SILENT TAGS: RESPONSE-CONDITIONED EXPRESSIVE SPEECH WITH A FROZEN LANGUAGE BACKBONE

Anonymous authors
Paper under double-blind review

ABSTRACT

With the rapid development of omni large language models, users can converse directly with assistants through speech. However, most existing models render every response in the same flat prosody, regardless of whether the user is distressed, excited, or amused, limiting the user experience in human-machine interaction. To inject expressiveness, the prevailing approach trains the language backbone, namely the Thinker, to jointly predict paralinguistic tags alongside text. We find that this consistently degrades reasoning on benchmarks such as MMLU and VoiceBench. A key challenge therefore lies in how to make a speech assistant expressive without compromising the reasoning ability of its language backbone. To address this challenge, we propose Silent-Tags, a Thinker-agnostic framework that leaves the Thinker entirely frozen and trains only the Talker. Our key insight is that a well-trained omni Thinker already encodes, in its hidden states, contextual signals required for emotional and paralinguistic inference. Silent-Tags equips the Talker with lightweight cross-attention layers that selectively tap these hidden states for rich emotional and paralinguistic cues, supervised by a tag-controllable TTS teacher together with a segment-level auxiliary classifier. Silent-Tags introduces no extra Thinker decoding steps and no tag-token generation at inference time; the only added overhead is the Talker's cross-attention into already-cached Thinker states. On ParaS2SBench and EmergentTTS-Eval, Silent-Tags consistently outperforms explicit-paradigm baselines on expressiveness. Code and inference examples can be found at `https://anonymous-3911.pages.dev/`.

1 INTRODUCTION

Speech is the most direct channel of human interaction: words convey what we mean, while paralinguistic cues such as emotion, laughter, hesitation, speaking rate, and vocal style convey how we mean it (Reece et al., 2023). Recent progress in omni-modal speech-to-speech (S2S) systems has made natural, end-to-end spoken interaction increasingly practical. Motivated by the shift toward real-time omni-modal interaction exemplified by GPT-4o (OpenAI, 2024), recent open-source systems such as Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), MiniCPM-o 4.5 (Yao et al., 2024), and Freeze-Omni (Wang et al., 2024) increasingly adopt a *Thinker–Talker* architecture: a language-centric Thinker interprets the user input and plans the response, while a Talker renders that response into speech. Yet current omni models remain largely **insensitive to the paralinguistic context** of the user input. For example, a response meant to console a distressed user may be rendered with nearly the same prosody as one meant to celebrate good news. Recent benchmarks make this gap quantitative. On ParaS2SBench (Yang et al., 2025), leading S2S models perform on par with a cascaded Whisper-LLM-TTS pipeline that largely discards non-verbal acoustic cues from the user input; on EmergentTTS-Eval (Manku et al., 2025), they trail far behind tag-controllable TTS systems on prosody, irony, whispering, and laughter. Together, these results suggest that current S2S systems can speak fluently, but still struggle to speak appropriately.

In contrast, modern TTS systems offer fine-grained control over expressive synthesis: CosyVoice 2 (Du et al., 2024), Step-Audio-EditX (Yan et al., 2025), and Fish-Speech (Liao et al., 2024) accept inline tags, emotion labels, and natural-language style instructions, controlling dozens of emotions and speaking styles with high fidelity. Existing approaches to expressive omni-modal speech include language-backbone adaptation and **explicit-tag generation**; these are related but distinct design choices. Empathy Omni (Wang et al., 2025) attaches a token-synchronous emotion

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.

trajectory via LoRA (Hu et al., 2022); OpenOmni (Luo et al., 2025) applies DPO (Rafailov et al., 2023) on the Thinker for emotional alignment; recent SFT and RL recipes for Kimi-Audio (Ding et al., 2025) train the Thinker to emit explicit paralinguistic tag tokens before each response segment. These approaches couple expressive adaptation to updates of the language backbone. In our evaluated explicit-tag fine-tuning settings, expressive gains are accompanied by regressions on MMLU (Hendrycks et al., 2021) and VoiceBench (Chen et al., 2026). This motivates separating speech rendering from language-model adaptation. Competition between training objectives is one possible explanation for the regressions, rather than a mechanism established by these scores alone.

A simpler alternative is to keep the Thinker frozen and fine-tune its Talker on expressive speech. This is an important control: stronger teacher-generated data may improve delivery even without a new conditioning mechanism. Our same-data Talker baseline confirms that it does, but leaves a gap to ***Silent Tags*** under a common reconstructed ParaS2SBench protocol (Section 4.2). Direct source-audio affect conditioning is another meaningful alternative, since it can learn how to respond to a user's emotion rather than merely copying it. The distinction we study is *response-conditioned control*: for the same source affect, different replies may require an apologetic, concerned, or firm delivery. We therefore ask whether reading the frozen Thinker's response representations provides useful control beyond expressive data and source affect alone.

Layer-wise probing studies show that the accessibility of linguistic and acoustic information varies with depth (Pasad et al., 2021; Jawahar et al., 2019). This motivates examining intermediate representations rather than treating the final language-model state as the only useful speech-conditioning signal. Our central hypothesis is that a pre-trained omni Thinker's response representations contain useful contextual signals for paralinguistic rendering. These signals need not match the surface sentiment of individual words, as in sarcasm or reassurance. We train the Talker to read them, without assuming that this readout can repair a pragmatic interpretation that the frozen Thinker failed to encode.

Silent Tags is built around this hypothesis. We first use a strong LLM to split each dialogue response into clause-level segments and assign each segment a contextually appropriate paralinguistic tag. CosyVoice 2 (Du et al., 2024) renders the tag schedule with inline control tokens. Mixed-tone responses such as *soothing* followed by *encouraging*, or *hesitant* followed by *confident*, become explicit training targets. A multi-stage filter removes inconsistent annotations, failed tag realisations, low-quality audio, and over-represented tags, retaining roughly 40% of raw samples and yielding a $\sim$90k-pair expressive S2S corpus covering 48 paralinguistic categories.

We then train **only the Talker**. Additional cross-attention layers read selected lower-layer hidden states from the frozen Thinker, while per-segment distillation and auxiliary classification losses align the resulting control vectors with teacher-derived tag embeddings. At inference, both the auxiliary classifier and explicit tag inputs are removed: the Talker conditions directly on the Thinker's hidden states and generates expressive speech without emitting tag tokens. We call each such cross-attention readout a **silent tag**: a continuous hidden-space signal that serves the role of an explicit paralinguistic tag without ever appearing in the token sequence.

Silent tags are recomputed at every Talker decoding step from the same frozen Thinker cache. Because the Talker's query state evolves across codec steps, the architecture permits control to vary within a reply without inline tag generation. The intended behavior is, for example, to render a reply to *"I just lost my job, but I have an interview tomorrow"* with a soothing opening followed by encouragement. Whether the generated waveform actually realizes and sustains such a transition is an empirical question, not a guarantee of cross-attention alone. We test it by holding source audio fixed while varying the response trajectory and evaluating the resulting waveforms.

Contributions.

- We propose ***Silent Tags***, a Thinker-frozen framework that distils explicit paralinguistic supervision into a Talker-side, response-conditioned readout. It leaves the Thinker's parameters and decoding path unchanged, requires no explicit tag-token generation, and discards the auxiliary classifier at inference; cross-attention adds Talker-side computation.
- We construct a paralinguistically-annotated S2S corpus tailored to this framework: $\sim$90k high-quality context–response pairs covering 48 paralinguistic categories, built by combining LLM-

based segment-level tagging with tag-controllable TTS synthesis and a four-stage filter that retains roughly 40% of raw synthesised samples.

- We evaluate expressiveness under a common reconstructed ParaS2SBench protocol with same-data Talker and source-affect controls. A fixed-source, varying-response test additionally probes whether delivery follows the generated reply, with independent waveform-level measurements of style realization, switching, and persistence.

2 Related Work

End-to-end omni-modal speech models. Speech-interaction systems use different couplings between language modeling and audio generation. GPT-4o (OpenAI, 2024) demonstrated that a single autoregressive backbone can interleave text and audio tokens in real time. Open-source successors such as Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), Kimi-Audio (Ding et al., 2025), GLM-4-Voice (Zeng et al., 2024), LLaMA-Omni 2 (Fang et al., 2025), and Freeze-Omni (Wang et al., 2024) differ in how aggressively the Talker shares parameters with the Thinker, but they all rely on the LLM backbone to plan textual content and on a speech head or decoder to render audio. Related speech and audio language models include Mini-Omni (Xie & Wu, 2024), Moshi (Défossez et al., 2024), SpeechGPT (Zhang et al., 2023), AudioPaLM (Rubenstein et al., 2023), VITA (Fu et al., 2024), and LLaMA-Omni (Fang et al., 2024); their streaming capabilities and module boundaries differ. With few exceptions, these models are trained on speech data in which the response prosody is incidental rather than contextually controlled, which is one source of the paralinguistic gap quantified on ParaS2SBench (Yang et al., 2025). ***Silent Tags*** is orthogonal to this architectural family: it leaves the Thinker unchanged and augments only the Talker.

