DISTILLING TEXT LMS INTO SPEECH-TO-SPEECH LMS BY MATCHING INTERNAL TEXT PREDICTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text language models (LMs) acquire broad linguistic and factual capabilities through large-scale pretraining, while comparable capabilities remain difficult to learn from speech alone. Recent work has shown that knowledge distillation from pretrained text LMs offers a promising way to transfer their linguistic and factual capabilities to speech models. However, existing distillation approaches have largely focused on speech-input, text-output models. This idea has yet to be explored for *speech-to-speech* models, which operate on both speech input and speech output. In this work we introduce a speech-to-speech pretraining method that creates an internal text interface for knowledge transfer while preserving the speech-only input and output. The model first consumes speech context without its transcription, switches to text prediction, and then uses the predicted text to lead future speech generation word by word. It also provides an interface for knowledge distillation: the internal text predictions can be directly distilled from the base text LM. To the best of our knowledge, this is the first application of text LM distillation to a speech-to-speech LM. Under matched training conditions, the proposed sequence construction substantially improves the generative perplexity of spoken continuations over speech-only and interleaved speech-text models, approaching the performance of a text-leading inner-monologue model with external ASR prompt transcription. More importantly, distillation improves factual performance under both text and speech evaluation, indicating better retention of knowledge from the base text language model.

1 INTRODUCTION

Large-scale text pretraining provides the foundation for acquiring broad linguistic and factual capabilities. A text language model can first learn from massive unlabeled text corpora and subsequently transfer these capabilities to downstream settings such as instruction following, dialogue,

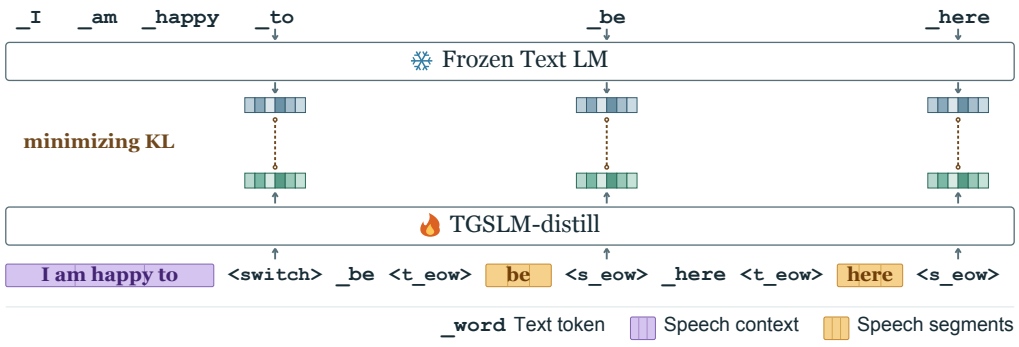


Figure 1: Our TGSLM architecture. Given a speech context, the model generates subsequent words sequentially, with each text word followed by its corresponding speech segments. In TGSLM-distill, the text-token predictions are additionally trained to match those of a frozen text LM.

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.

and agents (Ouyang et al., 2022). An analogous goal for speech language models (SLMs) is to acquire such capabilities directly from large-scale speech. However, scaling analyses suggest that comparable linguistic capabilities are harder to obtain from speech-only pretraining (Cuervo & Marxer, 2024; Maimon et al., 2025b). In particular, under matched training compute budgets, performance on linguistic evaluations improves more slowly with scale for speech than for text. This motivates making more direct use of text during speech pretraining, both as an explicit training signal and as an intermediate prediction target for knowledge distillation from pretrained text language models.

One way to use text is to incorporate it directly into the speech language modeling objective. Existing SLMs do so in different ways, leading to different tradeoffs. Interleaved models (Nguyen et al., 2025; Chou et al., 2023) train on sequences that alternate between speech and text spans from the same utterance, typically switching modalities at word boundaries, and can perform speech-to-speech inference without any text input. In this setting, text provides an additional training signal, but speech generation does not condition on text, resulting in worse linguistic performance (Chou et al., 2026). Text-leading, or inner-monologue, models (Défossez et al., 2024) instead generate text ahead of speech so text predictions directly condition subsequent speech generation. In prompted continuation, however, this typically requires the transcript of the speech prompt at inference to maintain the text-leading stream. Table 1 summarizes this tradeoff. This raises a question: how can we use text to guide linguistic prediction in SLMs without requiring prompt text at inference?

Model	Speech-only prompt	Text-guided generation
Text leading	✗	✓
Interleaved	✓	✗
TGSLM	✓	✓

Table 1: How text is used at inference in SLMs.

We introduce Text-Guided Speech LM (TGSLM), which approaches this question by separating future speech prediction into its linguistic content and its acoustic realization, as shown in the bottom part of Fig. 1. Rather than predicting future speech directly from the speech context, we decompose generation into first predicting the future linguistic content in text and then generating the corresponding speech conditioned on the predicted text. During training, we insert a switch token at a randomly selected word boundary in the speech sequence. Before the switch, the model consumes speech directly. After the switch, it generates the continuation word by word, first predicting each text word and then generating its corresponding spoken realization. At inference, the switch token explicitly triggers the transition to text-guided speech generation. This construction captures complementary advantages of interleaved and text-leading models: text directly guides future speech generation, while inference requires no transcription of the speech prompt.

The TGSLM prediction structure substantially improves generative perplexity (genPPL). TGSLM has a lower genPPL for spoken continuations than speech-only baselines, approaching the performance of a text-leading model without requiring the prompt transcript at inference. These results suggest that text provides a more effective target for learning the linguistic content of future speech, since the continuation is learned directly in text rather than through the speech prediction objective.

The explicit text prediction also provides a direct interface for distillation from the pretrained text LM. We distill the frozen text backbone into TGSLM by minimizing the KL divergence between their next-token distributions at text-prediction positions and refer to the resulting model as TGSLM-distill (combined with text replay, Sec. 3.3). Distillation improves factual performance under both text-to-text and speech-to-speech evaluation, showing that capabilities from the pretrained text model can be better maintained during speech adaptation.

Our main contributions are:

- We study a speech-to-speech pretraining formulation (TGSLM) in which we predict future text before its corresponding speech, while the input remains speech-only at inference. Under matched training conditions, this formulation substantially improves spoken continuation modeling over speech-only and interleaved baselines and approaches an inner-monologue model with an ASR-transcribed prompt text.

