# Bridging What the Model Thinks and How It Speaks: Expressive Speech Generation via Self-Aware Intent-Realization Alignment

Anonymous ACL submission


## Abstract

Speech Language Models (SLMs) exhibit strong semantic understanding, yet often fail to translate this capacity into expressive acoustic realization, producing speech with flattened prosody and misaligned emotion. We identify this mismatch as the **semantic understanding–acoustic realization gap**. Existing approaches typically rely on externally specified proxies, such as emotion labels or style prompts, which require annotations and struggle to capture dynamically evolving expressive intent throughout dialogue. To overcome these limitations, we propose **SASLM** (**S**elf-**A**ware **S**peech **L**anguage **M**odel), a proxy-free framework that bridges what the model thinks and how it speaks through self-aware intent–realization alignment: (1) *Intent-Aware Bridging* self-distills expressive intent from the model's own evolving semantic generation states via a Variational Information Bottleneck (VIB), thereby guiding expressive speech realization without external expressive supervision; while (2) *Realization-Aware Alignment* reflectively aligns generated acoustics with intended expression through self-reward optimization, progressively improving intent–realization consistency during speech generation. Despite using only 3B parameters and 800 hours of expressive speech data, SASLM achieves state-of-the-art performance on EchoMind among open-source systems, surpassing models over 10× larger and approaching commercial systems.[1]


## 1 Introduction

Speech Language Models (SLMs) have substantially advanced multimodal interaction by integrating speech and text within unified generative frameworks, achieving remarkable success in multimodal understanding (Xu et al., 2025a,b). However, stronger semantic intelligence does not necessarily yield better spoken interaction. Unlike text generation, spoken communication depends not only on *what* is said, but also on *how* it is acoustically realized (Zhu et al., 2022), requiring contextually appropriate emotional intensity (Scherer, 2003), prosody (Ladd, 2008), and spontaneous naturalness (Shriberg, 2005).

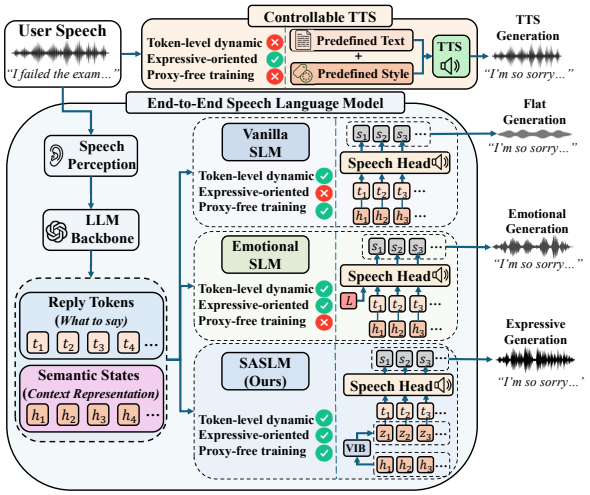


Figure 1: Comparison of context-aware expressive speech generation paradigms, categorized by their **expressive control signals**. Controllable TTS depends on external style references, whereas Vanilla SLMs naively reuse semantic hidden states $h_i$ to modulate text tokens $t_i$ into speech $s_i$. Unlike Emotional SLMs, which inject explicit emotion labels $L$ prior to synthesis, SASLM self-distills intrinsic expressive intent $z_i$ directly from $h_i$, eliminating proxy supervision.

While large-scale training endows current vanilla SLMs (Zeng et al., 2024; Xu et al., 2025a) with robust semantic capabilities, they still struggle to translate this intelligence into context-aware expressive speech interaction. As illustrated in Fig. 1, they directly reuse text-prediction-oriented semantic hidden states for speech realization, which are not inherently aligned with expressive acoustic patterns and often yield semantically correct but prosody-flattened "spoken text" (Tu et al., 2025;

[1]Demo samples are available on the anonymous project page: https://anonymous.4open.science/w/SASLM-7FD1/.

Zhan et al., 2025; Fang et al., 2025b). We refer to this mismatch as the **semantic understanding–acoustic realization gap**: semantic states that support strong understanding are not necessarily acoustically usable for expressive speech generation.

Existing approaches attempt to bridge this gap by externalizing expressive intent as explicit proxies. Cascaded systems (Li et al., 2024) typically use controllable TTS with style prompts inferred by speech understanding models, while emotional SLMs (Chen et al., 2025; Wang et al., 2025a; Gao et al., 2025c) internalize it as affective proxies via joint emotion–speech prediction. Although effective in constrained scenarios, these approaches struggle to capture dynamically evolving intent and require costly annotations. Moreover, their training paradigms remain largely open-loop, optimizing token-level imitation without explicitly verifying cross-modal consistency between the intended expression and its acoustic realization.

To address these limitations, we rethink context-aware expressive speech generation as a problem of self-aware intent–realization alignment. To this end, we propose **SASLM** (**S**elf-**A**ware **S**peech **L**anguage **M**odel), a proxy-free framework that intrinsically derives expressive intent from semantic generation dynamics, rather than relying on externally specified controls, and aligns generated acoustics with the intended expression via reflective self-feedback. This reformulates the task as a closed-loop process that jointly models semantic understanding and acoustic realization. Concretely, SASLM realizes this self-aware loop via two coupled mechanisms. *Intent-Aware Bridging* internally forms expressive intent without external supervision by using a Variational Information Bottleneck (VIB) to self-distill evolving semantic states, disentangling acoustically usable intent governing *how to speak* from lexical content already specified by generated tokens. *Realization-Aware Alignment* further enforces intent–realization consistency via a self-reward mechanism: SASLM acts as its own critic, scoring generated speech against its intended expression and optimizing itself using utterance-level rubric rewards.

Experiments on the EchoMind benchmark (Zhou et al., 2026) demonstrate that SASLM achieves state-of-the-art (SOTA) expressive performance among open-source models using only 3B parameters and 800 hours of expressive speech data, outperforming the 10× larger Qwen3-Omni-30B while approaching closed-source commercial systems. Our main contributions are as follows:

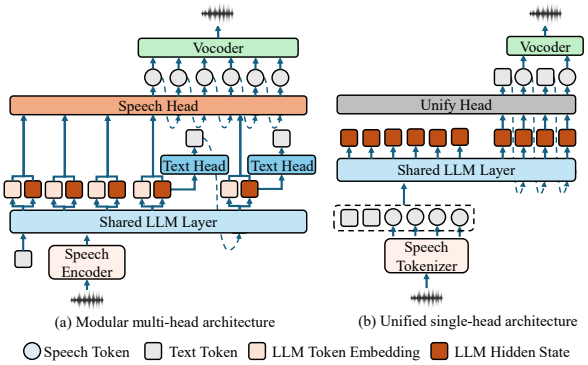


Figure 2: Illustration of existing end-to-end SLM architectures: (a) Modular multi-head architecture; (b) Unified single-head architecture.

(1) We identify the **semantic understanding–acoustic realization gap** in current SLMs and reformulate expressive speech generation as a problem of self-aware intent–realization alignment.

(2) We propose SASLM, a proxy-free framework that bridges semantic understanding and expressive acoustic realization through *Intent-Aware Bridging* and *Realization-Aware Alignment*.

(3) Extensive experiments show that SASLM achieves SOTA expressive-speech performance among open-source SLMs with only 3B parameters and 800 hours of expressive speech data.

## 2 Related Work

### 2.1 Semantic-to-Speech Generation in SLMs

Existing SLMs generally adopt two semantic-to-speech generation paradigms (Fig. 2). Modular multi-head models(Xu et al., 2025a; Fang et al., 2025b; Team et al., 2025) use a shared LLM backbone with dedicated speech encoders and generation heads, directly reusing LLM hidden states for acoustic generation. Unified single-head models(Zeng et al., 2024; Nguyen et al., 2025; Maimon et al., 2025) extend the LLM vocabulary with speech tokens and generate interleaved text–speech sequences autoregressively. Despite architectural differences, both paradigms mainly repurpose text-prediction-oriented semantic states for speech generation, without explicitly modeling acoustically usable expressive intent.

### 2.2 Proxy-based Expressive Speech Generation

Early work on expressive speech generation mainly focused on text-to-speech (TTS) systems (Du et al.,

2024; Anastassiou et al., 2024), where large-scale pretraining enables rich acoustic continuation conditioned on human-written text and predefined style prompts. These systems typically rely on externally specified controls, such as speaking styles or emotion descriptions, to guide acoustic realization.

More recently, expressive generation has been extended to SLMs. Lucy (Gao et al., 2025c) and EMOVA (Chen et al., 2025) employ discrete emotion tokens to control speech generation, while Empathy Omni (Wang et al., 2025a) introduces auxiliary emotion objectives to align generated speech with users' emotional states. Nevertheless, these methods still rely on externally proxy supervision, rather than dynamically modeling expressive intent as the dialogue unfolds in proxy-free manner.

### 2.3 Alignment for Speech Generation

Recent reinforcement learning advances (Guo et al., 2025), such as RLHF (Ouyang et al., 2022), AI feedback (Lee et al., 2024), self-rewarding (Fang et al., 2026), and rubric-based reward modeling (Gunjal et al., 2025), have not only enhanced text generation but also extended alignment techniques to the speech domain.

Existing text-to-speech systems commonly use preference optimization methods such as DPO (Rafailov et al., 2023; Du et al., 2024; Gao et al., 2025d), which rely on costly human-annotated preference pairs. Recent work further explores online reinforcement learning for speech generation, including reward modeling for acoustic quality (Gao et al., 2025a,b). In SLMs, SI-SDA (Wang et al., 2025b) leverages saliency-based pseudo-labels to improve speech understanding. However, these methods mainly target utterance-level TTS quality or speech-understanding ability, leaving semantic intent–expressive realization alignment in SLMs largely underexplored.

## 3 Self-Aware Speech Language Model

As shown in Fig. 3, SASLM bridges semantic understanding and acoustic realization through two key designs: an intent-aware expressive modeling architecture (Sec. 3.1) and a realization-aware alignment training paradigm (Sec. 3.2).

### 3.1 Intent-Aware Expressive Modeling Architecture

As illustrated in the left part of Fig. 3, SASLM follows the modular multi-head paradigm (Sec. 2.1) and consists of three key components: (1) a semantic understanding backbone for multimodal perception and text response generation; (2) a VIB-based intent-aware expressive modulation layer that distills latent expressive intents $\mathbf{z}_{y_i}$ from contextual semantic hidden states $\mathbf{h}_{y_i}$ to modulate token embeddings $\mathbf{e}_{y_i}$, yielding intent-aware fused embeddings $\mathbf{f}_{y_i}$; and (3) the speech head generates intent-modulated speech by predicting speech tokens from the fused embeddings and acoustic context $\mathbf{R}_G$, yielding context-aware expressive speech.

