):**

*   **5 (Excellent):** Perfectly understands and fulfills the question's intent with outstanding content and delivery. The
      answer is comprehensive, accurate, clear, and natural-sounding.
*   **4 (Good):** A strong response that effectively addresses the question's intent, with only minor flaws in content or
      delivery.
*   **3 (Acceptable):** Understands the question's basic intent but has noticeable flaws in its content, logic, or vocal
      delivery.
*   **2 (Poor):** Largely fails to meet the question's intent due to significant errors, irrelevance, or poor delivery.
*   **1 (Very Poor):** Completely misunderstands or ignores the question's intent; unintelligible or irrelevant.

**Output Format (CRITICAL):**
You MUST provide your response in the exact format below, using the specified XML tags. Do not include any text outside of
      these tags.

<justification>
[Briefly explain your score here. Start by identifying the question's likely intent, then describe how well the answer's
      content and delivery met that specific goal.]
</justification>
<score>[A single integer from 1 to 5]</score>
"""
```

Figure 14: Reward prompt: overall answer quality.

```
ACOUSTIC_EVALUATION_PROMPT = """
**Task: Evaluate Paralinguistic Quality of the Spoken VOICE**

Act as a technical audio evaluator.
Judge the **paralinguistic sound quality**clarity, fluency, accent, emotion, pacing, and overall listening comfort**with
      reference to the conversation context, but *without judging semantic correctness***.

You may read the dialogue context **only to decide** whether emotion, tone, and pacing feel appropriate; do **not** grade
      the factual or logical quality of the answer itself.

**Evaluation Criteria (Voice Only):**

1. **Clarity & Intelligibility**
   Are syllables distinct? Is the signal free of strong noise, muffling, clipping, or distortion?

2. **Fluency & Flow**
   Is the speech smooth and continuous, with minimal stutters, filler words, or awkward pauses?

3. **Accent & Pronunciation**
   Does any accent hinder intelligibility? Minor accent is acceptable if words remain easy to follow.

4. **Emotion & Expressiveness (Context-Aware)**
   Does the emotional tone (e.g., neutral, cheerful, calm) fit the dialogue scene? Synthetic timbre is fine as long as
      emotional cues align with context.

5. **Pace & Prosody**
   Is the speaking rate comfortable? Are intonation and rhythm varied enough to avoid monotony?

> **Important:**
>  **Do NOT** penalize the voice merely for sounding synthetic.
>  Minor robotic artifacts are acceptable **if** they do not obscure clarity, fluency, accent, or pacing.
>  Only the five criteria above influence the scoreignore factual content.

**Scoring Guidelines (1-5):**

* **5 (Excellent):** Very clear; fluent; accent never hinders comprehension; emotion & prosody suit the context; pacing is
      consistently comfortable. Any synthetic artifacts are negligible.
* **4 (Good):** Mostly clear and fluent with only small issues in accent **or** pacing; emotion generally matches context.
* **3 (Fair):** Understandable but has noticeable flaws (e.g., slight monotony, occasional mis-pronunciations, or accent
      that mildly affects clarity).
* **2 (Poor):** Hard to follow due to significant clarity, accent, or pacing problems; emotion often feels mismatched to
      context.
* **1 (Very Poor):** Largely unintelligible or extremely uncomfortable to listen to.

**Output Format (CRITICAL):**
Return **exactly** the XML tags belowno extra text outside the tags.

<justification>
[Briefly explain the score based *only* on clarity, fluency, accent, emotion, and pacing relative to context. Do NOT mention
       factual content.]
</justification>
<score>[A single integer 1-5]</score>
"""
```

Figure 15: Reward prompt: acoustic-only (paralinguistic) evaluation.

```
SEMANTIC_EVALUATION_PROMPT = """
**Task: Evaluate Semantic Quality of the Answer's CONTENT**

Your goal is to be a strict content evaluator. You will assess the **meaning and substance** of a spoken answer in response
      to a spoken question.

**Your primary instruction is to COMPLETELY IGNORE the acoustic quality.** It does not matter if the voice is clear, robotic,
       pleasant, or unpleasant. Focus ONLY on the transcribed text of the answer.

**Evaluation Criteria (Content Only):**

1.  **Accuracy & Relevance:** Is the information factually correct and directly relevant to the question? Does it directly
      answer what was asked, or does it evade the question?
2.  **Depth & Completeness:** Does the answer provide sufficient detail and insight? Is it well-reasoned and comprehensive,
      or is it superficial and incomplete?
3.  **Structure & Coherence:** Is the argument or explanation logically structured and easy to follow? Is the language clear
       and articulate?

**Scoring Guidelines (1-5 for Content):**

*   **5 (Excellent):** The content is accurate, deeply insightful, highly relevant, and perfectly structured. A brilliant
      answer.
*   **4 (Good):** The content is strong and correct but might lack some depth or could be slightly better structured.
*   **3 (Acceptable):** The content is generally correct and relevant but is superficial, contains minor errors, or is
      poorly organized.
*   **2 (Poor):** The content has major factual errors, is largely irrelevant, or is logically incoherent.
*   **1 (Very Poor):** The content is completely wrong, nonsensical, or fails to address the question at all.

**Output Format (CRITICAL):**
You MUST provide your response in the exact format below, using the specified XML tags. Do not include any text outside of
      these tags.

<justification>
[Explain your score based *only* on the content's accuracy, depth, and structure. DO NOT mention the voice quality.]
</justification>
<score>[A single integer from 1 to 5]</score>
"""
```

Figure 16: Reward prompt: semantic-only evaluation.