- We distill a pretrained text LM into TGSLM's intermediate text predictions. Distillation improves factual performance in both speech-to-speech and text-to-text evaluation, indicating better retention of the pretrained LM's capabilities during speech adaptation. To the

best of our knowledge, this is the first application of distillation from text LMs to speech-to-speech LMs.

- Additionally, we evaluate sensitivity to paralinguistic cues and find that models with internal text prediction and distillation remain comparably sensitive to emotional prosody when prompted with speech sharing the same linguistic content.

2 Related work

2.1 Pure Speech Language Models

Pure SLMs learn to model speech without using text as part of the objective. Early work on generative spoken language modeling showed that speech can be modeled autoregressively by discretizing self-supervised speech representations into "semantic" tokens (Lakhotia et al., 2021).[1] TWIST initializes a pure SLM from a pretrained text LM, which improves performance on linguistic evaluations(Hassid et al., 2023). SLMs over continuous speech features combine autoregressive generation with a generative head, using conditional flow matching (Chou et al., 2025) or diffusion (Rouard et al., 2026), to jointly learn linguistic and acoustic information. Llama-Mimi flattens semantic tokens and acoustic tokens (the remaining RVQ codebooks, which capture acoustic detail needed for waveform reconstruction) into a single autoregressive sequence (Sugiura et al., 2025). These approaches provide ways of learning pure speech models, either without text supervision or with text pretraining, but their performance on tasks involving linguistic content remains limited (Chou et al., 2026). This motivates introducing *explicit* text supervision during speech pretraining.

2.2 Speech-Text Language Models

Speech-text language models directly predict both speech and text sequences. Interleaved SLMs alternate between speech and text spans of the same utterance, typically switching modalities at word boundaries (Chou et al., 2023; Nguyen et al., 2025). In this formulation, text provides an additional training signal, but is not used explicitly when generating a speech continuation.

Another family is inner-monologue modeling, where text prediction precedes and guides speech generation. Spectron follows a speech prompt $\rightarrow$ prompt transcription $\rightarrow$ text continuation $\rightarrow$ speech continuation structure (Nachmani et al., 2024). Moshi uses a word-aligned inner monologue in which text precedes the corresponding generated audio (Défossez et al., 2024). FLM-Audio uses sentence-level monologues that let text run ahead of speech (Yao et al., 2025). Thinker-Talker architectures such as Qwen2.5-Omni (Xu et al., 2025) modularize a related text-before-speech dependency, with a separate Talker generating speech from the Thinker's text and hidden representations.

TGSLM combines complementary aspects of interleaved and inner-monologue models. At word boundaries, rather than switching between speech and text spans, TGSLM transitions to text-guided generation: similar to inner-monologue models, it first predicts the text and then generates the corresponding speech realization.

Moshi (Défossez et al., 2024) and FLM-Audio (Yao et al., 2025) use related speech-input, text-guided speech generation structures. However, they use it in dialogue post-training, where it is learned together with conversational behaviors such as turn-taking and full-duplex interaction. We instead use it in speech pretraining, isolating its effect on linguistic modeling from dialogue training.

2.3 Knowledge Transfer from Text Language Models

As text language models acquire broad linguistic and factual capabilities, a natural question is how to transfer these capabilities to speech models. Most prior distillation methods are speech-input, text-output models. BLSP-KD, DiVA, and SALAD follow this general paradigm, using speech–text pairs to align speech-conditioned predictions or representations with their text-conditioned counterparts (Wang et al., 2024; Held et al., 2025; Cuervo et al., 2026). SALAD further studies forgetting of text capabilities during speech adaptation and combines cross-modal distillation with text replay.

[1] This term refers to tokens that represent the linguistic (not only semantic) content.

Speech-to-speech models, however, produce no text distribution to distill into. Align-SLM transfers semantic knowledge from a text LLM through preference optimization: it generates multiple speech continuations, scores their semantic quality using an LLM, and constructs preference pairs for DPO (Lin et al., 2025). This allows knowledge transfer without introducing text prediction, but the teacher signal is mediated through generated speech and sequence-level preferences.

TGSLM instead introduces an explicit text prediction interface within a speech-to-speech model. This allows the pretrained text LM to directly supervise TGSLM's next-token predictions, while the model still takes speech as input and produces speech as its final output.

3 METHOD

3.1 SPEECH REPRESENTATION AND GENERATION OBJECTIVE

We follow the speech representation and generation objective of Flow-SLM (Chou et al., 2025; 2026). Given an input waveform, the Mimi tokenizer (Défossez et al., 2024) produces synchronized discrete and continuous speech representations. We use the first Mimi RVQ (Zeghidour et al., 2022) codebook as discrete semantic tokens $\mathbf{z} = (z_1, \ldots, z_M)$ and the pre-quantized Mimi features as continuous acoustic representations $\mathbf{x} = (x_1, \ldots, x_M)$.

At each speech position m, the Transformer (Vaswani et al., 2017) encodes the preceding context into a hidden state h_m. A discrete speech head autoregressively predicts the next $k = 4$ semantic tokens, $p_\theta(z_{m:m+k-1} \mid h_m) = \prod_{i=0}^{k-1} p_\theta^{(i)}(z_{m+i} \mid h_m, z_{m:m+i-1})$, with corresponding negative log-probability loss $\mathcal{L}_{\text{discrete}}$.

To model acoustic information beyond the discrete semantic tokens, a conditional flow-matching (CFM) (Lipman et al., 2023) head models $p_\theta(x_m \mid h_m, z_{m:m+k-1})$, conditioning the continuous representation on both the Transformer state and the predicted semantic tokens. Following (Chou et al., 2026), we use the x-prediction parameterization (Li & He, 2026) and optimize the corresponding CFM loss $\mathcal{L}_{\text{CFM}}$. The full speech-generation objective is

$$\mathcal{L}_{\text{speech}} = \mathcal{L}_{\text{discrete}} + \mathcal{L}_{\text{CFM}}. \tag{1}$$

3.2 TEXT-GUIDED SEQUENCE CONSTRUCTION

Fig. 1 illustrates the sequence construction. We first obtain word-level speech-text alignments using a Whisper-based alignment method (Yeh et al., 2025). For each training utterance, we randomly sample a word boundary and insert a <switch> token at that position. Speech preceding the switch is used only as context and incurs no loss.

After the switch, generation proceeds word by word. Let T_i denote the sequence of text tokens corresponding to the i^{th} word after the switch point, and let S_i denote its speech segment. The sequence is constructed as

$$S_{\text{context}} \rightarrow \texttt{<switch>} \rightarrow T_1 \rightarrow S_1 \rightarrow T_2 \rightarrow S_2 \rightarrow \cdots. \tag{2}$$

Thus, each continuation word is first predicted as text tokens and then generated as its corresponding speech segment before the model proceeds to the next word.