#### 3.1.1 Semantic Understanding Backbone

**Semantic Perception and Understanding.** Given a multimodal input $X$ comprising speech $X^S$ and/or text $X^T$, the model jointly generates a textual response $Y^T = \{y_1^T, \ldots, y_N^T\}$ and a speech response $Y^S = \{y_1^S, \ldots, y_M^S\}$. For the speech modality, a perception encoder $\phi_P$ extracts semantic representations $\mathbf{R}_P = \phi_P(X^S)$, which are fused with the text input to form a unified multimodal representation $\mathbf{E}_X$:

$$\mathbf{E}_X = \left[\mathrm{Embed}(X^T);\ \mathrm{Proj}(\mathbf{R}_P)\right], \qquad (1)$$

where $\mathrm{Embed}(\cdot)$ is the embedding lookup, $\mathrm{Proj}(\cdot)$ is linear projection and $[\cdot;\cdot]$ denotes concatenation.

The shared LLM backbone $\mathcal{H}$ then autoregressively models the multimodal context. At decoding step $i$, conditioned on $\mathbf{E}_X$ and the previously generated embeddings $\mathbf{e}_{y_{1:i}}$ with $\mathbf{e}_{y_i} = \mathrm{Embed}(y_i^T)$, it yields:

$$\mathbf{h}_{y_i} = \mathcal{H}(\mathbf{E}_X,\ \mathbf{e}_{y_{1:i}}). \qquad (2)$$

These hidden states capture contextual semantics and serve as shared representations driving both text and speech generation.

**Text Generation.** The text head decodes each hidden state $\mathbf{h}_{y_i}$ into a distribution over the text vocabulary to produce token $y_i^T$.

#### 3.1.2 Intent-aware Expressive Modulation

To let the model exploit its own semantic understanding for context-aware expressive realization, i.e., to be *aware of what it thinks*, we introduce a VIB-based encoder $q_\phi$ that self-distills the LLM hidden states $\mathbf{H}_Y = [\mathbf{h}_{y_1}, \ldots, \mathbf{h}_{y_N}]$ into latent expressive intents $\mathbf{Z}_Y$, as shown in Fig. 3. The bottleneck encourages $\mathbf{Z}_Y$ to retain expressive cues such as emphasis and prosody, while discarding lexical information already captured by the token embeddings $\mathbf{E}_Y$. The distilled intent then modulates $\mathbf{E}_Y$ to produce intent-aware fused embeddings $\mathbf{F}_Y$, which drive speech generation together

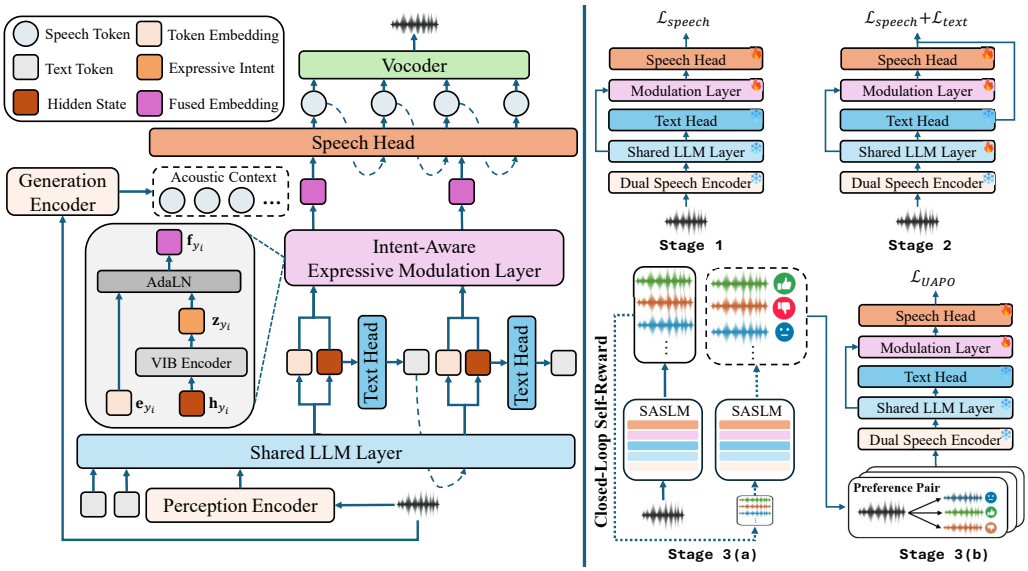


Figure 3: Overview of the SASLM framework. **Left:** Intent-aware architecture with VIB-driven expressive modulation. **Right:** Three-stage realization-aware training paradigm, progressing from acoustic bootstrapping to closed-loop self-reward optimization.

with the generation context $\mathbf{R}_G$ (Sec. 3.1.3). The ideal information-bottleneck objective is:

$$\mathcal{J} = \max_\theta \; I(\mathbf{F}_Y;\, Y^S \mid \mathbf{R}_G) - \beta\, I(\mathbf{Z}_Y;\, \mathbf{E}_Y), \quad (3)$$

where $\beta$ balances expressive acoustic fidelity (first term) and compression that suppresses lexical leakage from $\mathbf{E}_Y$ into $\mathbf{Z}_Y$ (second term).

Directly optimizing Eq. 3 is intractable. Since the VIB encoder takes $\mathbf{H}_Y$ as input, $\mathbf{Z}_Y$ is conditionally independent of $\mathbf{E}_Y$ given $\mathbf{H}_Y$, forming the Markov chain $\mathbf{E}_Y \to \mathbf{H}_Y \to \mathbf{Z}_Y$. By the Data Processing Inequality (DPI) (Cover, 1999), $I(\mathbf{Z}_Y;\mathbf{E}_Y) \le I(\mathbf{Z}_Y;\mathbf{H}_Y)$, so we penalize this tractable upper bound instead. Meanwhile, the realization term is lower-bounded by the conditional reconstruction loss of the speech head. Together, these yield the final variational surrogate objective (see Appendix B.2 for derivation):

$$\tilde{\mathcal{J}}_{\text{VIB}} = \max_{\theta,\phi} \underbrace{\mathbb{E}_{q_\phi}\left[\log p_\theta(Y^S \mid \mathbf{F}_Y, \mathbf{R}_G)\right]}_{\text{expressive realization}} - \beta \underbrace{D_{\text{KL}}\left(q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y) \,\|\, p(\mathbf{Z}_Y)\right)}_{\text{intent compression}}. \quad (4)$$

Since expressive intent unfolds smoothly over an utterance (Nooteboom, 2006; Skerry-Ryan et al., 2018), we model $p(\mathbf{Z}_Y)$ with a discrete-time Ornstein–Uhlenbeck (OU) prior (Uhlenbeck and Ornstein, 1930; Li et al., 2020), which captures temporally correlated latent dynamics that the standard i.i.d. Gaussian prior cannot represent:

$$p(\mathbf{z}_{y_i} \mid \mathbf{z}_{y_{i-1}}) = \mathcal{N}\left(\alpha \mathbf{z}_{y_{i-1}}, \sigma_p^2 \mathbf{I}\right), \quad \alpha \in (0,1), \quad (5)$$

where $\alpha$ controls temporal persistence and $\sigma_p^2$ is the per-step transition variance.

Paired with this OU prior, we instantiate the VIB encoder by parameterizing the variational posterior as a token-wise diagonal Gaussian, i.e., $q_\phi(\mathbf{z}_{y_i} \mid \mathbf{h}_{y_i}) = \mathcal{N}(\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2))$, whose mean and log-variance are predicted from the hidden state $\mathbf{h}_{y_i}$ by a lightweight MLP $f_\phi$. A differentiable intent is then drawn via the reparameterization trick:

$$\mathbf{z}_{y_i} = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}_i, \quad [\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i] = f_\phi(\mathbf{h}_{y_i}), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (6)$$

where $\odot$ denotes element-wise multiplication.

Given the OU prior (Eq. 5) and the Gaussian posterior (Eq. 6), the intent compression term (i.e., the KL divergence in Eq. 4) can be computed analytically at each time step $i$ as:

$$\mathcal{L}_{\text{KL}}(i) = \tfrac{1}{2} \sum_{k=1}^{d} \left[ \log \frac{\sigma_p^2}{\sigma_{i,k}^2} + \frac{\sigma_{i,k}^2 + (\mu_{i,k} - \alpha\, \mu_{i-1,k})^2}{\sigma_p^2} - 1 \right], \quad (7)$$

where $\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i \in \mathbb{R}^d$ are the per-dimension mean and standard deviation, with $\mu_{i,k}$ and $\sigma_{i,k}$ denoting their $k$-th components. This loss drives the expressive intent $\mathbf{z}_{y_i}$ to follow a temporally smooth distribution across timesteps. See Appendix B for the full derivation.

After extracting the latent intent $\mathbf{z}_{y_i}$ via the VIB encoder (Eq. 6), it is injected into the lexical token embeddings $\mathbf{e}_{y_i}$ through Adaptive Layer Normalization (AdaLN) (Peebles and Xie, 2023):

$$\mathbf{f}_{y_i} = \text{AdaLN}(\mathbf{e}_{y_i}, \mathbf{z}_{y_i}) = (1 + \gamma(\mathbf{z}_{y_i})) \odot \frac{\mathbf{e}_{y_i} - \mu}{\sigma} + \delta(\mathbf{z}_{y_i}) \quad (8)$$

where $\gamma(\cdot)$ and $\delta(\cdot)$ are zero-initialized linear projections, and $\mu, \sigma$ are the mean and standard deviation of $\mathbf{e}_{y_i}$. This lightweight modulation preserves lexical content while injecting token-level expressive guidance into speech conditioning.

#### 3.1.3 Intent-Modulated Speech Generation

The speech head autoregressively predicts speech tokens conditioned on the intent-modulated representations $\mathbf{f}_{y_i}$ and context $\mathbf{R}_G$:

$$P_\theta(y_j^S \mid y_{<j}^S, \mathbf{F}_Y, \mathbf{R}_G) = \text{SpeechHead}(y_{<j}^S, \mathbf{F}_Y, \mathbf{R}_G). \quad (9)$$

We train the model by minimizing the negative variational lower bound in Eq.4, which decomposes into an autoregressive speech reconstruction term and the VIB regularizer in Eq.7 that encourages temporally smooth expressive intent:

$$\mathcal{L}_{\text{speech}} = \underbrace{-\sum_{t=1}^{M} \log p_\theta(y_t^S \mid \mathbf{F}_Y, \mathbf{R}_G, y_{<t}^S)}_{\mathcal{L}_{\text{Recon}}} + \beta \underbrace{\sum_{i=1}^{N} \mathcal{L}_{\text{KL}}(i)}_{\mathcal{L}_{\text{VIB}}}. \quad (10)$$

To further adapt speech generation to the user's acoustic state, as the similar content with different paralinguistic cues may require distinct realizations (Wynn and Borrie, 2022), we introduce a generation encoder $\phi_G$ that directly encodes raw input speech: $\mathbf{R}_G = \phi_G(X^S)$. Unlike the semantics-oriented perception pathway in Fig. 2(a), $\mathbf{R}_G$ preserves continuous paralinguistic cues, providing complementary acoustic grounding for context-adaptive speech generation.