Expressive and tag-controllable TTS. Modern TTS systems support fine-grained paralinguistic control through emotion labels, style tokens, inline tags, or natural-language instructions. CosyVoice 2 (Du et al., 2024) and Fish-Speech (Liao et al., 2024) expose tag-based and instruction-based interfaces over a streaming LLM-based decoder; Step-Audio-EditX (Yan et al., 2025) extends this to over a dozen emotions, 32 speaking styles, and inline event tags inserted at arbitrary text positions; instruction-following TTS lines such as InstructTTS (Yang et al., 2024) and Voicebox (Le et al., 2023) similarly accept free-form descriptions. Foundational expressive synthesis works, including VALL-E (Chen et al., 2024), NaturalSpeech 2 (Shen et al., 2024), MegaTTS (Jiang et al., 2023), Seed-TTS (Anastassiou et al., 2024), and SpearTTS (Kharitonov et al., 2023), further show that high-quality and controllable expressiveness is largely tractable in the standalone TTS setting. We exploit these systems as *teachers*: a tag-controllable TTS synthesises expressive audio for paralinguistically-annotated dialogue, and we distil the underlying control signal into the Talker rather than asking the Thinker to learn it.

Paralinguistic control in omni-modal speech. A growing body of work brings expressive control into omni-modal speech models. Language-backbone adaptation and explicit tag-token generation are distinct choices, rather than interchangeable descriptions of this literature. Empathy Omni (Wang et al., 2025) augments the language backbone with a token-synchronous emotion trajectory and trains the Thinker via LoRA (Hu et al., 2022). OpenOmni (Luo et al., 2025) applies DPO (Rafailov et al., 2023) on the Thinker to align emotional outputs with human preferences. Recent SFT and GRPO recipes for Kimi-Audio (Ding et al., 2025) train the Thinker to emit explicit paralinguistic tag tokens before each response segment, yielding strong gains on ParaS2SBench but at the cost of measurable reasoning regressions.

3 Method

3.1 Overview

We introduce ***Silent Tags***, a Thinker-frozen framework that endows omni-modal speech models with paralinguistic expressiveness without modifying the language backbone. As shown in Figure 1, ***Silent Tags*** consists of three ingredients: (i) a paralinguistically annotated speech-to-speech corpus (§3.2); (ii) a Talker-side cross-attention mechanism that reads from selected lower-layer Thinker representations (§3.3); and (iii) a Talker-only objective combining speech-generation likelihood with segment-level distillation and an auxiliary paralinguistic supervision signal (§3.4).

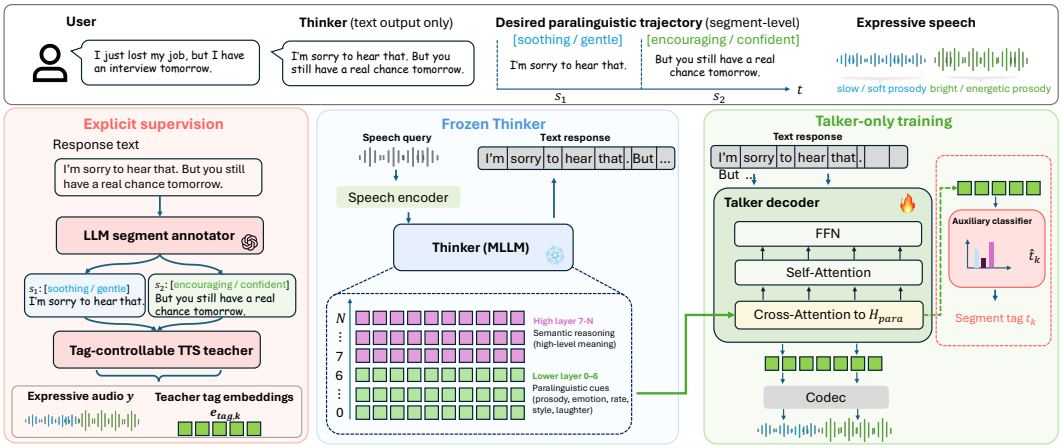


Figure 1: Overview of ***Silent Tags***. The Thinker remains frozen; the Talker learns a readout from selected shallow states without inference-time tag generation. The supervision panel is schematic: CosyVoice 2 synthesizes waveforms and a separate Fish Audio S2 encoder extracts style targets. Layer annotations illustrate the readout motivation, not an exclusive separation of semantic and paralinguistic information.

Silent tag (definition). A *silent tag* $\mathbf{z}_s^{(i)}$ is a Talker block's cross-attention readout from frozen Thinker response states at codec step s: a continuous control signal that performs the functional role of an explicit paralinguistic tag while (i) never being emitted as a token, (ii) never altering the Thinker's decoding path, and (iii) requiring no auxiliary tag-prediction head at inference. All gradients flow only into the Talker.

3.2 Paralinguistically-Annotated Data Construction

Since existing S2S corpora are rarely annotated for dialogue-grounded prosody, we construct a paralinguistically annotated corpus via LLM-based segment tagging, tag-controllable TTS synthesis, and multi-stage audio–tag consistency filtering. The resulting dataset contains ∼90k high-quality context–response pairs across 48 paralinguistic categories and supports all experiments in Section 4.

Step 1: LLM-based segment-level tag annotation. Given a conversational corpus of (context, response) pairs, we prompt the GPT-5.1 API (OpenAI, 2025) to split each response into clause-level segments and to assign each segment a tag $t_k \in \mathcal{T}$, where $\mathcal{T}$ covers (i) emotion (e.g., *sad*, *joyful*, *angry*, *anxious*), (ii) speaking style (e.g., *gentle*, *excited*, *whisper*), and (iii) social or vocal events (e.g., *[Laughter]*, *[Sigh]*, *[Hesitation]*). The prompt explicitly allows adjacent segments to differ when the dialogue calls for it (for example, opening with *soothing* and pivoting to *encouraging*), so that responses with mixed paralinguistic content are first-class training targets rather than edge cases. The output is a tag schedule $\mathcal{S}_i = ((b_1, t_1), (b_2, t_2), \ldots, (b_{K_i}, t_{K_i}))$, where b_k is the token offset at which segment k begins. This yields the intermediate corpus $\mathcal{D} = \{(c_i, r_i, \mathcal{S}_i)\}_{i=1}^{N}$.

Step 2: Tag-controllable TTS synthesis. We render each annotated response with CosyVoice 2 (Du et al., 2024), splicing the tag schedule into the prompt as inline tokens of the form $[t_1]\, r_{[b_1:b_2)}\, [t_2]\, r_{[b_2:b_3)}\, \ldots$, which returns an expressive waveform $\mathbf{a}_i$. The reported teacher-alignment grid has frame rate $f_{\text{codec}} = 25$ Hz; text offsets b_k map to frame indices $s_k = \lfloor \tau_{b_k} \cdot f_{\text{codec}} \rfloor$, where τ_{b_k} is the aligned token start time. This grid does not specify each backbone's native codec-token mapping (Appendix **??**). Per-segment tag embeddings $\mathbf{e}_{\text{tag},k} \in \mathbb{R}^{d_{\text{tag}}}$ ($d_{\text{tag}} = 512$) are extracted by a separate Fish Audio S2 (Liao et al., 2024) style encoder applied to $\mathbf{a}_i$, decoupling the distillation target from the synthesiser's own embedding space.

Step 3: Multi-stage quality filtering. The raw output of Steps 1 and 2 is noisy: occasional context-inappropriate tags, TTS realisation failures, and waveform artefacts each become label-noise gradients that bias the silent-tag readout. We therefore apply a four-stage filter combining annota-

tion self-consistency, tag–audio consistency via an independent paralinguistic classifier, acoustic-quality scoring with UTMOS (Saeki et al., 2022) and a Whisper-based (Radford et al., 2023) ASR transcript-recovery check, and tag-distribution balancing; the filtering stages and corpus counts are summarized in Appendix B.2. After filtering, roughly 40% of raw synthesised samples survive, yielding a final corpus of $\sim$90k high-quality (context, expressive response) pairs:

$$\mathcal{D} = \big\{(c_i,\, r_i,\, \mathcal{S}_i,\, \mathbf{a}_i,\, \{\mathbf{e}_{\text{tag},i,k}\}_{k=1}^{K_i})\big\}_{i=1}^{N}, \quad N \approx 9{\times}10^4.$$

This corpus provides supervision at three granularities: discrete segment tags, dense per-segment style embeddings, and acoustic waveforms, all of which we use jointly during training.