At text positions, we apply the standard autoregressive cross-entropy loss to the tokens in T_i, conditioned on the full preceding speech-text context. At speech positions, we apply $\mathcal{L}_{\text{speech}}$. Only the output head corresponding to the current modality is trained at each position.

To mark the end-of-word boundaries between text and speech generation, we introduce <t-eow> and <s-eow> tokens. The model predicts <t-eow> after the final text token of each word and <s-eow> after the corresponding speech segment. We treat <s-eow> as an additional symbol in the vocabulary of the first discrete-token prediction $p_\theta^{(0)}$.

3.3 KNOWLEDGE DISTILLATION FROM THE TEXT LM

The explicit text predictions provide a direct interface for knowledge distillation from the pretrained text LM. We use the original backbone LM, before speech adaptation, as a frozen teacher. At each

text-prediction position, the frozen teacher conditions on the ground-truth text corresponding to the speech context up to that position, while the student is conditioned on its speech context (prompt) and the speech-text history after the switch token. Our distillation loss is the forward KL divergence between their next-token distributions (Hinton et al., 2015):

$$\mathcal{L}_{\mathrm{KD}} = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}\left(p_{\mathrm{teacher}}(y_t \mid c_{<t}^{\mathrm{text}}) \parallel p_\theta(y_t \mid c_{<t}^{\mathrm{st}})\right)\right], \tag{3}$$

where $c_{<t}^{\mathrm{text}}$ and $c_{<t}^{\mathrm{st}}$ denote the teacher and student contexts before timestep t, and y_t denotes the text token at step t. T denotes the text token sequence length. We interpolate the standard text prediction loss and the distillation loss. The full training objective is

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{\mathrm{text}} + \lambda\mathcal{L}_{\mathrm{KD}} + \mathcal{L}_{\mathrm{speech}}, \tag{4}$$

where λ controls the distillation strength. Setting $\lambda = 0$ recovers training without distillation. We also train with text-only batches to mitigate forgetting. For replay-enabled models, a text-only batch is sampled with probability 0.2 and optimized using the standard next-token prediction objective.

3.4 Inference

At inference time, a prompt is provided as speech context. We append `<switch>` immediately after the speech prompt to trigger text-guided generation. The model then alternates between text and speech generation, constructing the word-level text-speech alignment online.

Word boundaries are determined by a greedy criterion: A segment ends when its end-of-word symbol becomes the highest-ranked prediction. For speech, the current segment ends when

$$\arg\max_v p_\theta^{(0)}(v \mid h_m) = \texttt{<s-eow>}, \tag{5}$$

and the text segment ends when `<t-eow>` is the highest-ranked text token. The model generates text tokens until `<t-eow>`, switches to speech generation, where it samples discrete speech tokens and generates the corresponding continuous representations until `<s-eow>`, and then returns to text generation for the next word. This process continues until the text stream produces an `<EOS>` token, after which generation terminates upon completing the corresponding final speech segment.

For likelihood-based evaluation (StoryCloze), we score only text-token positions in the continuations. Boundary handling is described in Appendix A.3.

4 Experiments

4.1 Data and Training

We train on the English subset of Emilia (He et al., 2024), adding the English subset of Emilia-YODAS for the scaled (TGSLM-Large) model. We filter alignments for quality (Appendix A.2). Training and decoding hyperparameters are in Appendix A.6. We use FineWeb 10B subset for text replay (Penedo et al., 2024).

For evaluation, we synthesize speech for StoryCloze, Known-1000, and a dataset we create called Gemini-Emo to test emotion sensitivity. For StoryCloze, following prior work (Lu et al., 2026; Cuervo et al., 2026; Maimon et al., 2025b), we synthesize the story context and candidate continuations with Kokoro-TTS using the `af_heart` voice.[2] For Known-1000, we synthesize the factual prefixes using NVIDIA Magpie TTS with the `Sofia` voice[3].

To test emotion sensitivity, we construct Gemini-Emo: 96 triplets of happy, sad, and angry utterances synthesized with emotional prosody, sharing a linguistically neutral spoken prefix that serves as the prompt (details and an example in Appendix A.5).

To avoid tokenizer artifacts at the end of prompts, we trim trailing silence and pad each waveform to a multiple of the tokenizer frame duration (details in Appendix A.4).

[2] `https://huggingface.co/hexgrad/Kokoro-82M`

[3] `https://huggingface.co/nvidia/magpie_tts_multilingual_357m`

4.2 Model Architecture

All models are initialized from pretrained Qwen3 base LMs (Yang et al., 2025): Qwen3-0.6B for TGSLM variants and Qwen3-1.7B for the scaled TGSLM-Large. Unless otherwise specified, **TGSLM** is trained without distillation or text replay, and **TGSLM-distill** uses both. We report λ when it varies.

4.3 Evaluation

Generative perplexity (genPPL). We use genPPL to evaluate the linguistic quality of open-ended speech continuations, following (Chou et al., 2026; Maimon et al., 2025a). Given a speech prompt, each model generates five continuations. We transcribe the generated speech with Whisper `small.en` (Radford et al., 2023) and score the continuation transcript with OLMo-3 7B (OLMo Team, 2025), conditioned on the prompt transcript. Lower genPPL indicates that the generated continuation is more likely under the reference text LM. We score only the continuation tokens so that genPPL reflects continuation quality rather than the likelihood of the given prompt. To identify the continuation in the ASR transcription, we locate the prompt boundary by finding the minimum character edit distance between the ASR transcription and the prompt text, following (Chou et al., 2026). We use the intersection of utterances with non-empty continuations across all compared models, so that every model is evaluated on the same subset.

Auto-BLEU. Following prior work (Lakhotia et al., 2021; Lin et al., 2025; Maimon et al., 2025a), we additionally report bigram auto-BLEU to measure repetition. Auto-BLEU computes the fraction of generated N-grams ($N = 2$) that appear more than once within the utterance. This complements generative perplexity, since repetition can result in an artificially low perplexity under a reference language model (Holtzman et al., 2020; Welleck et al., 2020; Zhang et al., 2026).

StoryCloze. We evaluate spoken StoryCloze (Maimon et al., 2025a; Lu et al., 2026; Hassid et al., 2023) by comparing candidate likelihoods, computed over text-token predictions only (Sec. 3.4).

