SCENA: REFERENCE-DRIVEN MULTI-SPEAKER AUDIO SCENE GENERATION FROM IN-THE-WILD PRIORS

Anonymous authors
Paper under double-blind review

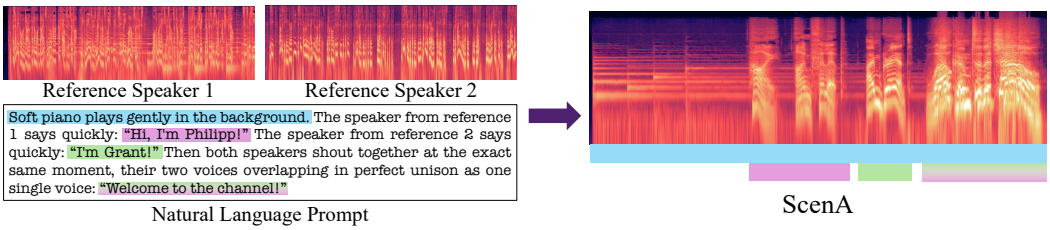

Figure 1: Our SCENA framework transforms free-form natural language prompts and a set of reference voices into rich, multi-speaker conversational scenes. The prompt alone determines which reference speaks where, with no per-turn tags, transcripts, or identity encoders. This natural language interface enables complex human interactions, including overlapping speech, spontaneous paralinguistic events, and scene-level ambient sound.

ABSTRACT

Existing reference-conditioned multi-speaker systems bind speakers to utterances through structured supervision: per-turn tags, multi-stream transcriptions, or learnable speaker embeddings. These systems operate within speech-only pipelines that produce clean vocal sequences without the ambient texture of real conversations. Our method, SCENA, conditions a text-to-audio flow-matching foundation model, pretrained on large-scale in-the-wild data, directly on multiple reference voices and a free-form natural language prompt that describes an entire multi-speaker audio scene. Leveraging such a foundational model allows us to inherit its capacity for natural, non-studio audio: background noise, room acoustics, overlapping dialogue, and spontaneous paralinguistic events, while adding multi-speaker control without any per-turn structure. Concretely, reference latents are concatenated into the model's token sequence and distinguished by lightweight learned reference-slot embeddings. However, we identify a critical obstacle to this approach: the *Reference Shortcut*. During training under standard noise schedules, the model can identify the matching reference by acoustic similarity to the noisy target, bypassing the text prompt entirely. We address this with a high-noise-biased timestep distribution that forces the model to rely on the text prompt for speaker assignment. We first evaluate broad scene-generation capabilities, demonstrating control over vocalizations, discrete events, emotional delivery, ambience, and simultaneous speech. Additionally, we evaluate SCENA on the CoVoMix2-Dialogue benchmark, where it outperforms specialized speech-only multi-speaker systems. Our results demonstrate the advantage of using a general-purpose audio model conditioned on a free-form scene description, rather than passing structured dialog scripts through a speech-only pipeline.

1 INTRODUCTION

A real multi-speaker audio scene is more than ordered turns of speech: voices overlap, laughter cuts in, and the room itself shapes how it all sounds. However, existing speech generation systems capture little of this. Zero-shot voice cloning models address only the single-speaker case (Wang et al., 2023; Du et al., 2024; Chen et al., 2024b; Casanova et al., 2024), leaving multi-speaker conversations to be

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.

assembled segment-by-segment outside the model. Recent dialogue-TTS systems generate multi-turn conversations directly (Zhang et al., 2026; Peng et al., 2025b; Zhu et al., 2025a; Zhang et al., 2025b), but bind speakers to utterances through *structured* supervision (per-turn speaker tags, multi-stream transcriptions, or learnable speaker-turn embeddings). All are speech-only pipelines, producing clean vocal tracks stripped of the ambient texture real conversations carry.

We take a different starting point with SCENA, our flow-matching framework for multi-speaker audio scene generation. Flow-matching (Lipman et al., 2023; Liu et al., 2023b) text-to-audio foundation models (HaCohen et al., 2026), pretrained on large-scale in-the-wild audio, already capture the texture of natural sound scenes. What they lack is a way to tie specific voices to specific roles. SCENA adds this capability with a deliberately minimal interface: reference latents are concatenated with the input latents and distinguished by lightweight learned reference-slot embeddings. A single free-form natural language prompt describes the entire scene: who speaks, what is said, and what else is in the room. No per-turn tags, no multi-stream transcripts, and no identity encoders or reference-side adapters (Ye et al., 2023; Wang et al., 2024b); the prompt alone determines which reference speaks where (Figure 1).