3.3 Model Architecture

Notation. T: Thinker text length (full response). S: codec sequence length. K: clause-level segments per response. $|\mathcal{T}| = 48$: tag vocabulary. d, d': Thinker and Talker hidden sizes. L: total number of Thinker layers; L_{para}: selected shallow-layer cutoff (set to 6 below). $d_{\text{tag}} = 512$: teacher style-embedding dimension (Sec. 3.2).

Thinker (frozen). The Thinker is a pre-trained omni-modal LLM that consumes the user's speech and autoregressively generates a textual response; we evaluate Qwen2.5-Omni (Xu et al., 2025a), Kimi-Audio (Ding et al., 2025), and Qwen3-Omni (Xu et al., 2025b) backbones. For an input query $\mathbf{x}$, it produces a stack of hidden states $\mathbf{H}_{\text{thinker}} = \{\mathbf{H}_{1:T}^{(l)}\}_{l=0}^{L}$ with $\mathbf{H}_{1:T}^{(l)} \in \mathbb{R}^{T\times d}$. Decoding proceeds in two phases: in **Phase 1** the Thinker generates the full response r of length T; in **Phase 2** the Talker autoregressively emits codec tokens, using the cached Thinker states as static memory for Talker-side key/value projections. All Thinker parameters are *frozen* throughout.

Layer selection. Layer-wise probing studies motivate examining which hidden states expose useful signals (Jawahar et al., 2019; Pasad et al., 2021). On frozen Qwen2.5-Omni states, our probes predict Laughter, Sigh, and Hesitation (Figure 2). Their averaged F1 peaks near layer 4 at 78.7%, compared with the external Whisper-Large-v3 reference of 73.4%, and falls below this reference beyond layer 6. We use layers 0–6 as an empirical readout range, denoted $\mathbf{H}_{\text{para}} = \{\mathbf{H}_{1:T}^{(l)}\}_{l=0}^{L_{\text{para}}}$. This notation identifies selected states, not a universal partition into paralinguistic and semantic layers. The three-event probe demonstrates decodability in this setting; it neither establishes coverage of all 48 tags nor isolates acoustic cues from correlated response-text information. The reference score motivates the cutoff but does not prove it optimal for generation. Appendix E.2 reports earlier text, upper-layer, and shuffled-state controls; their interpretation does not require a universal layer partition.

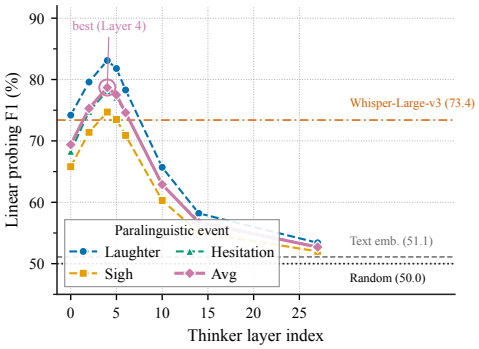


Figure 2: Per-layer linear-probe F1 (%) on three paralinguistic events (*Laughter*, *Sigh*, *Hesitation*) across the Thinker stack. The horizontal line marks Whisper-Large-v3 as an external acoustic-model reference.

Talker with cross-attention. The Talker is an autoregressive language model that emits discrete audio codec tokens conditioned on the Thinker's text and on paralinguistic features read from its lower hidden states. Let $\mathbf{Q}_s^{(i)} \in \mathbb{R}^{d'}$ denote the query in Talker block i at codec step s. The depth-paired readout is

$$\mathbf{z}_s^{(i)} = \text{CrossAttn}_i\big(\mathbf{Q}_s^{(i)},\, \mathbf{H}_{1:T}^{(i)},\, \mathbf{H}_{1:T}^{(i)}\big), \quad i = 0, \ldots, 6.$$

Write $\mathbf{Z}_s = \{\mathbf{z}_s^{(i)}\}_{i=0}^{6}$ for the collection of readouts. Although the memory is static during Phase 2, each $\mathbf{Q}_s^{(i)}$ evolves with the Talker's autoregressive state; the same Thinker cache can therefore be re-attended with different weights across codec steps, permitting within-response changes without tag-token signaling. The Talker predicts codec tokens, which are decoded into a waveform by the audio decoder.

Cross-attention placement. We add a cross-attention sub-block to each of the bottom $L_{\text{para}}+1 = 7$ Talker layers, with a one-to-one depth mapping: Talker layer i attends to Thinker layer i for $i \in \{0, \ldots, L_{\text{para}}\}$. This supplies a depth-paired conditioning path. The added Q/K/V projections are randomly initialized and do not reuse the Thinker's attention projections; gradients update only Talker-side parameters.

3.4 Training Objective

Only Talker-side parameters are updated. Let $\bar{\mathbf{z}}_k$ be the mean of the silent tags within the codec span of segment k. The Talker is trained with $\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_1 \mathcal{L}_{\text{distill}} + \lambda_2 \mathcal{L}_{\text{aux}}$, where

$$\mathcal{L}_{\text{gen}} = -\sum_{s=1}^{S} \log p_\theta\big(y_s \mid y_{<s}, \mathbf{H}^{(L)}_{1:T}, \mathbf{Z}_s\big), \tag{1}$$

$$\mathcal{L}_{\text{distill}} = \frac{1}{K}\sum_{k=1}^{K} \big\| g(\bar{\mathbf{z}}_k) - \mathbf{e}_{\text{tag},k} \big\|_2^2, \tag{2}$$

$$\mathcal{L}_{\text{aux}} = -\frac{1}{K}\sum_{k=1}^{K} \log p_a(t_k \mid \bar{\mathbf{z}}_k). \tag{3}$$

$\mathcal{L}_{\text{gen}}$ is standard next-token prediction over the codec sequence $\mathbf{y}$, conditioned on the Thinker's top-layer states $\mathbf{H}^{(L)}_{1:T}$ as in the baseline Talker. $\mathcal{L}_{\text{distill}}$ aligns each segment's pooled control, projected by a 2-layer MLP g, with the teacher tag embedding. $\mathcal{L}_{\text{aux}}$ trains a linear classifier p_a over the tag set $\mathcal{T}$ and is discarded at inference. The two segment losses supervise segment averages rather than individual codec steps. λ_1 and λ_2 are tuned on a held-out development split.

4 Experiments

We ask whether Talker-only adaptation achieves competitive expressiveness (Q1), how it compares with Thinker-updating recipes on text-side reasoning (Q2), and whether the readout provides response-conditioned control beyond expressive training data and source affect (Q3).

4.1 Experimental Setup

Models and training. We instantiate ***Silent Tags*** on Qwen2.5-Omni (Xu et al., 2025a) and Kimi-Audio (Ding et al., 2025), and include Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) in the main comparison. Only Talker-side parameters receive gradients. Training uses the approximately 90k expressive pairs described in Section 3.2, with CosyVoice 2 synthesis and separate Fish Audio S2 embedding targets. The reported main recipe uses AdamW, 30k steps, learning rate 2×10^{-4}, batch size 256, and loss weights $\lambda_1 = 0.5$, $\lambda_2 = 0.2$. Data-stage counts and training details are collected in Appendix B.

Evaluation protocol. We use *our reconstructed ParaS2SBench*, not the original unreleased test set (Yang et al., 2025). The reconstruction contains 2,690 prompts across Emotion, Sarcasm, Age, Gender, IEMOCAP, and MELD. All systems are evaluated on the same reconstructed inputs, newly evaluated Age subset, and frozen judging pipeline, using Qwen3-32B-Instruct instead of the original ChatGPT-4.1 text judge. No original-paper baseline scores are reused. Overall is the equal-weight mean of the six category/corpus scores on a 1–5 scale. The Age evaluation measures adult/child input conditions, not speaking rate. Appendix A details benchmark composition and aggregation. We additionally report EmergentTTS-Eval (Manku et al., 2025) pairwise win rates and Thinker text-side results on VoiceBench (Chen et al., 2026) and MMLU (Hendrycks et al., 2021).