Factual knowledge. GenPPL averages over all continuation tokens and may not fully reflect whether specific factual knowledge is retained. We evaluate factual knowledge using Known-1000 (Meng et al., 2022),[4] which tests factual completion through open-ended continuation and more closely matches our generation setting than question-answering benchmarks. For each prompt, we generate five spoken continuations and transcribe them with Whisper `small.en`. Following Spectron (Nachmani et al., 2024), a sample is considered correct if the answer appears in its transcription. We count an example as correct under Pass@K if at least one of the first K generated samples is correct.

Emotion sensitivity. We evaluate emotion sensitivity on Gemini-Emo using the prompt triplets described in Sec. 4.1. For each generated sample, we apply the emotion2vec-plus-large (Ma et al., 2024) classifier to the generated continuation, with happy, angry, and sad emotion classes.

Acoustic quality and speaker similarity. We evaluate acoustic quality using UTMOS-v2 (Baba et al., 2024). We measure speaker similarity between the speech prompt and generated continuation using WavLM-large-TDNN, following (He et al., 2024; Le et al., 2023).[5]

5 Results

5.1 Text-guided speech generation improves genPPL

As shown in Table 2, TGSLM reduces genPPL from 151.7 for the speech-only baseline to 58.9 at the same model and data scale, while interleaved training does not improve over the speech-

[4] `https://rome.baulab.info/data/dsets/known_1000.json`

[5] `https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification`

Table 2: Generation-based metrics on the Emilia test set: genPPL, bigram auto-BLEU (BLEU), speaker similarity (Spk), and UTMOS-v2 (UTMOS) (Baba et al., 2024); see Sec. 4.3 for details. For the **Whisper prompt transcript** row in text-leading references, the model is conditioned on a transcript of the speech prompt produced by Whisper `small.en`. **TGSLM-Large-distill** uses $\lambda = 0.5$ with text replay. All rows are evaluated on the same subset: a shared subset of 1,325 prompts (of 1,500) with non-empty continuations from every model, including external models in Appendix A.7, covering 4,957 of 7,500 continuations.

Model	Direction	Params	Data	GenPPL ↓	BLEU↓	Spk ↑	UTMOS ↑
Ground truth	–	–	–	43.8	6.0	0.74	2.82
Speech-only baselines (no text guidance)							
Speech-only	S → S	0.8B	21.3k h	151.7	4.3	0.42	2.56
Interleaved	S → S	0.8B	21.3k h	164.1	5.4	0.37	2.36
TGSLM variants							
TGSLM	S → (T, S)	0.8B	21.3k h	58.9	3.9	0.36	2.41
+ self-generated prompt text	(S, T) → (T, S)	0.8B	21.3k h	62.2	3.8	0.38	2.40
+ distillation ($\lambda = 0.25$)	S → (T, S)	0.8B	21.3k h	56.9	4.3	0.36	2.41
+ distillation + text replay	S → (T, S)	0.8B	21.3k h	51.0	4.9	0.38	2.44
TGSLM-Large-distill	S → (T, S)	1.9B	135k h	45.0	5.7	0.45	2.68
Text-leading references							
Oracle prompt text	(T, S) → (T, S)	0.8B	21.3k h	40.7	5.4	0.42	2.50
Whisper prompt transcript	(T, S) → (T, S)	0.8B	21.3k h	57.1	6.5	0.41	2.50

only baseline.[6] Distillation and text replay further reduce genPPL to 56.9 and 51.0, respectively, indicating that transferring the text LM's distribution also improves open-ended continuation. Auto-BLEU remains relatively stable for TGSLM variants and speech-only baselines, indicating that the genPPL improvements do not come from more repetitive continuations.

We also test whether explicitly predicting the speech prompt text provides further benefit. In the **+ self-generated prompt text** row in Table 2, the prompt is represented as an alternating sequence $S_1 \to T_1 \to S_2 \to T_2 \to \cdots$, where each speech word segment is followed by its generated text transcription. After the model consumes the prompt, generation proceeds with standard TGSLM text-guided continuation (text before speech). At inference, where the prompt transcript is unavailable, the model consumes the speech prompt sequentially and identifies word boundaries online using the same `<s-eow>` criterion as in Eq. 5. This variant does not improve over TGSLM, suggesting that explicitly recovering the prompt text is not necessary to obtain the benefits of text-guided continuation.

We also compare against two text-leading references, conditioned on either the ground-truth or the Whisper-transcribed prompt text. With oracle prompt text, the text-leading model achieves lower genPPL (40.7) than TGSLM (58.9). However, replacing the oracle transcript with the Whisper-generated transcript degrades performance. TGSLM approaches the text-leading model conditioned on a Whisper transcript (58.9 vs. 57.1) without requiring prompt transcription. This suggests that explicitly providing prompt text is not necessary to obtain most of the benefit of text-guided generation. The more important factor is predicting future text before generating its corresponding speech.

For acoustic metrics, the controlled TGSLM models show slightly lower speaker similarity and UTMOS-v2 than the speech-only baseline, consistent with the tradeoff observed in prior work (Chou et al., 2026). The differences are modest, and the scaled TGSLM-Large-distill model recovers both metrics.

5.2 Distillation improves factual knowledge retention

Fig. 2 shows the results of ablation over text replay and distillation weight. On Known-1000 (left), text replay alone already improves factual performance relative to TGSLM, showing that maintain-

[6] Under matched update counts, the interleaved model's sequence budget is shared between speech and text tokens, so it sees roughly half the speech per update compared with the speech-only and TGSLM models.

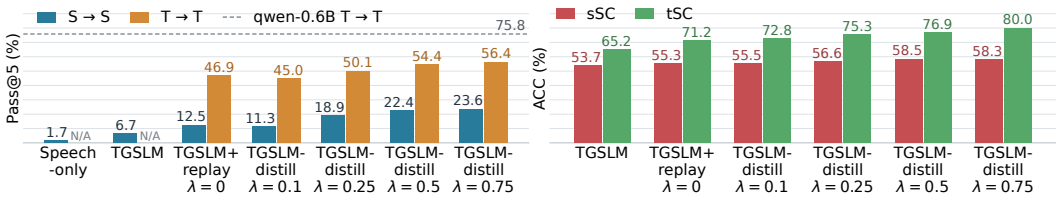


Figure 2: Ablation of text replay and distillation weight λ. TGSLM uses neither. TGSLM + replay adds text replay only ($\lambda = 0$); TGSLM-distill adds both. Left: Pass@5 (%) on Known-1000. Right: accuracy (%) on spoken (sSC) and text (tSC) StoryCloze (Hassid et al., 2023).