### 3.2 Realization-Aware Alignment Training Paradigm

We adopt a progressive realization-aware alignment training paradigm that gradually establishes the connection between semantic understanding and expressive speech generation.

#### 3.2.1 Stage 1: Acoustic Bootstrapping

In this stage, the model establishes basic speech generation ability conditioned on latent intent representations and acoustic context. Only the modulation layer and speech head are trainable, learning to map intent-modulated semantic representations into speech tokens via the combined objective in Eq. 10. This builds the initial pathway from semantic understanding to acoustic realization.

#### 3.2.2 Stage 2: Expressive Intent Grounding

Since the LLM backbone remains frozen in the first stage, its hidden states are primarily optimized for semantic modeling. We thus unfreeze its last layer to jointly optimize text prediction and speech reconstruction, grounding semantic representations in expressive speech generation without compromising the backbone's intelligence:

$$\mathcal{L}_{\text{text}} = -\sum_{i=1}^{N} \log p_\theta(y_i^T \mid y_{<i}^T, \mathbf{E}_X), \quad (11)$$

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{speech}}. \quad (12)$$

This joint optimization progressively shifts LLM hidden states from semantic-dominant to intent-aware acoustically usable representations.

#### 3.2.3 Stage 3: Closed-Loop Self-Reward Alignment

After the first two stages, training remains based on open-loop supervised reconstruction, lacking explicit cross-modal verification of whether generated speech faithfully realizes the expressive intent inferred from dialogue context. To address this, we introduce a closed-loop realization-aware alignment paradigm (details are provided in Algorithm 1, Appendix A) comprising two steps:

**Stage 3(a): Self-Reward Generation.** For each speech input $X^S$, SASLM generates a text response $Y^T$ together with multiple candidate speech rollouts $\{Y_k^S\}_{k=1}^K$. A TTS system additionally produces a context-unaware anchor response, as a minimum acceptable prospect (Ethayarajh et al., 2024) baseline $Y_\perp^S$. The SLM then evaluates each rollout using rubric-based criteria covering emotion, prosody, and naturalness, while WER is used as a hard intelligibility constraint (Zhang et al., 2025). Rollouts are then partitioned into the preferred set $\mathcal{Y}_w$ and the rejected set $\mathcal{Y}_l$ for subsequent alignment according to their scores relative to the anchor $Y_\perp^S$.

**Stage 3(b): Preference-based Alignment.** We further use the obtained preference pairs to align expressive speech generation via Utility-Anchored Preference Optimization (UAPO) (Wang et al., 2025c), which promotes preferred expressive realizations over context-unaware baseline generations and suppresses rejected responses:

$$\begin{aligned} u_\theta(x, y) &= \frac{1}{|y|} \sum_{i=1}^{|y|} \log \pi_\theta(y_i \mid x, y_{<i}) \\ \mathcal{L}_w &= -\sum_{y_w \in \mathcal{Y}_w} \log \frac{\exp\big(u_\theta(X, y_w)\big)}{\exp\big(u_\theta(X, Y_\perp^S)\big) + \sum_{y' \in \mathcal{Y}_w} \exp\big(u_\theta(X, y')\big)} \\ \mathcal{L}_l &= -\log \frac{\exp\big(u_\theta(X, Y_\perp^S)\big)}{\exp\big(u_\theta(X, Y_\perp^S)\big) + \sum_{y_l \in \mathcal{Y}_l} \exp\big(u_\theta(X, y_l)\big)} \\ \mathcal{L}_{\text{UAPO}} &= \mathbb{E}_{(X, \mathcal{Y}_w, \mathcal{Y}_l, Y_\perp^S) \sim \mathcal{B}}\big[\mathcal{L}_w + \mathcal{L}_l\big]. \end{aligned} \quad (13)$$

| Dataset | Stage | #Samples | Hours |
|---|---|---|---|
| InstructS2S (Fang et al., 2025a) | I | 200k | ∼2000 h |
| Genshin Dataset [2] | II | 220k | ∼603 h |
| EmoVoice-DB (Yang et al., 2025) | II | 66k | ∼192 h |
| EmoNet (Schuhmann et al., 2025) | III | 200k | ∼263 h |

Table 1: Datasets used in different training stages.

# 4 Experiments

## 4.1 Experimental Setups

**Model Configuration.** We adopt the Thinker module of Qwen2.5-Omni-3B (Xu et al., 2025a) as the semantic understanding backbone, which employs Whisper as the perception encoder and Qwen2.5-3B as the shared LLM layers with its text head. The speech head is initialized from CosyVoice2 (Du et al., 2024), reusing its S3 tokenizer as the generation encoder and its flow matching module and vocoder for waveform synthesis. Stage-specific datasets are used to support progressive realization-aware alignment, as summarized in Table 1. Dataset construction and training configurations are detailed in Appendix C.1.

**Evaluation Datasets.** We primarily evaluate expressive speech generation on EchoMind (Zhou et al., 2026), an empathy-oriented spoken dialogue benchmark covering seven emotion categories. To verify that expressive generation does not degrade semantic understanding, we further evaluate on the multimodal audio benchmark MMAU (Sakshi et al., 2025) and assess empathetic response quality on EchoMind provided in Appendix C.3.

**Baselines.** (1) **Controllable TTS**, representing oracle expressive control, including CosyVoice2 (Du et al., 2024) and GPT-4o-mini-TTS (Hurst et al., 2024), which synthesize speech from reference replies using predefined TTS prompts; (2) **Vanilla SLMs**, performing direct semantic-to-speech generation without explicit expressive modeling, including Qwen2.5-Omni (3B/7B) (Xu et al., 2025a), Qwen3-Omni-30B (Xu et al., 2025b), GLM-4-Voice (Zeng et al., 2024), Step-Audio2-Mini (Huang et al., 2025), Kimi-Audio (Ding et al., 2025), and UltraVoice-7B (Tu et al., 2025), which is trained on style-controlled expressive data; (3) **Emotional SLMs**, adopting proxy-based generation, represented by EMOVA-7B (Chen et al., 2025), which produces emotion prompts prior to synthesis; (4) **Commercial SLMs**, represented by Doubao [3] as an industrial reference.

## 4.2 Evaluation Protocol

**Utterance-level Objective Evaluation.** To evaluate expressive acoustic realization, we report four categories of objective metrics: (1) **Intelligibility**, measured by Word Error Rate (WER); (2) **Expressiveness**, measured by pitch variance (F0-Var) (Busso et al., 2009); (3) **Semantic–acoustic emotion alignment**, measured by EmoAlign with emotion2vec (Ma et al., 2024) to assess the consistency between text and speech emotions in the response; and (4) **Speech naturalness**, evaluated by Audiobox-Aesthetics (Tjandra et al., 2025), including Content Enjoyment (CE), Usefulness (CU), Production Complexity (PC), Quality (PQ), and their average score (Aes.).

**Conversation-level Subjective Evaluation.** Following (Tu et al., 2025), we employ Gemini-2.5-Flash (Comanici et al., 2025) as an SLM-as-Judge on a 1–5 scale to evaluate whether the generated speech exhibits contextually appropriate emotion (Emo.), prosody (Pro.), naturalness (Nat.), and overall interaction quality (Ovr.). Prompt details are provided in Appendix D.1.

## 4.3 Main Results

As shown in Table 2, SASLM achieves the best overall performance among open-source SLMs and rivals commercial systems such as GPT-4o-mini-TTS and Doubao, a finding further corroborated by human evaluation in Appendix D.2. With only 3B parameters, SASLM consistently surpasses much larger vanilla SLMs, including Qwen3-Omni-30B, delivering richer pitch dynamics (F0-Var: 63.44 vs. 49.76), stronger semantic–acoustic emotion alignment (EmoAlign: 35.61% vs. 25.03%), and higher subjective overall quality (4.33 vs. 4.25). While model scaling and improved semantic intelligence enhance content fidelity and robustness (e.g., Qwen3-Omni-30B's lower WER of 2.76), they do not automatically translate into expressive acoustic realization. This suggests that explicit modeling of expressive intent is more effective than relying solely on implicit semantic learning at scale.

Compared with Controllable TTS systems using oracle expressive prompts as an idealized alignment baseline, SASLM achieves higher subjective prosody (4.36 vs. 4.25), naturalness (4.49 vs. 4.34),

[2] https://www.modelscope.cn/datasets/aihobbyist/Genshin_Dataset

[3] https://www.volcengine.com/docs/6561/1594356

| Model Type | Model Name | Utterance-Level Objective Metrics | | | | Conversation-Level Subjective Metrics | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | WER% ↓ | F0-Var ↑ | EmoAlign% ↑ | CE / CU / PC / PQ / Aes. ↑ | Emo. ↑ | Pro. ↑ | Nat. ↑ | Ovr. ↑ |
| Controllable TTS (w/ Oracle Prompts) | CosyVoice2 | 3.80 | 51.20 | 39.07 | 6.21 / 7.09 / 1.42 / 7.67 / 5.60 | 4.15 | 4.25 | 4.34 | 4.23 |
| | GPT-4o-mini-TTS | 7.63 | 70.44 | 45.04 | 6.22 / 7.37 / 1.41 / 8.00 / 5.75 | 4.27 | 4.43 | 4.58 | 4.41 |
| Commercial SLM | Doubao | 1.12 | 68.58 | 42.95 | 6.21 / 7.38 / 1.53 / 7.94 / 5.72 | 4.22 | 4.50 | 4.59 | 4.44 |
| Vanilla SLM | Kimi Audio | 6.99 | 46.98 | 22.64 | 5.31 / 5.98 / **1.49** / 6.87 / 4.91 | 3.77 | 4.02 | 4.15 | 3.97 |
| | Step Audio2 Mini | 4.87 | 50.92 | 22.92 | 6.12 / 7.06 / 1.44 / 7.68 / 5.58 | 4.00 | 4.19 | 4.43 | 4.15 |
| | GLM-4-Voice | 8.88 | 43.71 | 23.84 | 6.22 / 6.94 / 1.42 / 7.57 / 5.54 | 3.95 | 4.11 | 4.29 | 4.09 |
| | UltraVoice-7B | 4.21 | 49.53 | 17.34 | 6.11 / 7.13 / 1.44 / 7.55 / 5.56 | 4.01 | 4.17 | 4.26 | 4.11 |
| | Qwen2.5-Omni-3B | 4.14 | 42.88 | 21.44 | 6.15 / 6.90 / 1.47 / 7.74 / 5.57 | 3.71 | 3.84 | 3.96 | 3.79 |
| | Qwen2.5-Omni-7B | 3.75 | 47.77 | 23.10 | **6.23** / 7.08 / 1.42 / 7.79 / 5.63 | 3.62 | 3.90 | 4.09 | 3.86 |
| | Qwen3-Omni-30B | **2.76** | 49.76 | 25.03 | **6.23** / 7.14 / 1.41 / 7.70 / 5.62 | 4.08 | 4.28 | 4.42 | 4.25 |
| Emotional SLM | EMOVA-7B | 10.09 | 47.98 | 30.58 | 5.68 / 6.45 / **1.49** / 6.84 / 5.12 | 4.10 | 4.01 | 4.14 | 4.06 |
| Expressive SLM | SASLM-3B (Ours) | 4.56 | **63.44** | **35.61** | 6.12 / **7.19** / 1.41 / **7.82** / **5.64** | **4.24** | **4.36** | **4.49** | **4.33** |

Table 2: Comparison of context-aware expressive speech generation. **Bold** and underline denote the best and second-best results among **open-source SLMs**, excluding commercial SLMs and controllable TTS systems.

and overall interaction quality (4.33 vs. 4.23), despite CosyVoice2 showing slightly better objective emotion alignment with ground-truth emotion labels (39.07% vs. 35.61%). This suggests that oracle labels mainly impose coarse-grained emotion consistency, while SASLM dynamically infers token-level, context-aware expressive intent for finer speech realization, enabling end-to-end SLMs to approach or even surpass oracle-prompted TTS in subjective quality. The similar trend between GPT-4o-mini-TTS and Doubao further supports the benefit of stronger end-to-end expressive modeling for context-aware speech generation.