Matched controls. *Same-data Talker FT* uses the expressive corpus without the added lower-layer Thinker readout. *Source-affect conditioning* uses the same 90k corpus, Talker, training objective, and optimization budget as ***Silent Tags***, but replaces the readout with an utterance-level affect embedding from source audio. It still receives the generated response text, so it can learn an empathetic input-to-output mapping rather than merely copying input affect. Explicit-tag SFT and GRPO

Table 1: **Common-protocol expressiveness** on our reconstructed ParaS2SBench (Overall, 1–5). All methods use the same reconstructed evaluation and Qwen3-32B-Instruct judge. "Update" and "Tags" denote Thinker fine-tuning and inference-time tag generation. Dashes indicate unreported configurations, not zero scores. Full category scores and additional baselines are in Table 6.

Model	Synthetic Speech				Real Speech		Avg
	Emotion	Sarcasm	Age	Gender	IEMOCAP	MELD	
Pipeline Baseline							
Whisper-GPT-TTS	3.126	2.951	3.093	2.884	3.574	3.372	3.167
Closed-Source Models							
GPT-4o-2024-11-20	3.712	2.978	3.301	3.391	3.851	3.520	3.459
Gemini-2.5-flash	3.863	3.248	3.391	3.437	3.865	3.709	3.586
Open-Source S2S Models (without paralinguistic control)							
Qwen2.5-Omni	3.656	2.904	3.221	3.216	3.642	3.584	3.370
Qwen3-Omni-30B-A3B	3.902	3.351	3.711	3.517	3.894	3.712	3.681
Kimi-Audio	2.701	2.629	3.203	3.090	1.392	1.153	2.361
Talker-side Controls (frozen Thinker, same data)							
Qwen2.5 Talker FT	3.886	3.378	3.693	3.518	3.827	3.645	3.658
Qwen2.5 Source-affect	4.041	3.602	3.853	3.640	4.040	3.736	3.819
Kimi Talker FT	3.881	3.469	3.883	3.590	3.872	3.533	3.705
Kimi Source-affect	4.032	3.819	4.015	3.732	4.033	3.611	3.874
Open-Source S2S Models (with explicit paralinguistic control)†							
Qwen2.5 Explicit-tag SFT	4.218	4.126	4.302	3.917	4.208	3.912	4.114
Qwen3 Explicit-tag SFT	4.492	4.401	<u>4.583</u>	4.219	<u>4.451</u>	**4.238**	<u>4.397</u>
Kimi-Audio SFT	4.121	4.261	4.461	3.528	4.162	3.284	3.969
Kimi-Audio GRPO	<u>4.516</u>	**4.506**	4.555	<u>4.228</u>	4.426	3.899	4.355
Ours (frozen Thinker, Talker-only training)							
Silent Tags (Qwen2.5)	4.236	4.079	4.261	3.998	4.183	3.931	4.115
Silent Tags (Qwen3)	**4.548**	<u>4.438</u>	**4.607**	**4.286**	**4.493**	<u>4.211</u>	**4.430**
Silent Tags (Kimi)	4.284	4.112	4.315	4.036	4.217	3.689	4.109
Topline (oracle labels)							
GPT-TTS (w/ GT labels)	4.718	4.760	4.591	4.806	4.747	4.798	4.737

update the Thinker and generate paralinguistic tags at inference. Table 1 compares these choices within each backbone; full category scores, additional baselines, and the oracle-label reference are in Appendix C.

4.2 Main Results

Competitive expressiveness without Thinker updates (Q1). On the reconstructed evaluation, ***Silent Tags*** achieves Overall scores of 4.115, 4.109, and 4.430 on Qwen2.5-Omni, Kimi-Audio, and Qwen3-Omni. Its Qwen2.5 point estimate is essentially equal to explicit-tag SFT (4.114); on Kimi, it exceeds SFT (3.969) but remains below GRPO (4.355); on Qwen3, it is slightly higher than SFT (4.397). These results support competitive expressiveness, not universal superiority over explicit control. Small point-estimate differences should not be interpreted as statistically established improvements.

On EmergentTTS-Eval, ***Silent Tags*** achieves 62.3% on Qwen2.5 and 65.7% on Kimi. The Kimi variant exceeds the reported tag-prefix and GRPO baselines (60.5% and 63.6%). Appendix C.2 provides the full six-dimension breakdown. These are win rates against a neutral TTS baseline, not scores on the reconstructed ParaS2S scale. Appendix G separately examines synthesis teachers and style targets at a reported matched training size of 90k.

Reasoning preservation (Q2). Freezing the Thinker preserves its parameters and decoding path; speech-side quality must still be evaluated separately. Table 2 shows that, in the 90k explicit-tag SFT setting, Qwen2.5 MMLU drops from 68.1 to 61.9 and VoiceBench from 72.41 to 66.34, whereas

Table 2: **Reasoning under expressive adaptation.** Mean VoiceBench (VB) and MMLU scores for the base models and 90k-data adaptation settings. Silent Tags (unfrozen) additionally updates the Thinker. Standard deviations are in Table 8.

Method	Qwen2.5		Kimi		Qwen3	
	VB	MMLU	VB	MMLU	VB	MMLU
Base	72.41	68.1	75.89	65.5	85.12	84.2
Explicit-tag SFT	66.34	61.9	73.26	60.7	81.05	81.1
Silent Tags (unfrozen)	73.21	64.8	74.95	65.1	84.86	83.9
Silent Tags (frozen)	72.41	68.1	75.89	65.5	85.12	84.2

Talker-only ***Silent Tags*** retains the base scores. The corresponding Kimi and Qwen3 results show the same direction of change under explicit-tag SFT; full statistics are in Appendix D. These results characterize the evaluated adaptation recipes, rather than proving a universal reasoning cost for every explicit-control method. Appendix D also reports 30k/60k explicit-tag SFT settings; the trends are not uniformly monotonic.

Allowing ***Silent Tags*** to update the Thinker also changes text-side performance: MMLU decreases on all three backbones, while VoiceBench increases on Qwen2.5 but decreases on Kimi and Qwen3. Freezing therefore provides a specific preservation property, not a claim that every parameter update harms every benchmark. Jointly reading Tables 1 and 2 makes the practical trade-off explicit: competitive expressive rendering can be obtained without altering the language backbone's measured text-side behavior.

4.3 Conditioning and Design Choices

Beyond expressive data and source affect (Q3). Same-data Talker FT already improves expressiveness to 3.658 on Qwen2.5 and 3.705 on Kimi; source-affect conditioning further improves these to 3.819 and 3.874. ***Silent Tags*** exceeds same-data FT by 0.457 and 0.404, and source-affect conditioning by 0.296 and 0.235, respectively. Thus, both expressive supervision and source affect are useful, but neither control recovers the full benefit of the response-conditioned readout in this comparison. Category-level differences alone do not determine whether speaker conditioning, data distribution, benchmark difficulty, or annotation agreement explains a particular gain. Earlier text-only, top-layer, and shuffled-state controls are reported separately in Appendix E.2; they support the readout direction without establishing a fully capacity-matched causal comparison.

4.4 Does Delivery Follow the Generated Response?

Aggregate expressiveness does not establish that a model changes its delivery with the response being rendered. An affect encoder can learn, for example, to answer an angry user calmly. The distinguishing question is whether the *same* acoustic input can support different delivery trajectories when the assistant's reply changes. We therefore evaluate response-conditioned control separately from overall expressiveness.

Fixed-source, varying-response test. For each of 100 source utterances, we construct two responses with different semantic and affect trajectories while holding the input audio fixed. One response to an angry source may initially match its urgency and then become concerned; another may take a calm, apologetic, or safety-oriented trajectory. The source-affect embedding is identical within a pair, but the desired delivery differs. We compare the same-data Talker, source-affect conditioning, explicit-tag SFT, and ***Silent Tags***. All tested affect-transition pairs (20 in total) have zero occurrences in the 90k training corpus under normalized tag-pair matching. This reduces the possibility of recovering the target simply by memorizing an observed transition, although it does not establish the absence of every semantically similar training example.

Waveform-level measurement. We score generated speech rather than similarity to the teacher embedding used for distillation. An independent paralinguistic classifier, obtained by fine-tuning Whisper-large-v3 on disjoint labeled speech, measures segment target accuracy, boundary-switch

Table 3: **Fixed-source, varying-response control.** For each of 100 source utterances, two responses require different delivery trajectories. Metrics are computed on generated waveforms by an independent classifier; all values are percentages (↑). The 20 tested transition pairs are absent from the training set under normalized tag-pair matching.