Our experiments show that naïvely training this design fails to learn which reference speaks where, due to a previously unrecognized failure mode of reference-conditioned flow matching that we call the *reference shortcut*. We find that under the standard logit-normal timestep distribution (Esser et al., 2024), the noised target retains enough acoustic information for the model to pick the matching reference by similarity. This shortcut sidesteps the text prompt entirely, and yields low training loss but catastrophic inference. At test time, generation begins from pure noise, where the shortcut is unavailable and text is the only signal that can resolve which voice goes where. Text, however, is precisely the signal the model has learned to ignore. To locate the noise levels where the shortcut works, we train a small probe on frozen audio features and find that it can match references to targets by similarity across the entire low-to-moderate noise range, exactly where standard training concentrates. We close this shortcut by placing sufficient training mass at high noise, where the target is uninformative and text is the only binding signal. We use a Beta+Uniform mixture as one practical realization, while our noise-schedule ablation shows more generally that binding-aware metrics improve monotonically as we shift training mass toward higher noise.

We evaluate SCENA in both its broader scene-generation setting and the conventional two-speaker dialogue setting. The scene evaluation measures control over vocalizations, discrete events, emotional delivery, ambience, and simultaneous speech generated jointly with the referenced voices. We also restrict the same interface to standard two-speaker transcripts and evaluate on the public COVOMIX2-DIALOGUE benchmark. Even in this specialized setting, SCENA outperforms dialog systems on every speaker-binding metric as well as transcription accuracy and overall speaker similarity. Section 4.4 reports the scene controls, and the supplementary material provides corresponding audio examples. To support independent verification, we provide the trained SCENA checkpoint, inference code, and the evaluation implementation used for the reported metrics.

2 RELATED WORK

Multi-speaker speech generation. Zero-shot voice cloning has converged on two architectural families. Autoregressive zero-shot TTS systems (VALL-E (Wang et al., 2023; Chen et al., 2024a), Seed-TTS (Anastassiou et al., 2024), Spark-TTS (Wang et al., 2025), CosyVoice 2/3 (Du et al., 2024; 2025), VoiceStar (Peng et al., 2025a), XTTS (Casanova et al., 2024), MiniMax-Speech (Zhang et al., 2025a)) autoregressively generate discrete speech tokens from a single speaker reference clip. Flow-matching and diffusion generators (Voicebox (Le et al., 2023), E2-TTS (Eskimez et al., 2024), F5-TTS (Chen et al., 2024b), NaturalSpeech 3 (Ju et al., 2024), MegaTTS 3 (Jiang et al., 2025), ZipVoice (Zhu et al., 2025b), StyleTTS 2 (Li et al., 2023)) produce mel-spectrograms or latents in a single non-autoregressive pass. Both families condition on a *single* reference; multi-speaker conversations are obtained post hoc by synthesizing each speaker's segments independently and concatenating, an arrangement that is incompatible with overlapping speech, shared acoustic environments, and scene-level descriptions. A more recent wave of dialog-TTS systems generates multi-turn conversations directly (Ju et al., 2025; Zhang et al., 2024; 2025b; Zhu et al., 2025a; Peng et al., 2025b; Xie et al., 2025c;b;a; Zhang et al., 2026; Yu et al., 2025; Nari Labs, 2025), but binds speakers to utterances through *structured* supervision: per-turn speaker tags (e.g., `[S1]`/`[S2]`), multi-stream transcriptions, learnable speaker-turn embeddings, or LLM-generated annotations.

Reference-conditioned generation. Beyond zero-shot TTS, reference conditioning has been studied across audio generation more broadly. Audiobox (Vyas et al., 2023) is closest to our setting: a flow-matching audio model that conditions on a text caption together with a single voice prompt to jointly synthesise speech, sound, and music. MusicGen (Copet et al., 2023) pairs text with a melody reference for music generation; AudioLDM 2 (Liu et al., 2023a) conditions sound generation on text together with an audio prompt. All of these condition on a *single* reference. We are, to our knowledge, the first to address the multi-reference audio setting where natural language alone determines speaker assignment.