SASLM also outperforms emotional SLMs that rely on external expressive proxies. Compared with EMOVA-7B, which conditions on self-generated emotion labels, SASLM achieves notable gains in prosody (4.36 vs. 4.01), overall quality (4.33 vs. 4.06), and emotional intent alignment (EmoAlign: 35.61% vs. 30.58%). These results indicate that intrinsic latent intent provides richer, more fine-grained expressive guidance than discrete emotion-centric discrete proxies under explicit supervision.

## 5 Analysis

### 5.1 Architectural View: Intent Modulation Meets Acoustic Grounding

Table 3 ablates two architectural factors after the first two training stages: (1) the **intent modulation strategy** $\mathcal{M}(\mathbf{e}_{y_i}, \mathbf{h}_{y_i})$, for which we compare a progression of increasingly expressive designs: a conversational TTS-style baseline $\pi_1$ using only content embeddings $\mathbf{e}_{y_i}$, naive element-wise fusion $\oplus$ (Xu et al., 2025a), adaptive modulation via AdaLN, and our proposed VIB+AdaLN via a self-distill-then-modulate scheme and (2) the **context-grounding representation** $\mathbf{R}_G$, where

Figure 4: Evaluation accuracy dynamics during UAPO training with different reward models.

$\varnothing$, *Semantic*, and *Acoustic* denote no context, perception-pathway representations ($\mathbf{R}_G = \mathbf{E}_X$), and generation-encoder acoustic features ($\mathbf{R}_G = \phi_G(X^S)$), respectively.

For intent modulation, directly adding semantic hidden states yields marginal gains over $\pi_1$ (Ovr. 4.20 vs. 4.17), but increases WER from 3.61% to 8.29%. AdaLN mitigates this instability (WER 6.50%), whereas the proposed VIB-based AdaLN achieves the best overall quality (Ovr.: 4.31) with a lower WER of 4.74%. This suggests that VIB improves the expression–fidelity trade-off by disentangling expressive intent from lexical content for more stable expressive speech generation.

For context grounding, removing context yields a cleaner but more conservative setting, slightly improving WER (4.19%) but reducing context-appropriate overall quality (4.21). Acoustic grounding improves the overall score to 4.31 with only a modest WER increase (4.74%), underscoring its role in context-adaptive expressive realization. In contrast, semantic grounding severely harms intelligibility (WER 11.97%) without improving overall quality (4.20), suggesting that injecting semantic features alone perturbs acoustic modeling and

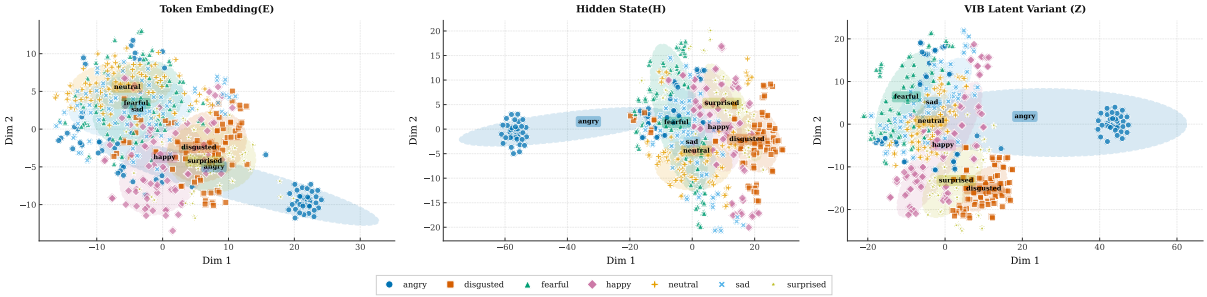


Figure 5: t-SNE visualization of emotion clustering across token embeddings ($\mathbf{e}$), LLM hidden states ($\mathbf{h}$), and distilled VIB latent variables ($\mathbf{z}$). Linear-probe accuracies, computed on the original representations rather than the t-SNE projections, are 18.1%, 41.4%, and 64.5%, respectively.

| $\mathcal{M}(\mathbf{e}_{y_i}, \mathbf{h}_{y_i})$ | $\mathbf{R}_G$ | WER%↓ | Emo.↑ | Pro.↑ | Nat.↑ | Ovr.↑ |
|---|---|---|---|---|---|---|
| *(A) Ablation on Modulation Strategy* | | | | | | |
| $\pi_1$ | Acoustic | **3.61** | 4.06 | 4.22 | 4.30 | 4.17 |
| $\oplus$ | Acoustic | 8.29 | 4.04 | 4.28 | 4.31 | 4.20 |
| AdaLN | Acoustic | 6.50 | 4.15 | 4.28 | 4.34 | 4.22 |
| *(B) Ablation on Context Grounding* | | | | | | |
| VIB + AdaLN | $\varnothing$ | 4.19 | 4.10 | 4.23 | 4.33 | 4.21 |
| VIB + AdaLN | Semantic | 11.97 | 4.08 | 4.24 | 4.36 | 4.20 |
| VIB + AdaLN | Acoustic | 4.74 | **4.21** | **4.34** | **4.45** | **4.31** |

Table 3: Ablation study on the modulation strategy $\mathcal{M}(\mathbf{e}_{y_i}, \mathbf{h}_{y_i})$ and context grounding $\mathbf{R}_G$ in Eq. 9.

destabilizes expressive realization. Full results are provided in Appendix C.4.

### 5.2 Optimization View: Efficient Alignment via Closed-Loop Self-Reward

| Strategy | RM Size | WER(%)↓ | Emo.↑ | Pro.↑ | Nat.↑ | Ovr.↑ |
|---|---|---|---|---|---|---|
| SA-SLM(SFT) | — | 4.74 | 4.21 | 4.34 | 4.45 | 4.31 |
| + Self-Reward | 3B | 4.56 | 4.24 | 4.36 | 4.49 | 4.33 |
| + Oracle-Reward | 30B | **4.21** | **4.30** | **4.41** | **4.54** | **4.38** |

Table 4: Comparison of reward strategies on Intelligibility (WER) and expressive speech quality.

As shown in Table 4, realization-aware alignment with both the 3B Self-Reward and 30B Oracle-Reward scorers consistently outperforms the open-loop SFT baseline, confirming its effectiveness. Notably, expressive quality and intelligibility improve simultaneously during alignment (Ovr. 4.31 → 4.38; WER 4.74% → 4.21%), suggesting that closed-loop optimization mitigates the expressivity–stability trade-off and scales reliably.

As shown in Fig. 4, the 30B critic converges faster and more stably, indicating stronger discrimination of high-quality rollouts and reduced preference-label noise, consistent with the human preference-accuracy results in Appendix D.2.2. Notably, the 3B critic also improves steadily, showing that smaller models can still provide effective alignment signals through the self-reward mechanism, enabling understanding-guided expressive realization without compromising intelligibility.

### 5.3 Representational View: How VIB Distills Expressive Intent

To examine how VIB reshapes expressive representations, Fig. 5 compares emotion separability across token embeddings $\mathbf{e}$, hidden states $\mathbf{h}$, and VIB latent intents $\mathbf{z}$. Token embeddings show heavily overlapping clusters with near-random linear-probe accuracy (18.1%), while hidden states improve separability (41.4%), indicating that LLM semantic understanding contains expressive cues but remain entangled with noise. Notably, VIB latents achieve the clearest clustering and highest accuracy (64.5%), showing that the OU-constrained VIB distills more discriminative expressive intent from entangled semantic states, yielding representations better aligned with expressive acoustic realization.

## 6 Conclusion

In this paper, we identify the gap between semantic understanding and acoustic realization in current SLMs and propose SASLM to bridge it via self-aware intent–realization alignment. With only 3B parameters and 800 hours of expressive data, SASLM achieves SOTA among open-source SLMs, surpassing the 10× larger Qwen3-Omni-30B in F0-Var (63.44 vs 49.76) and overall score (4.33 vs 4.25). Closed-loop alignment further boosts quality (4.31 → 4.38) while reducing WER (4.74% → 4.21%), easing the expressivity–stability trade-off. SASLM also outperforms proxy-based models like EMOVA-7B and rivals commercial SLM such as Doubao, showing that intent modeling efficiently enables expressive generation.

## Limitations

Although SASLM improves expressive speech generation without external emotion labels, several limitations remain. First, our evaluation mainly focuses on English empathetic dialogues in EchoMind, leaving multilingual, cross-cultural, and domain-specific expressive behaviors underexplored. Second, stronger expressive synthesis may make synthetic voices more persuasive, increasing risks of impersonation, emotional manipulation, and deceptive audio. Deployments should therefore incorporate synthetic-speech disclosure, watermarking or traceability when feasible, and safeguards for emotionally sensitive use cases.

## Ethics Statement

AI-assisted tools are used only for writing polishing and language refinement. All scientific contributions, experiments, conclusions, and citations are manually developed and verified by the authors. All training and evaluation data are from publicly available or open-source datasets. We strictly follow the original licenses, usage restrictions, and ethical requirements specified by the dataset providers, without applying additional special processing such as extra personal-information filtering.

## References


Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):582–596.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, and 1 others. 2025. Emova: Empowering language models to see, hear and speak with vivid emotions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5455–5466.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

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*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025a. Llama-omni: Seamless speech interaction with large language models. In *International Conference on Learning Representations*.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025b. Llama-omni 2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18617–18629.