Table 3: Known-1000 evaluation. Pass@K(%) measures whether the answer appears in at least one of K sampled continuations. Direction S $\rightarrow$ (T, S) denotes a speech prompt followed by a text-guided speech continuation. **TGSLM-Large-distill** uses a distillation weight of $\lambda = 0.5$ and text replay.

Model	Direction	Params	Data	Pass@1	Pass@5
Text-LM reference					
Qwen3-0.6B backbone (Yang et al., 2025)	T $\rightarrow$ T	0.6B	–	46.8	77.9
Qwen3-1.7B backbone (Yang et al., 2025)	T $\rightarrow$ T	1.7B	–	60.8	86.6
Our models					
TGSLM-distill	S $\rightarrow$ (T, S)	0.8B	45k h	9.1	22.4
TGSLM-Large-distill	S $\rightarrow$ (T, S)	1.9B	135k h	16.8	37.9
External pretrained models					
With text supervision					
SpiritLM Base (Nguyen et al., 2025)	S $\rightarrow$ S	7B	570k h	8.6	23.5
	S $\rightarrow$ T			15.1	31.8
MiMo-Audio-7B-Base (Xiaomi LLM-Core Team, 2025)	S $\rightarrow$ (T, S)	7B	>100M h	26.0	49.5
Without text supervision					
SLAM-Large (Maimon et al., 2025a)	S $\rightarrow$ S	1.3B	81.0k h	1.2	5.5
Llama-Mimi-1.3B (Sugiura et al., 2025)	S $\rightarrow$ S	1.3B	240k h	0.2	0.6
Llama-Mimi-8B (Sugiura et al., 2025)	S $\rightarrow$ S	8B	240k h	0.9	2.9

ing exposure to text during speech adaptation mitigates part of the loss of pretrained knowledge. Distillation provides an additional gain beyond replay, with performance generally improving as the distillation weight grows. A similar trend is observed on StoryCloze (right).

Importantly, these improvements are not limited to speech-to-speech generation. Distillation also improves text-to-text performance, even though the distillation loss during speech training is applied to text predictions conditioned on speech context. This suggests that distillation is not merely teaching the model to map speech contexts to better next-token predictions. It also constrains speech adaptation in a way that better preserves the capabilities of the original text backbone. A similar effect is observed in SALAD (Cuervo et al., 2026), where cross-modal distillation reduces the degradation of the original text capabilities after speech adaptation.

Table 3 compares the scaled TGSLM with distillation and text replay against external models. Scaling up the distilled model improves Known-1000 Pass@5 from 22.4% to 37.9%. Despite using a smaller language model and less speech data, TGSLM-Large-distill exceeds SpiritLM Base on both speech-to-speech and speech-to-text results.[7] Together with the controlled ablation in Fig. 2, this suggests that distillation is an efficient way to retain factual knowledge in speech-to-speech models.

The contrast is larger for pretrained models without text supervision. SLAM-Large and the Llama-Mimi models obtain Pass@5 below 6%, and increasing the size of Llama-Mimi from 1.3B to 8B parameters yields only a small gain (0.6% to 2.9%). This suggests that scaling pretraining without

[7]This comparison is not controlled for data distribution, architecture, base LM, or tokenizers.

Table 4: Gemini-Emo three-way emotion classification accuracy (%, higher is better) on continuation using emotion2vec-plus-large (Ma et al., 2024). We report 95% percentile bootstrap confidence intervals from 1000 resamples of the 96 triplets. $\pm h$ denotes the half-width interval. Generation failures and empty continuations are counted as incorrect.

Model	Happy	Sad	Angry	Macro
Original prefix only	64.6 $\pm$9.4	92.7 $\pm$4.7	77.1 $\pm$8.3	78.1 $\pm$4.3
Our models				
Speech-only	54.2 $\pm$5.2	76.9 $\pm$4.3	36.7 $\pm$4.5	55.9 $\pm$2.7
TGSLM ($\lambda = 0$)	57.7 $\pm$4.9	72.1 $\pm$4.6	43.3 $\pm$5.0	57.7 $\pm$2.8
TGSLM-distill ($\lambda = 0.5$)	60.0 $\pm$4.4	77.3 $\pm$4.3	34.8 $\pm$4.7	57.4 $\pm$2.6
TGSLM-Large-distill ($\lambda = 0.5$)	56.0 $\pm$4.6	75.4 $\pm$4.3	43.5 $\pm$5.4	58.3 $\pm$2.6
External models with limited acoustic information				
SLAM-Large (Maimon et al., 2025a)	65.4 $\pm$4.2	34.4 $\pm$4.6	1.9 $\pm$1.1	33.9 $\pm$1.9
SpiritLM Base (Nguyen et al., 2025)	75.4 $\pm$4.4	23.8 $\pm$4.3	2.7 $\pm$1.6	34.0 $\pm$1.8
SpiritLM Expressive (Nguyen et al., 2025)	66.0 $\pm$4.9	76.7 $\pm$4.0	9.2 $\pm$2.9	50.6 $\pm$2.2
External models with richer acoustic information				
MiMo-Audio Base (Xiaomi LLM-Core Team, 2025)	73.1 $\pm$4.2	45.8 $\pm$5.2	25.6 $\pm$4.6	48.2 $\pm$2.5
Llama-Mimi 1.3B (Sugiura et al., 2025)	46.5 $\pm$4.9	87.7 $\pm$4.0	45.0 $\pm$4.8	59.7 $\pm$2.3
Llama-Mimi 8B (Sugiura et al., 2025)	46.0 $\pm$4.8	88.5 $\pm$3.2	46.5 $\pm$5.0	60.3 $\pm$2.3

text supervision is a less efficient way to retain factual knowledge than text-guided generation with distillation.

5.3 Distillation has limited impact on emotion sensitivity

A natural question is whether distillation reduces sensitivity to paralinguistic information as the model is trained to mimic the behavior of a text LM. We evaluate this using Gemini-Emo together with an emotion classifier, with results shown in Table 4. TGSLM and TGSLM-distill achieve similar emotion classification accuracy, showing limited degradation in emotion sensitivity from distillation in this evaluation. TGSLM-Large-distill performs slightly worse than Llama-Mimi, but remains broadly comparable.

Models based on more limited acoustic representations (**External models with limited acoustic information** row) show substantially more class-imbalanced emotion predictions, which we attribute to the loss of acoustic information during tokenization, as these models use either only semantic tokens or semantic tokens plus pitch tokens.