Method	Segment target	Boundary switch	Style persistence
Same-data Talker FT	56.8	52.4	61.5
Source-affect conditioning	68.3	61.2	70.1
Explicit-tag SFT	79.1	74.6	77.4
Silent Tags	83.4	79.8	82.0

Table 4: Human evaluation on held-out dialogue turns. Ratings use a 1–5 Likert scale and are averaged over three independent annotators per system output; higher is better. Trajectory consistency is evaluated on the mixed-trajectory subset. Overall quality is rated directly rather than computed from the other dimensions. The TTS teacher uses ground-truth tags and serves as an oracle reference. More detail can be found in A.5

Model	Naturalness	Para. Approp.	Semantic Faith.	Traj. Consist.	Overall
Base omni model	3.72	3.18	**4.53**	3.05	3.41
Explicit-tag baseline	3.91	4.12	4.21	4.03	4.07
Silent Tags	**4.26**	**4.38**	4.48	**4.31**	**4.35**
TTS teacher w/ GT tags	4.41	4.55	4.50	4.46	4.49

accuracy, and style persistence. These respectively assess whether the intended style is realized, whether the style switches near an annotated boundary, and whether it is maintained away from boundaries. The classifier is not used for model training, teacher distillation, projection-head fitting, checkpoint selection, or data filtering. Windowed predictions use 400-ms windows with a 100-ms hop. Classifier calibration, exact boundary criteria, and source-level uncertainty remain unreported; Appendix A.3 separates the known protocol from these limitations.

Results and interpretation. Table 3 shows that source-affect conditioning improves over same-data Talker FT by 11.5 points in segment accuracy, 8.8 in boundary switching, and 8.6 in persistence. Source affect is an effective control signal, not a weak alternative that only transfers emotion. ***Silent Tags*** further improves these metrics by 15.1, 18.6, and 11.9 points over source-affect conditioning. It also has higher point estimates than explicit-tag SFT, by 4.3, 5.2, and 4.6 points, respectively. Gains concern both *which* style is rendered and *when* it changes or persists within a response.

The paired setup removes variation in input affect as an explanation for the different target trajectories. Because the source-affect baseline also receives response text, its lower scores do not imply that it is theoretically unable to model response-dependent delivery. Rather, the comparison supports the usefulness of a dedicated response-conditioned readout in the evaluated architecture: the Talker can access representations of the current reply at each codec step instead of relying only on an utterance-level affect signal and its standard text pathway. This is evidence for the proposed control route, not proof that all alternative audio-conditioned architectures are insufficient. It also does not isolate shallow-layer information from added capacity or an additional token-level text pathway; matched readout controls are needed for that attribution.

5 Conclusion

We presented ***Silent Tags***, a Talker-only approach that reads response-conditioned paralinguistic signals from a frozen Thinker. Segment-level teacher supervision trains the readout without requiring explicit tag generation at inference. Under a common reconstructed evaluation, the method achieves competitive expressiveness while retaining the Thinker's parameters and text-side behavior. The fixed-source experiment further supports a dedicated response-conditioned pathway: style realization, switching, and persistence improve over a source-affect-plus-text control. This motivates treating delivery as a property of the reply being rendered, not only of the user's affect.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing the manuscript, general code editing, and preparing figures and charts. We take responsibility for the final content of this work, including all text, code, and figures produced with AI assistance.

REFERENCES

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

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, 2008.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *CoRR*, abs/2406.05370, 2024.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 2026.

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

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. Technical report, Kyutai, September 2024. URL `http://kyutai.org/Moshi.pdf`.

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

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

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

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. *arXiv preprint arXiv:2505.02625*, 2025.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. Vita: Towards open-source interactive omni multimodal llm, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, 2019.

Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, Zejun Ma, and Zhou Zhao. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *CoRR*, abs/2306.03509, 2023.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Trans. Assoc. Comput. Linguistics*, 11:1703–1718, 2023.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijiang Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Min Yang, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, et al. Openomni: Advancing open-source omnimodal large language models with progressive multimodal alignment and real-time self-aware emotional speech synthesis. *arXiv preprint arXiv:2501.04561*, 2025.

Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. Emergenttts-eval: Evaluating tts models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. *arXiv preprint arXiv:2505.23009*, 2025.

OpenAI. Hello GPT-4o. `https://openai.com/index/hello-gpt-4o/`, 2024.

OpenAI. GPT-5.1: A smarter, more conversational ChatGPT. `https://openai.com/index/gpt-5-1/`, 2025.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2023.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197, 2023.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara N. Sainath, Johan Schalkwyk, Matthew Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirovic, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Havnø Frank. Audiopalm: A large language model that can speak and listen. *CoRR*, abs/2306.12925, 2023.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Interspeech*, 2022.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Haoyu Wang, Guangyan Zhang, Jiale Chen, Jingyu Li, Yuehai Wang, and Yiwen Guo. Empathy omni: Enabling empathetic speech response generation through large language models. *arXiv preprint arXiv:2508.18655*, 2025.

Xiong Wang, Yangze Li, Chaoyou Fu, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*, 2024.

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming, 2024. URL https://arxiv.org/abs/2408.16725.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Chao Yan, Boyong Wu, Peng Yang, Pengfei Tan, Guoqiang Hu, Li Xie, Yuxin Zhang, Fei Tian, Xuerui Yang, Xiangyu Zhang, et al. Step-audio-editx technical report. *arXiv preprint arXiv:2511.03601*, 2025.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive TTS in discrete latent space with natural language style prompt. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:2913–2925, 2024.

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

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 15757–15773. Association for Computational Linguistics, 2023.

A Evaluation Protocol

A.1 Reconstructed ParaS2SBench

Our evaluation follows the task description of ParaS2SBench (Yang et al., 2025), but uses reconstructed prompts and a substituted text judge. The original benchmark's test audio, filtering scripts, and judge rubric were not publicly available at evaluation time. Consequently, the results should not be interpreted as scores on the original unreleased test set.

Composition. The reconstruction contains 1,200 synthetic prompts: 300 each for Emotion, Sarcasm, Age, and Gender. Real-speech inputs comprise 709 prompts from IEMOCAP (Busso et al., 2008) and 781 from MELD (Poria et al., 2019), giving 2,690 prompts in total. The corrected Age subset uses adult/child input conditions, not speaking-rate control. All retained systems are reevaluated on this Age subset; earlier speaking-rate scores are not relabeled or reused.

Judging pipeline. The same prompt-audio manifest and frozen pipeline are used for every system in Table 6. Whisper-large-v3 (Radford et al., 2023) extracts the input and output transcripts. The benchmark label specifies the input style, and AudioReasoner extracts the output speaking style. A common rubric with Qwen3-32B-Instruct produces scores on a 1–5 scale. This differs from the original benchmark's ChatGPT-4.1 text judge. No baseline score measured using the original paper's prompts or judge is reused in our common-protocol table.

Aggregation. Let m_j be the aggregate score for each of the four synthetic dimensions and two real-speech corpora. We report

$$\text{Overall} = \frac{1}{6}\sum_{j=1}^{6} m_j.$$

Thus, Emotion, Sarcasm, Age, Gender, IEMOCAP, and MELD have equal weight despite different sample counts. Overall is neither a pooled sample average nor an equal-weight average of a four-category synthetic mean and a two-corpus real mean. Category-level scores are reported in Table 6.

A.2 Matched Conditioning Controls

The same-data Talker FT control receives the expressive training corpus without the added cross-attention readout from lower Thinker layers. The source-affect control uses the same 90k corpus, Talker architecture, response-text pathway, optimization budget, and training objective as ***Silent Tags***, but replaces the Thinker readout with an utterance-level affect embedding from the input audio. Importantly, it also receives the generated response text. It can learn an empathetic input-to-output affect mapping and is not restricted to reproducing the input emotion.

A.3 Fixed-Source, Varying-Response Evaluation

Paired construction and training-set check. The evaluation uses 100 source utterances, each paired with two responses that differ in semantics and intended affect trajectory. The source audio is unchanged within each pair. The 20 evaluated affect-transition pairs have zero occurrences in the 90k training corpus under normalized tag-pair matching. This check concerns the evaluated tag-pair patterns, not a guarantee of semantic disjointness from all training responses.

Independent audio-side evaluation. Metrics are computed directly on generated waveforms by a paralinguistic classifier obtained by fine-tuning Whisper-large-v3 on disjoint labeled speech. This classifier is not used for Silent Tags training, teacher distillation, projection-head fitting, checkpoint selection, or synthesis-data filtering. It is distinct from an evaluation projection into teacher-embedding space.

- **Segment target accuracy:** whether generated response segments realize their intended paralinguistic style according to the audio classifier.

Table 5: Blinded pairwise human preference between *Silent Tags* and the strongest same-backbone explicit-tag baseline. Values are percentages of individual listener judgments. Listeners judge which output better matches the dialogue context and intended response while preserving its meaning; ties are permitted.