Wenkai Fang, Shunyu Liu, Yang Zhou, Kongcheng Zhang, Tongya Zheng, Kaixuan Chen, Mingli Song, and Dacheng Tao. 2026. Serl: Self-play reinforcement learning for large language models with limited data. *Advances in Neural Information Processing Systems*, 38:103706–103738.

Changfeng Gao, Zhihao Du, and Shiliang Zhang. 2025a. Differentiable reward optimization for llm-based tts system. *arXiv preprint arXiv:2507.05911*.

Changfeng Gao, Yabin Li, Keyu An, Zhifu Gao, Zhihao Du, Han Zhao, and Xiangang Li. 2025b. Explore the reinforcement learning for the llm-based asr and tts system. *arXiv preprint arXiv:2509.18569*.

Heting Gao, Hang Shao, Xiong Wang, Chaofan Qiu, Yunhang Shen, Siqi Cai, Yuchen Shi, Zihan Xu, Zuwei Long, Yike Zhang, and 1 others. 2025c. Lucy: Linguistic understanding and control yielding early stage of her. *arXiv preprint arXiv:2501.16327*.

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F. Chen. 2025d. 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.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.

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, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, and 1 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

D Robert Ladd. 2008. *Intonational phonology*. Cambridge University Press.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. 2020. Scalable gradients for stochastic differential equations. In *International conference on artificial intelligence and statistics*, pages 3870–3882. PMLR.

Yinghao Aaron Li, Xilin Jiang, Jordan Darefsky, Ge Zhu, and Nima Mesgarani. 2024. Style-talker: Finetuning audio language model and style-based text-to-speech model for fast spoken dialogue generation. *arXiv preprint arXiv:2408.11849*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760.

Gallil Maimon, Michael Hassid, Amit Roth, and Yossi Adi. 2025. Scaling analysis of interleaved speech-text language models. *arXiv preprint arXiv:2504.02398*.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, and 1 others. 2025. Spirit-lm: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52.

S.G. Nooteboom. 2006. *The prosody of speech: melody and rhythm*, pages 640–673. Basil Blackwell Limited.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, pages 53728–53741.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*.

Klaus R Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2):227–256.

Christoph Schuhmann, Robert Kaczmarczyk, Gollam Rabby, Felix Friedrich, Maurice Kraus, Kourosh Nadi, Huu Nguyen, Kristian Kersting, and Sören Auer. 2025. Emonet-voice: A fine-grained, expert-verified benchmark for speech emotion detection. *arXiv preprint arXiv:2506.09827*.

Elizabeth Shriberg. 2005. Spontaneous speech: How people really talk and why engineers should care. In *Proceedings of Interspeech 2005*, pages 1781–1784.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *International Conference on Machine Learning*, pages 4693–4702.

Tongyi Fun Team, Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, and 1 others. 2025. Fun-audio-chat technical report. *arXiv preprint arXiv:2512.20156*.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, and 1 others. 2025. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*.

Wenming Tu, Guanrou Yang, Ruiqi Yan, Wenxi Chen, Ziyang Ma, Yipeng Kang, Kai Yu, Xie Chen, and Zilong Zheng. 2025. Ultravoice: Scaling fine-grained style-controlled speech conversations for spoken dialogue models. *arXiv preprint arXiv:2510.22588*.

George E Uhlenbeck and Leonard S Ornstein. 1930. On the theory of the brownian motion. *Physical review*, 36(5):823.

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

Shaowen Wang, Xinyuan Chen, and Yao Xu. 2025b. Self-improvement for audio large language model using unlabeled speech. *arXiv preprint arXiv:2507.20169*.

Xiaobo Wang, Zixia Jia, Jiaqi Li, Qi Liu, and Zilong Zheng. 2025c. Adaptive preference optimization with uncertainty-aware utility anchor. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19204–19225.

Camille J. Wynn and Stephanie A. Borrie. 2022. Classifying conversational entrainment of speech behavior: An expanded framework and review. *Journal of Phonetics*, 94:101173.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, and 18 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, and Xie Chen. 2025. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10748–10757.

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

Jun Zhan, Mingyang Han, Yuxuan Xie, Chen Wang, Dong Zhang, Kexin Huang, Haoxiang Shi, Dongxiao Wang, Tengtao Song, Qinyuan Cheng, and 1 others. 2025. Vstyle: A benchmark for voice style adaptation with spoken instructions. *arXiv preprint arXiv:2509.09716*.

Xueyao Zhang, Chaoren Wang, Huan Liao, Ziniu Li, Yuancheng Wang, Li Wang, Dongya Jia, Yuanzhe Chen, Xiulin Li, Zhuo Chen, and 1 others. 2025. Speechjudge: Towards human-level judgment for speech naturalness. *arXiv preprint arXiv:2511.07931*.

Li Zhou, Lutong Yu, You Lyu, Yihang Lin, Zefeng Zhao, Junyi Ao, Yuhao Zhang, Benyou Wang, and Haizhou Li. 2026. EchoMind: An interrelated multi-level benchmark for evaluating empathetic speech language models. In *Proceedings of the International Conference on Learning Representations*.

Qingxiaoyang Zhu, Austin Chau, Michelle Cohn, Kai-Hui Liang, Hao-Chuan Wang, Georgia Zellou, and Zhou Yu. 2022. Effects of emotional expressiveness on voice chatbot interactions. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–11.

**Algorithm 1** Closed-Loop Self-Reward Alignment

**Input:** Dataset $\mathcal{D}$, WER threshold $\tau$, TTS model $\theta_{\text{TTS}}$, SLM $\theta_{\text{SLM}}$, prompts $p_{\text{emo}}$, $p_{\text{pro}}$, $p_{\text{nat}}$, $p_{\text{asr}}$, sample size $N$
Initialize replay buffer $\mathcal{B} \leftarrow \emptyset$
*# Stage 3(a): Self-Reward*
**for** each $X^S \in \mathcal{D}$ **do**
  *# Generate the text response and N speech rollouts*
  $Y^T,\ \{Y_k^S\}_{k=1}^N \leftarrow \theta_{\text{SLM}}(X^S)$
  $Y_\perp^S \leftarrow \theta_{\text{TTS}}(Y^T)$  *// Context-unaware anchor*
  *# Generate rubrics for each dimension*
  $\mathcal{R}_d \leftarrow \theta_{\text{SLM}}(X^S,\ Y^T,\ p_d), \quad d \in \{\text{emo, pro, nat}\}$
  $\mathcal{R} \leftarrow \{\mathcal{R}_{\text{emo}},\ \mathcal{R}_{\text{pro}},\ \mathcal{R}_{\text{nat}}\}$
  *# Evaluate rollouts and anchor against rubrics*
  **for** each $Y \in \{Y_k^S\}_{k=1}^N \cup \{Y_\perp^S\}$ **do**
    $s_{\text{emo}},\ s_{\text{pro}},\ s_{\text{nat}} \leftarrow \theta_{\text{SLM}}(Y,\ \mathcal{R}) \in \{0,\ 0.5,\ 1\}^3$
    $\text{wer} \leftarrow \text{WER}\big(\theta_{\text{SLM}}(Y,\ p_{\text{asr}}),\ Y^T\big)$
    $r(Y) \leftarrow 0$ **if** wer $> \tau$ **else** $(s_{\text{emo}} + s_{\text{pro}} + s_{\text{nat}})$
  **end for**
  $\mathcal{G}_+ \leftarrow \{Y_k^S \mid r(Y_k^S) > r(Y_\perp^S)\}$
  $\mathcal{G}_- \leftarrow \{Y_k^S \mid r(Y_k^S) < r(Y_\perp^S)\}$
  **if** $\mathcal{G}_+ \neq \emptyset$ **and** $\mathcal{G}_- \neq \emptyset$ **then**
    $\mathcal{Y}_w \leftarrow \big\{y \in \mathcal{G}_+ \mid r(y) = \max_{y' \in \mathcal{G}_+} r(y')\big\}$
    $\mathcal{Y}_l \leftarrow \big\{y \in \mathcal{G}_- \mid r(y) = \min_{y' \in \mathcal{G}_-} r(y')\big\}$
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{(X^S,\ \mathcal{Y}_w,\ \mathcal{Y}_l,\ Y_\perp^S)\}$
  **end if**
**end for**
*# Stage 3(b): Offline Optimization*
**for** each $(X^S,\ \mathcal{Y}_w,\ \mathcal{Y}_l,\ Y_\perp^S) \in \mathcal{B}$ **do**
  Compute $\mathcal{L}_{\text{UAPO}}$ via Eq. (13) and update $\theta_{\text{SLM}}$
**end for**

# A Detailed Algorithm for Self-Reward Alignment

In this section, we provide the detailed pseudocode for the closed-loop self-reward alignment paradigm introduced in Stage 3. Algorithm 1 outlines the complete pipeline, including data construction, rubric-based evaluation, and preference pair selection. Specifically, we adopt the English subset of EmoNet (Schuhmann et al., 2025) as the question source. Given each question, SASLM first specifies the intended expression along three dimensions—emotional intensity, prosody, and naturalness. The generated rollouts are then scored against predefined rubrics to assess their adherence to the intended expression. The scoring template for the emotion dimension is illustrated in Figure 8.

# B Derivation of the VIB-Driven Disentanglement Objective

Section 3.1.2 introduced the ideal information bottleneck objective (Eq. 3) and invoked the Data Processing Inequality (DPI) (Cover, 1999) to derive a tractable surrogate (Eq. 4). This appendix presents the complete derivation, including the Markov chain formulation, the variational bounds, and the closed-form KL divergence under an Ornstein–Uhlenbeck (OU) (Uhlenbeck and Ornstein, 1930) prior.

## B.1 The Ideal Objective and the DPI Surrogate

The intent-aware architecture aims to extract a pure expressive intent $\mathbf{Z}_Y$ that maximizes acoustic realization fidelity ($Y^S$) while discarding lexical redundancy already encoded in the text embeddings ($\mathbf{E}_Y$). This yields the ideal information bottleneck objective:

$$\mathcal{J} = \max_\theta \ I(\mathbf{F}_Y;\, Y^S \mid \mathbf{R}_G) - \beta\, I(\mathbf{Z}_Y;\, \mathbf{E}_Y) \quad (14)$$

where $\mathbf{F}_Y$ is the intent-modulated representation, $\mathbf{R}_G$ the acoustic context, and $\beta$ the compression strength.