We view this as a preliminary study of paralinguistic sensitivity rather than a comprehensive evaluation of the generated continuation, as Gemini-Emo uses a single emotion classifier and primarily probes whether acoustic emotion cues in the prompt are reflected in the generated speech. It does not test whether distillation toward a text-only teacher shifts the lexical content toward neutral content.

6 Conclusion

We introduce TGSLM, a speech-to-speech LM that uses internally predicted text to guide speech generation while maintaining speech-only input and output at inference. The TGSLM internal text construction substantially improves generative perplexity and provides an interface for text-LM distillation, which improves factual knowledge retention in both speech-to-speech and text-to-text evaluation. An emotion sensitivity evaluation suggests that distillation has limited impact on this aspect of generation.

TGSLM currently relies on word-level speech-text alignments to construct the training sequences. We have not studied whether this alignment requirement can be eliminated. One possible direction is to predict text and speech at the sentence level, which could provide text guidance without fine-grained alignment. We also focus on pretraining in this work and leave post-training, such as instruction tuning or dialogue adaptation, for future work.

AI USE STATEMENT

In this work, we used AI tools to generate synthetic evaluation data, assist with implementation, and with experimental code. In particular, we used AI to construct synthetic text and speech data as described in the paper, and coding agents to assist with implementation and debugging. We also used AI tools to help draft and edit the paper for clarity and readability, as well as for brainstorming and code-related assistance. The authors reviewed and tested AI-assisted code, and verified all experimental results. The authors reviewed and revised the AI-assisted manuscript text. We take responsibility for the final content of this work.

REFERENCES

A. Amini, A. Banaszak, H. Benoit, A. Böök, T. Dakhran, S. Duong, A. Eng, F. Fernandes, M. Härkönen, A. Harrington, et al. LFM2 technical report. *arXiv preprint arXiv:2511.23404*, 2025.

K. Baba, W. Nakata, Y. Saito, and H. Saruwatari. The T05 system for the VoiceMOS challenge 2024: transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *Proc. SLT*, 2024.

J.-C. Chou, C.-M. Chien, W.-N. Hsu, K. Livescu, A. Babu, A. Conneau, A. Baevski, and M. Auli. Toward joint language modeling for speech units and text. In *Findings of EMNLP*, 2023.

J.-C. Chou, J. Zhou, and K. Livescu. Flow-SLM: joint learning of linguistic and acoustic information for spoken language modeling. In *Proc. ASRU*, 2025.

J.-C. Chou, J. Zhou, and K. Livescu. Quantifying the generation modality gap in speech-text language models. *arXiv preprint arXiv:2609.14743*, 2026.

S. Cuervo and R. Marxer. Scaling properties of speech language models. In *Proc. EMNLP*, 2024.

S. Cuervo, S. Seto, M. de Seyssel, R. Bai, Z. Gu, T. Likhomanenko, N. Jaitly, and Z. Aldeneh. Closing the gap between text and speech understanding in LLMs. In *Proc. ICLR*, 2026.

A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer. 8-bit optimizers via block-wise quantization. In *Proc. ICLR*, 2022.

M. Hassid, T. Remez, T. A. Nguyen, I. Gat, A. Conneau, F. Kreuk, J. Copet, A. Defossez, G. Synnaeve, E. Dupoux, et al. Textually pretrained speech language models. In *Advances in Neural Information Processing Systems*, 2023.

H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi, et al. Emilia: an extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *Proc. SLT*, 2024.

W. Held, Y. Zhang, M. Li, W. Shi, M. J. Ryan, and D. Yang. Distilling an end-to-end voice assistant without instruction training data. In *Proc. ACL*, 2025.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. In *Proc. ICLR*, 2020.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

J. Hwang, M. Hira, C. Chen, X. Zhang, Z. Ni, G. Sun, P. Ma, R. Huang, V. Pratap, Y. Zhang, et al. TorchAudio 2.1: advancing speech recognition, self-supervised learning, and audio processing components for PyTorch. In *Proc. ASRU*, 2023.

K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, et al. Voicebox: text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, 2023.

T. Li and K. He. Back to basics: let denoising generative models denoise. In *Proc. CVPR*, 2026.

G.-T. Lin, P. G. Shivakumar, A. Gourav, Y. Gu, A. Gandhe, H. y. Lee, and I. Bulyko. Align-SLM: textless spoken language models with reinforcement learning from AI feedback. In *Proc. ACL*, 2025.

Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *Proc. ICLR*, 2023.

Y.-J. Lu, Y. Gaur, W. Zhou, B. Muller, J. Villalba, N. Dehak, L. Zettlemoyer, G. Ghosh, M. Lewis, S. Iyer, et al. Latent speech-text transformer. In *Proc. ICLR*, 2026.

Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen. emotion2vec: self-supervised pre-training for speech emotion representation. In *Findings of ACL*, 2024.

G. Maimon, A. Elmakies, and Y. Adi. SLAMming: training a speech language model on one GPU in a day. In *Findings of ACL*, 2025a.

G. Maimon, M. Hassid, A. Roth, and Y. Adi. Scaling analysis of interleaved speech-text language models. In *Proc. COLM*, 2025b.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

E. Nachmani, A. Levkovitch, R. Hirsch, J. Salazar, C. Asawaroengchai, S. Mariooryad, E. Rivlin, R. Skerry-Ryan, and M. Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered LLM. In *Proc. ICLR*, 2024.

T. A. Nguyen, B. Muller, B. Yu, M. R. Costa-Jussa, M. Elbayad, S. Popuri, C. Ropers, P.-A. Duquenne, R. Algayres, R. Mavlyutov, et al. SpiRit-LM: interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025.

OLMo Team. OLMo 3. *arXiv preprint arXiv:2512.13961*, 2025.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

G. Penedo, H. Kydlíček, L. Ben Allal, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, and T. Wolf. The FineWeb datasets: decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, 2024.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023.

S. Rouard, M. Orsini, A. Roebel, N. Zeghidour, and A. Défossez. Continuous audio language models. In *Proc. ICLR*, 2026.

I. Sugiura, S. Kurita, Y. Oda, and R. Higashinaka. Llama-Mimi: exploring the limits of flattened speech language modeling. *arXiv preprint arXiv:2509.14882*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

C. Wang, M. Liao, Z. Huang, and J. Zhang. BLSP-KD: bootstrapping language-speech pre-training via knowledge distillation. *arXiv preprint arXiv:2405.19041*, 2024.

S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston. Neural text generation with unlikelihood training. In *Proc. ICLR*, 2020.