Subset	Prefer *Silent Tags*	Prefer explicit-tag	Tie
All samples	**63.7**	25.8	10.5
Single-tag samples	**59.2**	29.4	11.4
Mixed-trajectory samples	**68.4**	22.1	9.5

- **Boundary-switch accuracy:** whether the classifier's predicted style changes from the pre-boundary target to the post-boundary target near the annotated transition. Predictions use 400-ms windows with a 100-ms hop.
- **Style persistence:** the fraction of non-boundary windows classified as the intended style within a segment.

All three metrics in Table 3 are percentages. The evaluation is separate from the earlier 200-utterance trajectory diagnostics; their sample sets and metric variants must not be combined. In particular, the earlier internal diagnostic's ± 2-codec-frame tolerance is not a specified tolerance for this waveform-level test.

A.4 Complementary Evaluation

EmergentTTS-Eval (Manku et al., 2025) reports GPT-4o-based pairwise win rates against a neutral TTS baseline across Emotion, Emphasis, Question, Irony, Whispering, and Laughter. Its scale differs from the reconstructed ParaS2SBench score; the two are reported separately. VoiceBench (Chen et al., 2026), MMLU (Hendrycks et al., 2021), and GSM8K (Cobbe et al., 2021) concern Thinker text-side performance and do not by themselves establish the intelligibility or appropriateness of generated speech.

A.5 Human Evaluation

Automatic judges provide scalable measurements of paralinguistic expressiveness, but may not fully capture whether generated speech sounds natural and contextually appropriate to human listeners. We therefore conduct a blinded human evaluation of Silent Tags and the principal comparison systems. The study measures both absolute perceptual quality and direct pairwise preference.

Evaluation set. We randomly sample 200 dialogue turns from the held-out test split, with no overlap with the training or development data. The sample includes both single-tag responses and mixed-trajectory responses in which the intended speaking style changes across clauses. For every dialogue turn, all evaluated systems receive the same dialogue context and response content. We compare four systems: (i) the original Qwen2.5-Omni model without additional paralinguistic control; (ii) Explicit-tag baseline, the strongest same-backbone explicit-tag baseline; (iii) Silent Tags; and (iv) the tag-controllable TTS teacher conditioned on the ground-truth tag schedule. The final system uses privileged control information and is included only as an oracle reference, not as a deployable end-to-end S2S baseline.

Annotators and presentation. We recruit 15 fluent English listeners with self-reported normal hearing. The generated audio clips are anonymized, and their presentation order is randomized so that annotators are not informed of the system identity. Each system output for each dialogue turn is evaluated independently by three annotators. Annotators are shown the dialogue context and intended response text before listening to the generated response.

Absolute-rating protocol. Annotators assign a score from 1 to 5 for each of the following dimensions:

- Naturalness: whether the generated speech is fluent, smooth, and free from noticeable synthesis artifacts.

- Paralinguistic appropriateness: whether its emotion, speaking style, and vocal events are appropriate for the dialogue context and response.
- Semantic faithfulness: whether the spoken output preserves the meaning of the intended response text without material omissions, additions, or distortions.
- Overall quality: the annotator's holistic assessment of the generated response, considering both what is spoken and how it is delivered.

For mixed-trajectory responses, annotators additionally rate **trajectory consistency**, measuring whether the delivery realizes and sustains the intended clause-level change in paralinguistic style. This dimension is evaluated only on the mixed-trajectory subset and is not assigned to single-tag responses.

For each system–dialogue pair, we first average the three annotator ratings. The reported score for each dimension is then the mean across the applicable dialogue turns. Overall quality is collected as a separate holistic rating and is not computed as an average of the other dimensions.

Pairwise-preference protocol. We additionally compare Silent Tags directly with Explicit-tag baseline on the same 200 dialogue turns. For each turn, annotators hear the two anonymized outputs in randomized order and select the output that better matches the dialogue context and intended response while preserving its semantic content. Annotators may also select a tie when neither output is clearly preferable. Each pair is judged independently by three annotators.

Table 5 reports the percentage of individual annotator judgments assigned to Silent Tags, the explicit-tag baseline, or a tie. Results are reported over all 200 turns and separately for the single-tag and mixed-trajectory subsets. The subset analysis is specified in advance to examine whether the relative advantage changes when a response requires an intra-utterance transition rather than a single speaking style.

In the absolute-rating evaluation, Silent Tags obtains the highest scores among deployable systems for naturalness, paralinguistic appropriateness, trajectory consistency, and overall quality. Its semantic-faithfulness score of 4.48 remains close to the base model's 4.53, suggesting that the expressive gains do not coincide with a large loss in perceived response meaning. The ground-truth-tag TTS teacher remains the strongest system overall, as expected from its access to the oracle style schedule.

In the pairwise comparison, Silent Tags is preferred in 63.7% of all judgments, compared with 25.8% for the explicit-tag baseline and 10.5% ties. The preference for Silent Tags increases from 59.2% on single-tag responses to 68.4% on mixed-trajectory responses. This difference is consistent with the hypothesis that response-conditioned continuous control is particularly useful when delivery must change within an utterance. The human results establish a perceptual preference for the generated speech, but do not by themselves isolate whether the improvement arises specifically from shallow-layer representations, the additional response-conditioned pathway, or another architectural difference.

B Data Construction and Training Configuration

B.1 Corpus Stages and Supervision

The initial text QA/dialogue pool contains approximately 4.2M pairs. This is not the size of the expressive speech corpus used for training. In the main CosyVoice 2 setting, approximately 220k samples are synthesized and approximately 90k survive filtering (about 40.9%). The retained examples cover 48 paralinguistic categories and carry response text, clause boundaries, segment tags, expressive waveforms, and style-embedding targets.

GPT-5.1 provides the clause-level tag schedule; CosyVoice 2 synthesizes the speech; a separate Fish Audio S2 style encoder supplies the 512-dimensional distillation targets. The synthesis teacher and target encoder are distinct components. Fish Audio S2 is not the synthesis teacher of the main configuration.

B.2 Filtering Stages

The pipeline applies four types of quality control:

1. **Annotation self-consistency:** check the segment-level tag assignments for consistency.
2. **Tag–audio consistency:** use a paralinguistic classifier to assess whether synthesized speech realizes the intended tags.
3. **Acoustic and lexical quality:** combine UTMOS (Saeki et al., 2022) quality scoring with Whisper-based (Radford et al., 2023) transcript recovery.
4. **Tag-distribution balancing:** downsample over-represented categories to limit dominance by frequent styles.

These stages remove different sources of supervision noise. The reported retention rate is an aggregate outcome and does not specify the individual acceptance thresholds.

B.3 Reported Talker Training Recipe

The reported main-setting training recipe uses AdamW (Loshchilov & Hutter, 2019) for 30k steps, a learning rate of 2×10^{-4} with cosine decay, a global batch size of 256, and 16 H100-80GB GPUs. The distillation and auxiliary-loss weights are $\lambda_1 = 0.5$ and $\lambda_2 = 0.2$, selected on a held-out development split. Only Talker-side parameters receive gradients. The auxiliary classifier is discarded at inference.

Cross-attention is inserted in the bottom seven Talker layers, with layer i attending to frozen Thinker layer i for $i = 0, \ldots, 6$. The added Q/K/V projections are randomly initialized; these projections are not shared with the Thinker's own attention projections. The segment-level projection into the teacher embedding space is a two-layer MLP. The baseline top-layer conditioning pathway is preserved. Teacher/encoder sensitivity and residual distribution confounds are discussed in Appendix G.

C Extended Expressiveness Results

C.1 Full Common-Protocol Comparison

Table 6 expands the compact comparison in the main text with per-category scores, pipeline and closed-source baselines, and an oracle-label reference. Closed-source systems are not assigned a known Thinker-update or internal tag-generation status. GPT-TTS with ground-truth labels has privileged control information and is not a deployable end-to-end comparison.

Table 6: Results on our **reconstructed ParaS2SBench**, not the original unreleased test set. All rows use the same reconstructed prompts, newly evaluated Age subset, acoustic-analysis pipeline, and Qwen3-32B-Instruct judge; no baseline scores are borrowed from the original paper. Scores range from 1 to 5. Overall is the equal-weight mean of the four synthetic dimensions and the two real-speech corpora, rather than a sample-weighted mean. "Update" indicates Thinker fine-tuning; "Tags" indicates explicit paralinguistic tag-token generation by the Thinker at inference. N/A denotes an inapplicable or undisclosed property.