Since $I(\mathbf{Z}_Y;\, \mathbf{E}_Y)$ is intractable, we exploit the fact that $\mathbf{Z}_Y$ is sampled solely from the LLM hidden states $\mathbf{H}_Y$ via the VIB encoder $q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y)$, with no direct access to $\mathbf{E}_Y$. The resulting conditional independence $p(\mathbf{Z}_Y \mid \mathbf{H}_Y, \mathbf{E}_Y) = p(\mathbf{Z}_Y \mid \mathbf{H}_Y)$ defines the Markov chain:

$$\mathbf{E}_Y \to \mathbf{H}_Y \to \mathbf{Z}_Y. \quad (15)$$

By the DPI, information cannot increase along a Markov chain, which guarantees that:

$$I(\mathbf{Z}_Y;\, \mathbf{E}_Y) \le I(\mathbf{Z}_Y;\, \mathbf{H}_Y), \quad (16)$$

so $I(\mathbf{Z}_Y;\, \mathbf{H}_Y)$ serves as a tractable upper bound for penalization.

**Physical Interpretation of the Surrogate.** Substituting this upper bound into the objective yields the surrogate:

$$\tilde{\mathcal{J}} = \max_{\theta,\phi} \ I(\mathbf{F}_Y;\, Y^S \mid \mathbf{R}_G) - \beta\, I(\mathbf{Z}_Y;\, \mathbf{H}_Y) \quad (17)$$

Penalizing $I(\mathbf{Z}_Y;\, \mathbf{H}_Y)$ rather than $I(\mathbf{Z}_Y;\, \mathbf{E}_Y)$ is in fact desirable: it compels $\mathbf{Z}_Y$ to discard not only redundant lexical tokens (already captured in $\mathbf{E}_Y$) but also irrelevant contextual semantics, retaining *only* the expressive cues essential for acoustic realization.

## B.2 Variational Bounds for the Surrogate Objective

As exact mutual information is intractable for high-dimensional neural representations, we derive variational bounds for both terms in $\tilde{\mathcal{J}}$.

**Lower Bound for Expressive Realization.** By applying the factorization

$$p(\mathbf{F}_Y, Y^S \mid \mathbf{R}_G) = p(Y^S \mid \mathbf{F}_Y, \mathbf{R}_G)\, p(\mathbf{F}_Y \mid \mathbf{R}_G),$$

the conditional mutual information admits the decomposition:

$$I(\mathbf{F}_Y; Y^S \mid \mathbf{R}_G) = \mathbb{E}_p\big[\log p(Y^S \mid \mathbf{F}_Y, \mathbf{R}_G)\big] + H(Y^S \mid \mathbf{R}_G) \quad (18)$$

where $\mathbb{E}_p$ denotes the expectation over the empirical data distribution. Since the true posterior $p(Y^S \mid \mathbf{F}_Y, \mathbf{R}_G)$ is unknown, we introduce a parameterized decoder $p_\theta(Y^S \mid \mathbf{F}_Y, \mathbf{R}_G)$ (i.e., the speech generation head). By the non-negativity of the KL divergence, $D_{\mathrm{KL}}(p \,\|\, p_\theta) \geq 0$, we directly obtain:

$$\mathbb{E}_p\big[\log p(Y^S \mid \mathbf{F}_Y, \mathbf{R}_G)\big] \geq \mathbb{E}_p\big[\log p_\theta(Y^S \mid \mathbf{F}_Y, \mathbf{R}_G)\big] \quad (19)$$

Substituting this into Eq. (18) yields the Variational Lower Bound (VLB). Since $H(Y^S \mid \mathbf{R}_G)$ is independent of $\theta$, maximizing this bound is equivalent to minimizing the negative log-likelihood. Applying autoregressive factorization over the length-$M$ speech sequence yields the reconstruction loss:

$$\mathcal{L}_{\mathrm{Recon}} = -\sum_{t=1}^{M} \log p_\theta\big(y_t^S \mid \mathbf{F}_Y, \mathbf{R}_G, y_{<t}^S\big) \quad (20)$$

Note that $\mathbf{R}_G = \phi_G(X^S)$, produced by an independent perception encoder, lies outside the Markov chain $\mathbf{E}_Y \to \mathbf{H}_Y \to \mathbf{Z}_Y$, leaving the DPI argument unaffected.

**Upper Bound for Intent Compression.** For the second term $I(\mathbf{Z}_Y; \mathbf{H}_Y)$, the mutual information is defined as:

$$I(\mathbf{Z}_Y; \mathbf{H}_Y) = \mathbb{E}_{p(\mathbf{Z}_Y, \mathbf{H}_Y)}\Big[\log \frac{p(\mathbf{Z}_Y \mid \mathbf{H}_Y)}{p(\mathbf{Z}_Y)}\Big] \quad (21)$$

Since the true posterior $p(\mathbf{Z}_Y \mid \mathbf{H}_Y)$ is generally intractable, we introduce a variational encoder $q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y)$ to approximate it. Multiplying the log-term by $\frac{q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y)}{q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y)} = 1$ and splitting the expectation yields:

$$\begin{aligned} I(\mathbf{Z}_Y; \mathbf{H}_Y) = {} & \mathbb{E}_{p(\mathbf{Z}_Y, \mathbf{H}_Y)}\Big[\log \frac{q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y)}{p(\mathbf{Z}_Y)}\Big] \\ & + \mathbb{E}_{p(\mathbf{H}_Y)}\Big[D_{\mathrm{KL}}\big(p(\mathbf{Z}_Y \mid \mathbf{H}_Y) \,\|\, q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y)\big)\Big]. \end{aligned} \quad (22)$$

By the non-negativity of the KL divergence, i.e., $D_{\mathrm{KL}}(p \,\|\, q_\phi) \geq 0$, dropping the second term gives the inequality:

$$I(\mathbf{Z}_Y; \mathbf{H}_Y) \leq \mathbb{E}_{p(\mathbf{Z}_Y, \mathbf{H}_Y)}\Big[\log \frac{q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y)}{p(\mathbf{Z}_Y)}\Big] \quad (23)$$

Factorizing $p(\mathbf{Z}_Y, \mathbf{H}_Y) = p(\mathbf{H}_Y)\, p(\mathbf{Z}_Y \mid \mathbf{H}_Y)$ and using $q_\phi$ as the sampling distribution (via the reparameterization trick) for the inner expectation, we arrive at the Variational Upper Bound (VUB):

$$I(\mathbf{Z}_Y; \mathbf{H}_Y) \leq \mathbb{E}_{p(\mathbf{H}_Y)}\Big[D_{\mathrm{KL}}\big(q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y) \,\|\, p(\mathbf{Z}_Y)\big)\Big] \quad (24)$$

Minimizing this upper bound restricts the information capacity of $\mathbf{Z}_Y$, forcing it to discard redundant content from $\mathbf{H}_Y$.

### B.3 Ornstein–Uhlenbeck Prior and Tractable KL Divergence

**Choice of the OU Prior.** Specifying the prior $p(\mathbf{Z}_Y)$ is essential for computing the VUB in Eq. (24), as it serves as the reference distribution that regularizes the encoder $q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y)$. Standard VAEs typically assume an independent standard normal prior, which ignores the temporal continuity inherent in expressive speech (e.g., prosody and emotion evolve smoothly). To capture this temporal inertia, we model the prior as a discrete-time Ornstein–Uhlenbeck (OU) process—a mean-reverting first-order autoregressive, AR(1), process. The joint prior factorizes as:

$$p(\mathbf{Z}_Y) = \prod_{i=1}^{M} p(\mathbf{z}_{y_i} \mid \mathbf{z}_{y_{i-1}}). \quad (25)$$

where $p(\mathbf{z}_{y_i} \mid \mathbf{z}_{y_{i-1}}) = \mathcal{N}(\alpha \mathbf{z}_{y_{i-1}}, \sigma_p^2 \mathbf{I})$.

Here, $\alpha \in [0, 1)$ controls the temporal smoothness, and $\sigma_p^2$ is the prior variance.

Assuming the variational encoder extracts the expressive intent $\mathbf{z}_{y_i}$ from a Gaussian posterior $q_\phi(\mathbf{z}_{y_i} \mid \mathbf{h}_{y_i}) = \mathcal{N}(\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2))$ via the reparameterization trick:

$$\mathbf{z}_{y_i} = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (26)$$

the sequence-level KL divergence admits a chain-rule decomposition:

$$\begin{aligned} & D_{\mathrm{KL}}(q_\phi(\mathbf{Z}_Y \mid \mathbf{H}_Y) \,\|\, p(\mathbf{Z}_Y)) \\ & = \sum_{i=1}^{M} \mathbb{E}_{q_\phi(\mathbf{z}_{y_{i-1}} \mid \mathbf{h}_{y_{i-1}})}\Big[D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{y_i} \mid \mathbf{h}_{y_i}) \,\|\, p(\mathbf{z}_{y_i} \mid \mathbf{z}_{y_{i-1}})\big)\Big]. \end{aligned} \quad (27)$$

**Mean Substitution and the Stop-Gradient Operator.** Assuming diagonal-Gaussian parameterizations for the posterior $q_\phi(\mathbf{z}_{y_i} \mid \mathbf{h}_{y_i})$ and the

conditional prior $p(\mathbf{z}_{y_i} \mid \mathbf{z}_{y_{i-1}})$, the inner KL divergence for a $d$-dimensional latent space can be computed analytically:

$$\begin{aligned}\mathcal{L}_{\mathrm{KL}}(i) &= D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{y_i} \mid \mathbf{h}_{y_i}) \,\|\, p(\mathbf{z}_{y_i} \mid \mathbf{z}_{y_{i-1}})\big) \\ &= \frac{1}{2}\sum_{k=1}^{d}\left[\log\frac{\sigma_p^2}{\sigma_{i,k}^2} + \frac{\sigma_{i,k}^2 + (\mu_{i,k} - \alpha z_{y_{i-1},k})^2}{\sigma_p^2} - 1\right].\end{aligned} \tag{28}$$

Which is exactly equal to Eq. 7. When evaluating the outer expectation over $q_\phi(\mathbf{z}_{y_{i-1}} \mid \mathbf{h}_{y_{i-1}})$, only the squared difference term depends on $\mathbf{z}_{y_{i-1}}$:

$$\begin{aligned}&\mathbb{E}_{q_\phi(\mathbf{z}_{y_{i-1}}|\mathbf{h}_{y_{i-1}})}\big[(\mu_{i,k} - \alpha z_{y_{i-1},k})^2\big] \\ &= (\mu_{i,k} - \alpha\mu_{i-1,k})^2 + \alpha^2\sigma_{i-1,k}^2.\end{aligned} \tag{29}$$

For computational simplicity, we approximate this expectation via mean substitution, absorbing the scaled variance $\alpha^2\sigma_{i-1,k}^2$ into the prior variance $\sigma_p^2$. Crucially, we apply a stop-gradient operator, $\mathrm{sg}[\cdot]$, to the previous mean $\boldsymbol{\mu}_{i-1}$:

$$\mathbb{E}_{q_\phi(\mathbf{z}_{y_{i-1}}|\mathbf{h}_{y_{i-1}})}\big[\alpha\,\mathbf{z}_{y_{i-1}}\big] \approx \alpha\,\mathrm{sg}[\boldsymbol{\mu}_{i-1}]. \tag{30}$$

Aggregating this per-step KL divergence over the sequence yields the closed-form Variational Information Bottleneck (VIB) loss: $\mathcal{L}_{\mathrm{VIB}} = \sum_{i=1}^{M}\mathcal{L}_{\mathrm{KL}}(i)$.