Xiaomi LLM-Core Team. MiMo-Audio: audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Y. Yao, X. Li, X. Jiang, X. Fang, N. Yu, W. Ma, A. Sun, and Y. Wang. FLM-Audio: natural monologues improves native full-duplex chatbots via dual training. *arXiv preprint arXiv:2509.02521*, 2025.

S.-L. Yeh, Y. Meng, and H. Tang. Whisper has an internal word aligner. In *Proc. ASRU*, 2025.

N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi. SoundStream: an end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507, 2022.

S. Zhang, Z. Chen, H. He, L. Du, and Z. Lan. Low perplexity is repetition: a one-dimensional self-conditioning attractor in continuous diffusion LMs. *arXiv preprint arXiv:2607.00588*, 2026.

A APPENDIX

A.1

A.2 ALIGNMENT FILTERING RULES

First, we require the final aligned word to reach at least 80% of the audio duration. Second, we require that no more than 5% of aligned words have effectively zero duration. We exclude utterances whose transcripts exceed 7.5 Qwen3-0.6B text tokens per second.

A.3 LIKELIHOOD MEASURE

For likelihood-based evaluation, we consume the given text and speech continuation using the same alternating procedure and accumulate likelihood only at text-token positions. Because the text-speech alignment is reconstructed online, the strict top-1 boundary criterion can occasionally fail to trigger enough speech boundaries, leaving some text tokens unscored. Only in this case, we use a fallback boundary criterion and accept `<s-eow>` whenever it appears among the top-3 predictions of the first discrete speech-token head. This fallback is used only for likelihood-based scoring (StoryCloze) and is not applied during generation.

A.4 VAD PROTOCOL FOR KNOWN-1000

We observe that abruptly ending a synthesized waveform can introduce tokenizer artifacts in the final speech frames, creating an artificial discontinuity at the end of the input. To reduce this effect, we first trim trailing silence using voice activity detection (VAD) (Hwang et al., 2023) and then pad each waveform so that its duration is a multiple of the tokenizer frame duration. We use a frame duration of 80 ms for Mimi and 40 ms for HuBERT-based (Hsu et al., 2021) representations (Maimon et al., 2025a; Nguyen et al., 2025).

A.5 GEMINI-EMO PROMPTS

For Gemini-Emo, we first use Gemini-3.7-Flash to generate triplets of sentences expressing happy, sad, and angry emotions while sharing the same linguistically neutral prefix. Table 5 shows an example. We construct the shared prefix so that its text alone does not indicate the target emotion. We then use Gemini-3.1-Flash-TTS-Preview to synthesize each sentence with the corresponding emotional prosody. Detailed prompts are provided in Table 6. We obtain word-level speech-text alignments using the same Whisper-based alignment method (Yeh et al., 2025) and extract the prefix as the speech prompt for continuation generation. We classify the full synthesized utterances with emotion2vec-plus-large (Ma et al., 2024), restricting the prediction to the happy, sad, and angry classes, and retain only triplets in which all three utterances are classified as their intended emotions. This results in 96 triplets.

Table 5: An example from Gemini-Emo (`item_01`). Each sentence combines the shared prefix with the corresponding suffix and is synthesized in the indicated emotion. Evaluation uses only the cropped spoken prefix, whose words are identical across emotions.

Shared prefix	I noticed that the bread in the kitchen
Emotion	**suffix**
Happy	has baked into the most wonderful golden loaf!
Sad	has turned stale and completely moldy already today.
Angry	was left uncovered on the counter again!

A.6 TRAINING HYPERPARAMETERS

To reduce storage requirements, we pre-encode the Emilia-YODAS audio into Mimi RVQ tokens using 12 codebooks and decode them back to waveform on the fly during training. Table 7 provides detailed training and decoding hyperparameters.

Table 6: Gemini-Emo language-model and TTS prompt specification. The shared TTS row applies to all three emotion prompts. Each prompt inserts one emotion-specific director instruction between the scene and transcript sections.

Prompt purpose	Model	Prompt template
Text generation	`gemini-3.7-flash`	Create exactly 25 English speech-continuation stimulus items. For every item: write one shared prefix of 7–12 words as a natural, affect-neutral, syntactically unfinished conversational turn; write emotion-specific 5–14-word suffixes that complete that prefix; form each carrier as `PREFIX + `` '' + SUFFIX`; use varied, ordinary ASR-friendly language; and return the supplied item schema. The generation ran in batches through 500 items.
Shared TTS interface	`gemini-3.1-flash-tts-preview` voice: `Schedar`	Synthesize speech for the transcript below. **AUDIO PROFILE:** Use one consistent adult speaker identity and a standard American English accent. **SCENE:** This is a natural conversation with one familiar listener in a quiet room. **DIRECTOR'S NOTES:** `<EMOTION-SPECIFIC DIRECTOR'S NOTES>` Speak the transcript exactly once and verbatim. Do not read these directions. Do not add, remove, or replace words. Do not laugh, cry, sigh, gasp, whisper, shout, or add nonverbal sounds. **TRANSCRIPT:** `<TRANSCRIPT>`
Happy director's notes		Convey a small, private happiness, as if sharing good news with someone you trust. Keep an easy moderate pace and soft-to-medium energy. Let a subtle smile warm the tone and add light upward inflections. Sound understated but unmistakably pleased, not breathy, cute, or theatrical. Apply these cues from the first word, including the neutral lead-in. Keep the performance conversational, coherent, and moderately expressive rather than exaggerated.
Sad director's notes		Convey strong genuine sorrow while trying to remain composed. Use a soft voice, slow measured pacing, lowered resonance, and pronounced downward intonation. Sound deeply sad but never cry, sob, or become theatrical. Apply these cues from the first word, including the neutral lead-in. Keep the performance conversational, coherent, and moderately expressive rather than exaggerated.
Angry director's notes		Sound clearly angry but fully controlled. Use a firm tense tone, clipped articulation, strong stress, deliberate pacing, and moderately high energy. Address the listener directly without shouting. Apply these cues from the first word, including the neutral lead-in. Keep the performance conversational, coherent, and moderately expressive rather than exaggerated.

A.7 FULL RESULTS

Full results for generation-based metrics for Emilia test set are in Table 8, and known-1000 results with post-trained models are in Table 9.

Table 7: Training hyperparameters and decoding parameters.