Model	Update	Tags	Synthetic Speech				Real Speech		Overall
			Emotion	Sarcasm	Age	Gender	IEMOCAP	MELD	
Pipeline and closed-source baselines									
Whisper-GPT-TTS	N/A	No	3.126	2.951	3.093	2.884	3.574	3.372	3.167
GPT-4o-2024-11-20	N/A	N/A	3.712	2.978	3.301	3.391	3.851	3.520	3.459
Gemini-2.5-flash	N/A	N/A	3.863	3.248	3.391	3.437	3.865	3.709	3.586
Base open-source models									
Qwen2.5-Omni	No	No	3.656	2.904	3.221	3.216	3.642	3.584	3.370
Qwen3-Omni-30B-A3B	No	No	3.902	3.351	3.711	3.517	3.894	3.712	3.681
GLM-4-Voice	No	No	3.450	2.747	2.929	2.954	2.938	3.105	3.021
LLaMA-Omni 2	No	No	3.539	3.131	3.174	3.064	3.456	3.429	3.299
Freeze-Omni	No	No	2.319	2.664	2.864	2.857	2.858	3.026	2.765
Kimi-Audio	No	No	2.701	2.629	3.203	3.090	1.392	1.153	2.361
Matched Talker-side controls									
Qwen2.5 Same-data Talker FT	No	No	3.886	3.378	3.693	3.518	3.827	3.645	3.658
Qwen2.5 Source-affect conditioning	No	No	4.041	3.602	3.853	3.640	4.040	3.736	3.819
Kimi Same-data Talker FT	No	No	3.881	3.469	3.883	3.590	3.872	3.533	3.705
Kimi Source-affect conditioning	No	No	4.032	3.819	4.015	3.732	4.033	3.611	3.874
Explicit-tag methods									
Qwen2.5 Explicit-tag SFT	Yes	Yes	4.218	4.126	4.302	3.917	4.208	3.912	4.114
Qwen3 Explicit-tag SFT	Yes	Yes	4.492	4.401	4.583	4.219	4.451	4.238	4.397
Kimi-Audio SFT	Yes	Yes	4.121	4.261	4.461	3.528	4.162	3.284	3.969
Kimi-Audio GRPO	Yes	Yes	4.516	4.506	4.555	4.228	4.426	3.899	4.355
Silent Tags: frozen Thinker, Talker-only training									
Silent Tags (Qwen2.5)	No	No	4.236	4.079	4.261	3.998	4.183	3.931	4.115
Silent Tags (Qwen3)	No	No	4.548	4.438	4.607	4.286	4.493	4.211	4.430
Silent Tags (Kimi)	No	No	4.284	4.112	4.315	4.036	4.217	3.689	4.109
Oracle-label reference (not an end-to-end baseline)									
GPT-TTS with GT labels	N/A	No	4.718	4.760	4.591	4.806	4.747	4.798	4.737

C.2 EMERGENTTTS-EVAL BREAKDOWN

Table 7 reports the six individual dimensions in addition to the average discussed in the main text. These are pairwise win rates against a neutral TTS reference, not reconstructed ParaS2SBench scores.

Table 7: **EmergentTTS-Eval** (Manku et al., 2025): GPT-4o-based pairwise win rates (%) against a neutral TTS baseline. The explicit-control variants use tag conditioning; the Silent Tags variants keep the Thinker frozen.

Model	Emotion	Emphasis	Question	Irony	Whispering	Laughter	Avg
Pipeline baselines							
CosyVoice (neutral)	41.2	38.7	43.1	35.4	39.8	33.2	38.6
GPT-4o-mini-TTS	62.3	58.9	64.5	55.1	61.2	52.4	59.1
Base S2S models							
Qwen2.5-Omni	51.3	47.8	53.2	44.6	50.1	41.3	48.1
GLM-4-Voice	48.7	45.1	50.3	42.2	47.6	38.9	45.5
LLaMA-Omni 2	50.1	46.3	51.8	43.5	48.9	40.2	46.8
Kimi-Audio	46.5	43.2	48.7	40.1	45.3	37.8	43.6
Explicit paralinguistic control							
Kimi-Audio + Tag Prefix	63.8	59.4	65.7	57.2	62.5	54.3	60.5
Kimi-Audio GRPO	67.2	62.8	68.4	60.3	65.1	57.6	63.6
Silent Tags							
Silent Tags (Qwen2.5)	65.4	61.2	67.1	59.8	63.7	56.4	62.3
Silent Tags (Kimi)	69.3	64.7	70.2	62.5	67.4	59.8	65.7

The earlier conditioning and trajectory studies are reported separately in Appendices E.2 and F. Their reported runs are not pooled with this breakdown or with the common-protocol main comparison.

D THINKER-SIDE REASONING RESULTS

Table 8 reports text-side results for the base models and the 90k-data adaptation settings. Expressive adaptation with explicit-tag SFT reduces VoiceBench and MMLU relative to the corresponding base model in all three backbones. Updating the Thinker within Silent Tags also changes its text-side scores, whereas Talker-only training preserves them. This is an empirical comparison of the evaluated training recipes, not proof that one particular interference mechanism explains every regression.

Table 8: **Reasoning endpoints** (mean $\pm$ standard deviation over three reported seeds). Adapted variants use 90k training pairs; Base uses no additional adaptation data. No earlier-protocol ParaS2S scores are included in this table.

Backbone	Adaptation	VoiceBench	MMLU
Kimi-Audio	Base	75.89 ± 0.20	65.5 ± 0.2
	Explicit-tag SFT	73.26 ± 0.30	60.7 ± 0.4
	Silent Tags + Thinker update	74.95 ± 0.22	65.1 ± 0.3
	Silent Tags, Talker-only	75.89 ± 0.20	65.5 ± 0.2
Qwen2.5-Omni	Base	72.41 ± 0.22	68.1 ± 0.3
	Explicit-tag SFT	66.34 ± 0.32	61.9 ± 0.4
	Silent Tags + Thinker update	73.21 ± 0.24	64.8 ± 0.3
	Silent Tags, Talker-only	72.41 ± 0.22	68.1 ± 0.3
Qwen3-Omni	Base	85.12 ± 0.20	84.2 ± 0.2
	Explicit-tag SFT	81.05 ± 0.31	81.1 ± 0.4
	Silent Tags + Thinker update	84.86 ± 0.22	83.9 ± 0.3
	Silent Tags, Talker-only	85.12 ± 0.20	84.2 ± 0.2

Scope of preservation. Freezing the Thinker and retaining its decoding path prevents Talker training from changing Thinker parameters. Under otherwise identical input processing and text decoding, this preserves its text-side behavior. It does not guarantee that the final waveform transcribes perfectly or that the selected delivery is appropriate; these require separate speech-side evaluation.

D.1 SENSITIVITY TO EXPLICIT-TAG TRAINING-SET SIZE

Table 9: **Additional explicit-tag SFT data scales**: mean $\pm$ standard deviation over three reported seeds. The 90k endpoints and base scores appear in Table 8. No earlier ParaS2S coordinates are attached to these reasoning results.

Backbone	Training pairs	VoiceBench	MMLU
Kimi-Audio	30k	75.12 ± 0.22	64.2 ± 0.3
	60k	74.45 ± 0.26	62.5 ± 0.3
Qwen2.5-Omni	30k	71.43 ± 0.25	66.2 ± 0.6
	60k	68.12 ± 0.28	63.3 ± 0.4
Qwen3-Omni	30k	83.25 ± 0.24	82.4 ± 0.3
	60k	80.25 ± 0.27	79.8 ± 0.3

Table 9 adds the reported 30k and 60k settings. VoiceBench and MMLU are below the corresponding base scores at each of the three data scales, but the trend is not universally monotonic. In particular, Qwen3-Omni improves from 60k to 90k on both VoiceBench (80.25 to 81.05) and MMLU (79.8 to 81.1), while remaining below its base scores. These observations characterize the evaluated fine-tuning recipe, not an inevitable penalty that grows with every additional training example. Without matched common-protocol expressiveness coordinates for all these checkpoints, they do not define a complete reasoning–expressiveness frontier.

E CONDITIONING AND SUPERVISION DIAGNOSTICS

Protocol scope. The following layer and loss ablations retain results from the earlier evaluation configuration, preceding the common-protocol reevaluation in Table 1. Their ParaS2S scores are not directly comparable to the corrected Age-inclusive evaluation. They are retained as design diagnostics, not as additional common-protocol comparisons.

E.1 CROSS-ATTENTION AND SUPERVISION LOSSES

Table 10: **Earlier-protocol ablation study**: synthetic average on a 1–5 scale and VoiceBench text-side score. The original table attributes the variants to Qwen2.5-Omni. Dashes do not denote zero scores; the frozen Thinker's VoiceBench score is unchanged by Talker-only training. These results are not directly comparable to Table 1.