### B.4 Final Tractable Objective

Combining the reconstruction lower bound (Eq. (20)) and the intent compression upper bound (Eq. (7)), we arrive at the final tractable loss function for the VIB-driven disentanglement:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{Recon}} + \beta\,\mathcal{L}_{\mathrm{VIB}}. \tag{31}$$

By minimizing $\mathcal{L}_{\mathrm{total}}$, the model is forced to encode only the temporally smooth, acoustic-essential expressive cues into $\mathbf{Z}_Y$, successfully filtering out the redundant lexical information already present in $\mathbf{E}_Y$.

**Optimization Dynamics of the Stop-Gradient.** Beyond simplifying the expectation, the stop-gradient $\mathrm{sg}[\boldsymbol{\mu}_{i-1}]$ is essential for preventing representation collapse. Without it, the optimizer could trivially minimize the temporal penalty $(\mu_{i,k} - \alpha\,\mu_{i-1,k})^2$ by collapsing all $\boldsymbol{\mu}_i$ to a constant vector. The $\mathrm{sg}[\cdot]$ operator forces the current state $\boldsymbol{\mu}_i$ to actively track the historical trajectory of $\boldsymbol{\mu}_{i-1}$ without retroactively altering past representations. This ensures the OU prior acts strictly as a causal low-pass filter, enforcing temporal smoothness while preserving information capacity.

## C Experiments Details

### C.1 Training Configuration.

The learning rate of speech head is set to $1 \times 10^{-5}$, while the LLM backbone is updated at $1 \times 10^{-6}$ during Stage 2 expressive grounding. The OU-prior hyperparameters are fixed to $\alpha = 0.95$ and $\sigma_p = 0.5$. To stabilize VIB optimization, the KL coefficient $\beta$ is initialized to 0 during the first 10% warm-up steps and then annealed to 0.5 following a cosine schedule. During realization-aware alignment, we sample $K = 32$ candidate speech rollouts per input and use a WER threshold $\tau = 0.2$ as the intelligibility constraint. Each training stage uses 8 H20 GPUs and takes approximately two days to complete.

### C.2 Dataset Construction

The training corpus is built in two layers: a **text-level emotional QA completion** layer that yields complete emotion-labeled quadruples $(q, e_q, a, e_a)$, , where $q$ and $a$ denote the user query and assistant answer, respectively, and $e_q$ and $e_a$ denote their corresponding emotion labels, and an **emotion-controllable TTS** layer that synthesizes the corresponding audio.

**Text-level emotional QA completion.** Raw conversational sources are heterogeneous and frequently incomplete: some entries provide only an emotion-labeled reply, others only a query, and many provide no emotion label at all. We therefore design a set of symmetric LLM-based completion pipelines that fill the missing side from any observed side, producing a complete tuple $(q, e_q, a, e_a, c)$ where $c$ is a short audio instruction consumed by the downstream TTS. The completion is governed by three high-level constraints—*causal coherence*, *closed-vocabulary emotion labels*, and *controllable style instructions*—whose detailed definitions are given in the prompt in Figure 9. Each call returns a structured output, and malformed completions are discarded.

**Emotion-controllable TTS.** Given a complete tuple from the text layer, the missing side is synthesized by an emotion-controllable TTS module that conditions on both a reference voice and the audio instruction $c$, so that speaker identity and emotional style are decoupled and independently controllable. The audio corpus is built around two goals: (i) **speaker coverage**, where the user query side samples reference voices from a voice pool while the

| Model Name | Audio Understanding | | | | Empathetic Response | | | |
|---|---|---|---|---|---|---|---|---|
| | Sound | Music | Speech | Avg. | CF | RN | Col. | IR |
| Qwen2.5-Omni-3B | 30.24 | 49.25 | 39.04 | 39.51 | 4.54 | 4.66 | 4.51 | 3.24 |
| Qwen2.5-Omni-7B | **78.10** | **65.90** | **70.60** | **71.50** | **4.59** | 4.58 | **4.51** | **3.28** |
| **SASLM-3B (Ours)** | 29.94 | 49.25 | 39.63 | 39.60 | 4.56 | **4.65** | 4.40 | 3.24 |

Table 5: Comparison of different speech language models on audio understanding (MMAU) and empathetic response ability (EchoMind). **Bold** = best; underline = second-best.

assistant answer side uses a single fixed voice to preserve character consistency; and (ii) **emotion alignment**, requiring the paralinguistic realization of the synthesized speech to be consistent with the text-side emotion label. Under these objectives, we reuse the same TTS module to produce three types of audio data, serving *grounding*, *supervised training*, and *preference learning*, respectively.

### C.3 Impact of Expressive End-to-End Training on Semantic Intelligence

To verify that end-to-end training for expressive speech does not compromise semantic intelligence, we evaluate SASLM on two complementary benchmarks: EchoMind (Zhou et al., 2026) for empathetic dialogue quality, assessed along four dimensions—Context Fit (CF), Response Naturalness (RN), Colloquialism (Col.), and Implicit Recognition (IR)—and MMAU-v05.15.25 (Sakshi et al., 2025) for general audio understanding, reported as single-choice accuracy across the sound, music, and speech domains.

As shown in Table 5, SASLM achieves an MMAU average of 39.60, on par with its base model Qwen2.5-Omni-3B (39.51), indicating that our VIB-driven modulation incurs no catastrophic forgetting of audio understanding. On EchoMind, SASLM matches the 3B baseline across all four dimensions, with a marginal gain on CF (4.56 vs. 4.54) and parity on IR (3.24). Together, these results demonstrate that enhanced acoustic expressiveness can be attained without sacrificing—and in some aspects mildly complementing—the model's semantic and empathetic reasoning capabilities.

### C.4 Architecture Ablation: Context Grounding and Intent Modulation

Table 6 reports the complete ablation results across all combinations of fusion strategy and context grounding after the first two training stages, extending the summary in Section 5.1.

**Intent Modulation Strategy.** Table 6 isolates how semantic intent is injected into speech generation under acoustic grounding. The conversational TTS baseline ($\pi_1$) attains strong intelligibility (WER 3.61%) but limited expressiveness (Ovr. 4.17), confirming the absence of a pathway from *what the model thinks* to *how it speaks*. **Naive injection harms intelligibility with little expressive gain.** Direct summation ($\oplus$) entangles fast-varying lexical content with token-level acoustics, inflating WER to 8.29% while barely improving expressiveness (Ovr. 4.20). Vanilla AdaLN treats $\mathbf{h}_{y_i}$ as a modulation signal rather than an additive perturbation, reducing WER to 6.50%, yet lexical information still leaks without regularization. **VIB regularization resolves intent transmission failure.** VIB+AdaLN compresses $\mathbf{h}_{y_i}$ into a temporally smooth intent variable $\mathbf{z}_{y_i}$ via an OU prior, suppressing lexical fluctuations while preserving utterance-level intent. This yields the best expressiveness (Ovr. 4.31, Emo. 4.21, Pro. 4.34, Nat. 4.45) with near-baseline intelligibility (WER 4.74%), confirming that disentangling intent from content is essential for effective expressive modulation.

**Context Grounding.** We next vary the grounding signal $\mathbf{R}_G$ across all modulation strategies. Removing context ($\varnothing$) improves intelligibility but weakens expressiveness: under VIB+AdaLN, $\varnothing$ yields WER 4.19%/Ovr.4.21 versus 4.74%/4.31 with acoustic grounding, indicating that acoustic context contributes to both pronunciation stability and prosodic variation. **Semantic grounding is incompatible with reliable synthesis.** Replacing acoustic features with perception-pathway representations ($\mathbf{R}_G = \mathbf{E}_X$) yields the worst WER across all strategies (9.62%–15.40%), as such representations lack the fine-grained acoustic continuity needed for phoneme-level articulation. **Acoustic grounding is indispensable for jointly preserving intelligibility and expressiveness.** $\phi_G(X^S)$ supplies the speech head with continuous acoustic representations that retain phonetic detail while remaining decoupled from the LLM semantic pathway, better capturing the one-to-many nature of spoken interaction(Wynn and Borrie, 2022).

| Model Type | Architecture Design | | Intell. | Expressiveness | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\mathcal{M}(\mathbf{e}_{y_i}, \mathbf{h}_{y_i})$ | $\mathbf{R}_G$ | WER(%)↓ | F0-Var↑ | Emo.↑ | Pro.↑ | Nat.↑ | Ovr.↑ |
| Conversational TTS | $\pi_1$ | ∅ | **3.53** | 55.88 | 4.04 | 4.20 | 4.30 | 4.17 |
| | | Semantic | 9.62 | 61.16 | 4.02 | 4.18 | 4.28 | 4.16 |
| | | Acoustic | 3.61 | 60.42 | 4.06 | 4.22 | 4.30 | 4.17 |
| Naive Addition | $\oplus$ | ∅ | 7.91 | 57.00 | 4.06 | 4.22 | 4.32 | 4.19 |
| | | Semantic | 15.40 | 57.61 | 4.11 | 4.24 | 4.35 | 4.21 |
| | | Acoustic | 8.29 | 57.10 | 4.04 | 4.28 | 4.31 | 4.20 |
| Vanilla AdaLN | AdaLN | ∅ | 5.66 | 59.03 | 4.00 | 4.17 | 4.25 | 4.14 |
| | | Semantic | 15.13 | 61.14 | 4.02 | 4.19 | 4.27 | 4.15 |
| | | Acoustic | 6.50 | **62.24** | 4.15 | 4.28 | 4.34 | 4.22 |
| SASLM (Ours) | VIB+AdaLN | ∅ | 4.19 | 56.34 | 4.10 | 4.23 | 4.33 | 4.21 |
| | | Semantic | 11.97 | 56.43 | 4.08 | 4.24 | 4.36 | 4.20 |
| | | Acoustic | 4.74 | 58.37 | **4.21** | **4.34** | **4.45** | **4.31** |

Table 6: Ablation study on expressive-aware speech generation across two independent design choices in Eq. 9. **(1) Intent Modulation Strategy** $\mathcal{M}(\mathbf{e}_{y_i}, \mathbf{h}_{y_i})$: how semantic intent is injected for expressive generation, where $\pi_1$: $\mathcal{M}(\mathbf{e}_{y_i}, \mathbf{h}_{y_i}) = \mathbf{e}_{y_i}$ discards SLM semantic intent (conversational TTS-style baseline); $\oplus$: naive element-wise addition $\mathbf{e}_{y_i} + \mathbf{h}_{y_i}$ (Xu et al., 2025a); AdaLN (Wang et al., 2025a): adaptive modulation without VIB; VIB+AdaLN: our proposed disentangled modulation. **(2) Context Grounding** $\mathbf{R}_G$: how context is constructed during synthesis, where ∅ denotes no context grounding, *Semantic* denotes perception-pathway representations $\mathbf{R}_G = \mathbf{E}_X$, and *Acoustic* denotes generation-encoder acoustic features $\mathbf{R}_G = \phi_G(X^S)$.