Hyperparameter	**Architecture ablations**	**Distillation ablations**	**TGSLM-Large-distill**
Training			
Backbone	Qwen3-0.6B-Base	Qwen3-0.6B-Base	Qwen3-1.7B-Base
Optimizer updates	40,000; 50,000 for model with text replay	100,000	785,052 (665,000 + annealing)
Audio budget (s/GPU/update)	480	540	240
GPUs	4	4	4
Peak learning rate	5×10^{-4}	5×10^{-4}	3×10^{-4}
Warmup updates	5,000	5,000	5,000
End-of-cosine-decay / peak learning rate ratio	0.1	0.1	0.1
annealing	–	–	Linear annealing from 3×10^{-5} to 0
Bucket bin duration (s)	2	2	2
FineWeb replay probability	0; 0.2 for replay-enabled model	0.2	0.2
Replay tokens/GPU/update	2,048	2,048	2,048
Optimizer	AdamW8bit (Dettmers et al., 2022)		
Adam (β_1, β_2)	$(0.9, 0.98)$		
Adam ϵ	10^{-6}		
Weight decay	0.01		
Learning-rate schedule	Linear warmup followed by cosine decay followed by optional linear annealing		
Training precision	BF16 mixed precision		
Shared decoding settings			
Sampling temperature	0.8 for text, semantic tokens, and flow Gaussian		
Nucleus threshold p (Holtzman et al., 2020)	0.95		
Flow integration	Euler, 32 steps		
CFG (Ho & Salimans, 2022) scale	0.3		
Mimi decoder quantizers	16		
Samples per prompt	5		
Maximum new speech	10 s (Known-1000); 20 s (Emilia/Gemini-Emo)		

Table 8: Generation-based metrics on the Emilia test set: genPPL, bigram auto-BLEU (BLEU), speaker similarity (Spk), and UTMOS-v2 (UTMOS) (Baba et al., 2024); see Sec. 4.3 for details. For the **Whisper prompt transcript** row in text-leading references, the model is conditioned on a transcript of the speech prompt produced by Whisper `small.en`. All rows are evaluated on the same subset: a shared subset of 1,325 prompts (of 1,500) with non-empty continuations from every model, covering 4,957 of 7,500 continuations. **TGSLM-Large-distill** uses $\lambda = 0.5$ with text replay.

Model	Direction	Params	Data	GenPPL ↓	BLEU↓	Spk ↑	UTMOS ↑
Ground truth	–	–	–	43.8	6.0	0.74	2.82
Speech-only baselines (no text guidance)							
Speech-only	S → S	0.8B	21.3k h	151.7	4.3	0.42	2.56
Interleaved	S → S	0.8B	21.3k h	164.1	5.4	0.37	2.36
TGSLM variants							
TGSLM	S → (T, S)	0.8B	21.3k h	58.9	3.9	0.36	2.41
+ self-generated prompt text	(S, T) → (T, S)	0.8B	21.3k h	62.2	3.8	0.38	2.40
+ distillation ($\lambda = 0.25$)	S → (T, S)	0.8B	21.3k h	56.9	4.3	0.36	2.41
+ distillation + text replay	S → (T, S)	0.8B	21.3k h	51.0	4.9	0.38	2.44
TGSLM-Large-distill	S → (T, S)	1.9B	135k h	45.0	5.7	0.45	2.68
Text-leading references							
Oracle prompt text	(T, S) → (T, S)	0.8B	21.3k h	40.7	5.4	0.42	2.50
Whisper prompt transcript	(T, S) → (T, S)	0.8B	21.3k h	57.1	6.5	0.41	2.50
External models							
With limited acoustic information							
SpiritLM Base	S → S	7B	570k h	30.4	13.3	0.05	2.39
SpiritLM Expressive	S → S	7B	570k h	67.3	8.7	0.07	2.45
SLAM-Large	S → S	1.3B	81.0k h	53.8	10.2	0.07	3.46
With richer acoustic information							
(Chou et al., 2026) with oracle prompt text	(T, S) → (T, S)	0.5B	10k h	55.0	6.6	0.35	2.48
(Chou et al., 2026) with Whisper prompt text	(T, S) → (T, S)	0.5B	10k h	72.1	10.7	0.32	2.35
MiMo-Audio-7B-Base	S → (T, S)	7B	>100M h	14.9	20.3	0.40	3.03
Llama-Mimi-1.3B	S → S	1.3B	240k h	88.7	6.2	0.39	2.61
Llama-Mimi-8B	S → S	8B	240k h	53.2	7.0	0.41	2.59

Table 9: Known-1000 evaluation. Pass@K(%) measures whether the answer appears in at least one of K sampled continuations. Direction S $\rightarrow$ (T, S) denotes a speech prompt followed by a text-guided speech continuation. **TGSLM-Large-distill** uses a distillation weight of $\lambda = 0.5$ and text replay. Post-trained models are shown in gray because they are not directly comparable to pretrained models.

Model	Direction	Params	Data	Pass@1	Pass@5
Text-LM reference					
Qwen3-0.6B backbone (Yang et al., 2025)	T $\rightarrow$ T	0.6B	–	46.8	77.9
Qwen3-1.7B backbone (Yang et al., 2025)	T $\rightarrow$ T	1.7B	–	60.8	86.6
Our models					
TGSLM-distill	S $\rightarrow$ (T, S)	0.8B	45k h	9.1	22.4
TGSLM-Large-distill	S $\rightarrow$ (T, S)	1.9B	135k h	16.8	37.9
External pretrained models					
With text supervision					
SpiritLM Base (Nguyen et al., 2025)	S $\rightarrow$ S	7B	570k h	8.6	23.5
	S $\rightarrow$ T			15.1	31.8
MiMo-Audio-7B-Base (Xiaomi LLM-Core Team, 2025)	S $\rightarrow$ (T, S)	7B	>100M h	26.0	49.5
Without text supervision					
SLAM-Large (Maimon et al., 2025a)	S $\rightarrow$ S	1.3B	81.0k h	1.2	5.5
Llama-Mimi-1.3B (Sugiura et al., 2025)	S $\rightarrow$ S	1.3B	240k h	0.2	0.6
Llama-Mimi-8B (Sugiura et al., 2025)	S $\rightarrow$ S	8B	240k h	0.9	2.9
External post-trained models					
Qwen2.5-Omni-3B (Xu et al., 2025)	S $\rightarrow$ (T, S)	3B	$\sim$3.3M h†	32.9	55.1
Qwen2.5-Omni-7B (Xu et al., 2025)	S $\rightarrow$ (T, S)	7B	$\sim$3.3M h†	44.4	65.8
LFM2.5-Audio (Amini et al., 2025)	S $\rightarrow$ (T, S)	1.2B	472k h	35.2	40.3