Variant	Cross-attn	$\mathcal{L}_{\text{distill}}$	$\mathcal{L}_{\text{aux}}$	ParaS2S	VoiceBench
Base	✗	✗	✗	3.248	72.4
Upper layers 7–N	✓	✗	✗	3.612	—
Lower layers 0–6	✓	✗	✗	3.891	—
+ distillation	✓	✓	✗	4.026	—
+ auxiliary only	✓	✗	✓	3.966	—
+ both losses	✓	✓	✓	4.138	—

Reading lower rather than upper layers improves the earlier-protocol synthetic score by 0.279. Adding distillation to the lower-layer readout improves it by 0.135, adding auxiliary classification alone improves it by 0.075, and using both improves it by 0.247. These comparisons support complementary supervision within this configuration, not universal necessity of either loss. The probe-motivated 0–6 cutoff is an empirical choice, not a demonstrated generation or compute optimum.

E.2 TEXT, UPPER-LAYER, AND SHUFFLED-STATE CONTROLS

Table 11: **Earlier readout controls on Qwen2.5-Omni.** All five adapted variants use the same expressive corpus and training budget. EmergentTTS-Eval win rates are reported as mean ± standard deviation over three runs. These diagnostic runs are not the corrected common-protocol comparison in Table 1.

Variant	Conditioning source	Win rate (%) ↑
Same-data Talker FT	Standard pathway	54.3 ± 0.6
Text-only conditioning	Text/top-state pathway	55.6 ± 0.5
Shuffled Thinker states	Shuffled lower states	53.2 ± 0.7
Top-layer cross-attention	Upper Thinker states	57.8 ± 0.5
Silent Tags	Lower Thinker states (0–6)	62.3 ± 0.4

Silent Tags has higher reported EmergentTTS-Eval means than text/top-state conditioning, upper-layer cross-attention, and shuffled-state conditioning (Table 11). We omit the earlier ParaS2S column because its aggregation is not identified with the corrected main-table Overall metric.

These controls support the usefulness of the selected response representations, but do not fully isolate their contribution from capacity or conditioning granularity. The exact shuffle dimension and whether shuffling is applied in training, inference, or both are not documented. Shuffling may disrupt response-text information as well as affective information, so its effect cannot be attributed solely to the latter. Equal training data and budget also do not, by themselves, establish equal parameter count or computation. These are supporting diagnostics, not a replacement for the main common-protocol comparisons.

F RESPONSE-TRAJECTORY DIAGNOSTICS

This earlier Qwen2.5-Omni diagnostic evaluates 200 held-out utterances containing more than 400 annotated clause boundaries. It is distinct from the fixed-source test in Section 4.4: 200 utterances here do not denote the two responses to each of that test's 100 sources. We separate internal representation measurements from output-waveform measurements because similarity to a training target space is not independent evidence of perceptual control.

Table 12: **Earlier trajectory diagnostic**, with reported means and accompanying variability. Boundary accuracy is a percentage; persistence is a fraction. WER is reported as a point estimate. The $\pm$ values are not confidence intervals, and these observations must not be pooled with Table 3.

Internal conditioning-space measurements			
Method	Boundary acc. ↑	Teacher similarity ↑	Persistence ↑
Same-data Talker FT	51.6 ± 1.8	0.34 ± 0.05	0.63 ± 0.03
Explicit-tag baseline	73.4 ± 1.3	0.56 ± 0.03	0.77 ± 0.02
Silent Tags	75.2 ± 1.0	0.59 ± 0.012	0.80 ± 0.02
Independent output-waveform measurements			
Method	Boundary acc. ↑	Persistence ↑	WER ↓
Same-data Talker FT	53.6 ± 2.4	0.31 ± 0.05	5.14
Explicit-tag baseline	74.1 ± 2.1	0.51 ± 0.07	6.27
Silent Tags	74.4 ± 1.2	0.63 ± 0.019	5.18

Internal measurements. For each model, a diagnostic projection maps its conditioning representation into the teacher style space: the silent-tag readout, the explicit baseline's tag-conditioned Talker state, or the same-data baseline's Talker state. The diagnostic projection is fitted on a disjoint held-out calibration split and is not used for training the speech models. Internal boundary accuracy checks whether the nearest predicted conditioning label changes from the pre-boundary to the post-boundary target within ± 2 codec frames, approximately ± 80 ms on the reported 25-Hz grid. Internal persistence measures the fraction of non-boundary frames assigned the segment's target label. Teacher similarity is the cosine similarity between the projected state and the segment's style target, averaged over the evaluated frames/segments. These metrics are diagnostic, not a claim that hidden-space proximity guarantees audible style realization.

Waveform measurements. The audio-side classifier is based on Whisper-large-v3 fine-tuned independently on held-out labeled speech. It is not used for speech-model training, distillation, diagnostic-projection fitting, or checkpoint selection. Audio predictions use 400-ms windows and a 100-ms hop to assess the intended pre/post styles and persistence within segments. The internal ± 80-ms criterion must not be transferred to this waveform test: its exact boundary acceptance and exclusion rules are not specified by the window configuration. WER compares an ASR transcript of the generated waveform with the reference response text.

Interpretation and scope. Silent Tags and the explicit baseline have similar audio boundary point estimates (74.4 versus 74.1), while Silent Tags has higher reported audio persistence (0.63 versus 0.51). Its WER is close to same-data fine-tuning (5.18 versus 5.14), not lower. The initial account associates the internal trajectory results with a selected teacher/filter setting; the later expanded table does not identify that setting's checkpoint. We therefore treat these as separate supporting diagnostics, not a replication of the main configuration or a statistically established switching advantage. Raw trajectories and perceptual calibration are not supplied, so no boundary-aligned plot or human-listening result is inferred from the aggregate table.

G SYNTHESIS-TEACHER AND STYLE-TARGET SENSITIVITY

Matched-size comparison. The final sensitivity comparison reports 90k training pairs for every variant, replacing an earlier comparison that used 100k pairs for the stronger teachers. Samples are selected from each teacher's filtered pool by class-stratified sampling toward the CosyVoice 2 tag distribution. The comparison reports common annotation/tag schedules, data split, Talker architecture and initialization, optimizer, training budget, loss weights, evaluation prompts, decoding, and judge. The synthesis teacher and style-target encoder are varied separately where possible.

Table 13: **Reported matched-90k sensitivity on Qwen2.5-Omni.** Every row uses 90k pairs. TTS WER measures teacher synthesis; Talker WER measures generated responses. EmergentTTS-Eval scores are win rates (%). These point estimates belong to the sensitivity study, not additional rows of the common-protocol main table.

Synthesis teacher	Style target	TTS WER ↓	Talker WER ↓	Win rate ↑
CosyVoice 2	Fish Audio S2	2.55	5.18	62.3
Fish Audio S2	Fish Audio S2	2.61	5.21	62.0
CosyVoice 2	emotion2vec	2.55	5.27	61.5
CosyVoice 2	WavLM-Large*	2.55	5.33	60.8
Qwen-Audio-3.0-TTS	Fish Audio S2	1.35	2.30	63.9
ElevenLabs v3	Fish Audio S2	1.13	1.94	65.0
ElevenLabs v3	emotion2vec	1.13	1.96	64.3

*WavLM-Large with an emotion head.

Teacher synthesis. With the Fish Audio S2 target held fixed, the reported EmergentTTS-Eval win rate increases from 62.3 for CosyVoice 2 to 63.9 for Qwen-Audio-3.0-TTS and 65.0 for ElevenLabs v3. Teacher and Talker WER also decrease in these comparisons. This association is consistent with the importance of lexical fidelity in expressive supervision, but does not isolate lexical accuracy as the causal explanation for improved style control.

Style targets. With CosyVoice 2 synthesis held fixed, replacing the Fish target with emotion2vec or a WavLM-Large emotion head yields win rates of 61.5 and 60.8, respectively, compared with 62.3. With ElevenLabs synthesis, Fish and emotion2vec targets yield 65.0 and 64.3. The tested alternatives thus retain much of the reported expressive performance, without establishing that target encoders are interchangeable or that these differences are significant.

Residual confounds and protocol boundary. Equal training-set size removes the earlier 90k-versus-100k confound. However, complete per-tag counts and numerical distribution distances are not supplied, so exact category matching cannot be verified. Teacher-specific synthesis and filtering also change the retained waveforms and potentially the retained examples. These results are therefore sensitivity evidence, not a fully isolated teacher-quality experiment. No uncertainty estimates are attached to the final matched-size table; uncertainties from the earlier unequal-size runs must not be transferred to it. We omit its earlier ParaS2S diagnostic column because its aggregation has not been reconciled with the corrected Age-inclusive Overall score.