# D Evaluation Details

## D.1 Prompt Design for Conversation-level Subjective Evaluation

We employ the Gemini-2.5-Flash API for conversation-level subjective evaluation along three dimensions: *emotion and intensity*, *prosody*, and *naturalness*. All three prompts follow a unified two-stage protocol, with the emotion-and-intensity prompt shown in Fig. 10 as a representative example. The model first infers the expected delivery (target emotion and intensity) from the dialogue context and response text, then listens to the synthesized audio to judge whether the actual delivery matches this expectation. By decoupling *expected expression* from *acoustic execution*, this design promotes context-grounded judgments that go beyond surface acoustic quality to assess whether the speech faithfully conveys the dialogue's communicative intent.

## D.2 Human Evaluation

We recruited three annotators based on two criteria: (1) an IELTS score of 6.5 or above, ensuring sufficient English listening ability to judge fine-grained expressive cues; and (2) a computer-science background with research experience or basic research training, helping them understand the evaluation criteria and produce consistent judgments. During annotation, each audio clip is presented together with its subtitle so that annotators can focus on acoustic expressiveness rather than being affected by speech-recognition difficulty.All reported statistics are aggregated across the three annotators.

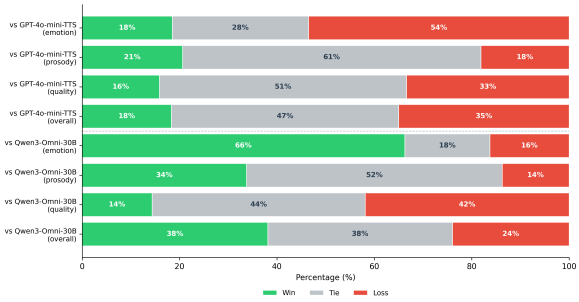


Figure 6: Pairwise human Win/Tie/Loss between our model and each baseline (GPT-4o-mini-TTS, Qwen3-Omni-30B) on the three perceptual dimensions, together with an aggregated *Overall* row per baseline. *Win* (green) indicates that our model is preferred, *Loss* (red) indicates that the baseline is preferred, and *Tie* (gray) indicates perceptually comparable.

### D.2.1 Speech Performance Assessment

To assess perceptual speech performance, we conduct a pairwise human comparison study. For each utterance, annotators independently judge our model against each baseline along three percep-

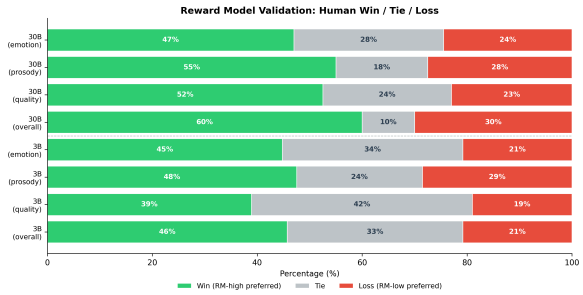


Figure 7: Human Win/Tie/Loss validation of reward models. Each row corresponds to a (model, dimension) configuration. *Win* (green) indicates that the RM-high sample is preferred by human raters, *Loss* (red) indicates that the RM-low sample is preferred, and *Tie* (gray) indicates that the two samples are judged perceptually comparable.

tual dimensions – *Emotion*, *Prosody*, and *Quality* – yielding a *Win*, *Loss*, or *Tie* verdict per (annotator, utterance, dimension) triplet; an aggregated *Overall* verdict is summarized as the average of the three per-dimension Win/Tie/Loss rates. Figure 6 reports the aggregated Win/Tie/Loss distributions, from which we observe that our model is clearly preferred over the strong open-source baseline Qwen3-Omni-30B (overall Net Win Rate $+14.2\%$) but trails the proprietary GPT-4o-mini-TTS (overall Net Win Rate $-16.7\%$); the remaining *Quality* deficit against both baselines is mainly caused by occasional acoustic artifacts (e.g. mild breaking, breathiness, unstable energy), which we attribute to the limited scale of our training data and which also explain the slightly higher WER of our model in the main results.

#### D.2.2 Reward-Model Validation

To verify whether reward-model preferences are aligned with human judgments, we conduct a blind pairwise comparison study. For each pair, the reward model assigns a higher score to one sample (RM-high) and a lower score to the other (RM-low). The two samples are then presented to annotators in a randomized order, while their RM-induced ranking is hidden, and annotators compare them on *Emotion*, *Prosody*, and *Quality* as a Win, Loss, or Tie. As shown in Figure 7, the reward-model-selected samples are generally preferred by human annotators, indicating a positive alignment between reward-model rankings and human perceptual judgments.

## E Prompt Template

**Prompt for Emotion Delivery Evaluation**

**Role**

You are a professional *Speech Quality Reviewer*. Your task is to assess whether a synthesized speech faithfully realizes the specified emotional directive, based strictly on its acoustic delivery.

**Inputs**

- **Emotion Directive**: The intended emotion and arousal level.
- **Target Transcript**: The textual content to be spoken.
- **Response Speech**: The synthesized audio to be evaluated.

**Core Principles**

- **Acoustic-Grounded**: Judgments must rely solely on *acoustic features* (pitch, intonation, rate, energy), **not** on textual semantics.
- **Anti-Exaggeration**: If the delivery sounds flat, describe it honestly. **Never** project emotions onto the speech to satisfy the directive.
- **Objective Description**: Specify both the *perceived emotion type* and the *arousal level* (low / medium / high).

**Scoring Rubric** (`status` $\in \{0, 0.5, 1\}$)

- **1 — Match**: Delivery fully aligns with the directive.
- **0.5 — Partial**: Target cues present but intensity insufficient (e.g., *rage* directive delivered as *mild annoyance*).
- **0 — Mismatch**: Delivery clearly contradicts the directive (e.g., *sad* read *cheerfully*).

**Output Format**

```
{
  "check_emotion_directive": {
    "instruction": "<copy of emotion_directive, verbatim>",
    "observation": "<objective acoustic-based description>",
    "status": <0 | 0.5 | 1>
  }
}
```

Figure 8: Prompt for evaluating emotion delivery in synthesized speech.

Prompt for Text-level Emotional QA Completion

**Role.** You are a dialogue designer crafting emotionally engaging, natural conversations.
**Task.** Given one side of a dialogue (e.g., Character B's emotion-labeled reply), generate the missing counterpart (Character A's preceding line) with its emotion label and an audio instruction specifying how it should be spoken.

**(1) Causal coherence.** The generated line must serve as a natural emotional cause of the observed side, such that its removal would render the reply incomplete. The model must justify this causality in a `causality_check` field, avoiding generic reactions (e.g., "What happened?") and lexical repetition of the observed side.

**(2) Closed-vocabulary emotion labels.** The emotion must be *a single word* drawn *verbatim* from a predefined closed vocabulary of 20 labels, covering 14 primary emotions (e.g., Happy, Sad, Angry, ...) and 6 extended ones (e.g., Gentle, Playful, ...). Compound labels, modifiers (e.g., "Very"), and out-of-vocabulary terms are forbidden.

**(3) Controllable style instruction.** The audio instruction is a 3–8 word descriptor ending with `<|endofprompt|>`, conveying exactly one emotion at one of two intensity levels (*Normal* or *High*). Softening modifiers (e.g., *slightly*, *mildly*), multi-emotion combinations, and indirect framings (e.g., "Barely holding back anger") are forbidden. Example: "Strong sorrowful tone`<|endofprompt|>`".

Figure 9: Prompt for text-level emotional QA completion. The figure illustrates the "answer-to-question" direction; other directions follow the same template by swapping the observed and generated sides.

**Prompt for Emotion and Intensity Evaluation**

**Objective**

Given a conversational context, a response text, and its synthesized speech, evaluate whether the speech delivery is emotionally appropriate and contextually calibrated.

**Inputs**

- **Speech Input**: User's utterance providing conversational grounding.
- **Response Text**: Target textual content to be synthesized.
- **Response Speech**: Generated speech of the response.

**Procedure**

1. **Expectation**: Infer the target emotion and intensity (e.g., empathetic, encouraging, calm, enthusiastic, neutral) from the input and response text *before* listening.
2. **Execution**: Listen to the response speech and compare actual delivery with expectation, checking for tonal appropriateness and intensity calibration.

**Scoring Rubric**

- **5 — Perfect**: Precise, evocative, perfectly calibrated.
- **4 — High**: Well aligned; minor issues in subtle transitions.
- **3 — Acceptable**: Basic tendency conveyed but generic or mildly mis-scaled.
- **2 — Poor**: Clearly inconsistent or substantially miscalibrated.
- **1 — Unacceptable**: Flat, inappropriate, or uncomfortable delivery.

**Output Format**

```
{
  "Expected_Emotion_and_Intensity": "...",
  "Execution_Analysis": "...",
  "Score": <1-5>
}
```

Figure 10: Prompt for evaluating emotion and intensity in synthesized speech.

**Guidelines for Human MOS Annotation**

**Task.** You will listen to a short speech utterance generated by a speech model in a multi-turn dialogue. The dialogue subtitle is provided. Please rate the audio on a 1–5 scale along the three perceptual dimensions below, and additionally provide a single *Overall* score that summarizes your holistic preference, where $1$ = very poor and $5$ = excellent.

**Dimensions.**

- **Emotion.** Is the emotional delivery appropriate for the content? Does it match what a speaker would naturally feel in this context?
- **Prosody.** Is the pitch, rhythm, stress and pause pattern rich and natural? Is the speech free of monotone or robotic delivery?
- **Quality.** Is the audio signal clean? Listen for glitches, distortion, breaking, breathiness, and unstable energy.
- **Overall.** A holistic judgment of how good this audio is as a response in this conversation, considering expressiveness, naturalness and clarity together.

**Procedure (MOS study).** Listen to each audio at least once with the subtitle visible, and provide one integer rating per dimension.

**Procedure (reward-model validation).**
The ordering of samples is randomized and the reward-model ranking is hidden from annotators. Two audio samples are presented as a pair. Considering the same perceptual criteria (Emotion, Prosody, and Quality), select

- **Win**: RM-high is preferred;
- **Loss**: RM-low is preferred;
- **Tie**: no clear perceptual preference.

Figure 11: Annotation guideline shown to human evaluators in both the MOS study and the reward-model validation study.