Parallel work in the image and video domains is informative for our design choices. IP-Adapter (Ye et al., 2023) introduces decoupled cross-attention dedicated to image references; InstantID (Wang et al., 2024a) and PhotoMaker (Li et al., 2024) use face-specific encoders for identity preservation. Multi-subject methods rely on additional structure beyond text: bounding-box layouts in MS-Diffusion (Wang et al., 2024b), segmentation maps in MuDI (Jang et al., 2024), localized cross-attention at training and image-augmented prompt tokens at inference in FastComposer (Xiao et al., 2023), and region-aware masked guidance for video in MAGREF (Deng et al., 2025). Closer in spirit to our setup, three image-domain methods condition transformers on references in context. OmniGen (Xiao et al., 2024) replaces each <|image_k|> marker in the prompt with that reference image's tokens, producing a single sequence that interleaves text and image references. UNO (Wu et al., 2025) concatenates references in the attention sequence with offset rotary positions, and In-Context LoRA (Huang et al., 2024) stitches references and target into a composite image; both bind references by matching descriptive subject names in the prompt (*the toy*, *the man with blond hair*) to the reference list. This requires references to be separately describable in language, which fails when they share content, as is typical in a multi-speaker dialog.

SCENA binds references through indexed mentions (*reference 1*, *reference 2*) that need not describe their content, and ties them to dynamic spans of a jointly generated multi-speaker output. This setting exposes a flow-matching-specific failure mode, the reference shortcut (§3.2), in which the model can bypass the prompt by matching references against the noised target. We close it by modifying the timestep distribution alone, rather than by adding an explicit binding mechanism.

Timestep distributions and noise schedules. The choice of noise schedule has a large effect on diffusion-model training. EDM (Karras et al., 2022) proposes log-normal sigma sampling for image generation; Stable Diffusion 3 (Esser et al., 2024) introduces the logit-normal timestep distribution for flow matching, with a mild upward shift at higher resolutions; Min-SNR weighting (Hang et al., 2023) reweights the loss by signal-to-noise ratio; and Simple Diffusion (Hoogeboom et al., 2023) shifts schedules toward higher noise for high-resolution images. All of these choices are motivated by *generation quality* or *training stability*.

3 METHOD

We describe our approach in four parts: the reference conditioning mechanism (§3.1), the reference shortcut and its analysis (§3.2), the timestep distribution that addresses it (§3.3), and the multi-reference dataset on which we train (§3.4). Figure 2 provides an overview.

3.1 REFERENCE CONDITIONING VIA LATENT CONCATENATION

Setup. We build on a pretrained audio diffusion transformer (HaCohen et al., 2026) that operates on a sequence of latent tokens. A target audio clip is encoded into a latent sequence $\mathbf{z}_0 \in \mathbb{R}^{N \times D}$ by the backbone's variational autoencoder (VAE), where N is the number of tokens and D is the latent dimension. A text prompt $\mathbf{c}$ is encoded by a frozen text encoder and conditions generation through cross-attention; reference clips are encoded by the same VAE as the target.

Reference injection. Given $K \leq K_{\max}$ reference speaker clips $\{r_1, \ldots, r_K\}$, each is encoded into a latent sequence $\mathbf{r}_k \in \mathbb{R}^{N_k \times D}$. The full input to the transformer is the concatenation:

$$\mathbf{z}_{\text{input}} = [\mathbf{z}_t;\ \mathbf{r}_1;\ \ldots;\ \mathbf{r}_K], \tag{1}$$

where $\mathbf{z}_t$ is the noised target at timestep t. Only the target is noised; the reference latents $\mathbf{r}_k$ are passed clean. All tokens participate in the transformer's self-attention, allowing the model to attend freely between target and reference tokens.

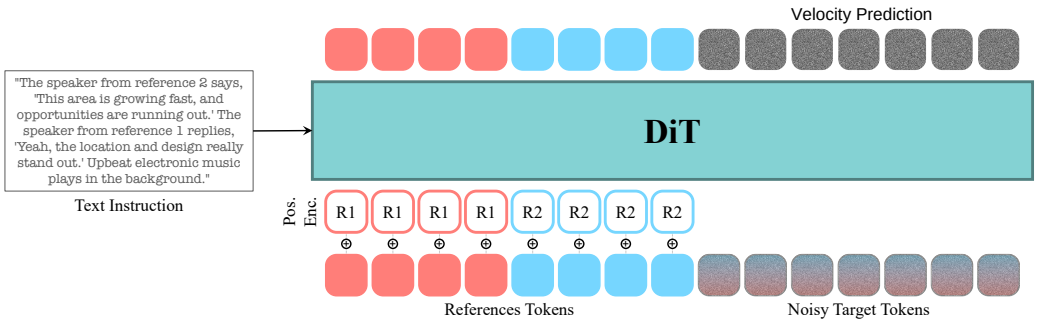

Figure 2: Our framework utilizes a DiT to synthesize reference-driven conversational scenes. Reference latents (red, blue) are concatenated with noisy target tokens, and learned reference-slot embeddings ($R1, R2$) distinguish speakers. Text conditions the model via cross-attention, and training with a flow-matching loss enables transformation into high-fidelity conversational audio.

Reference-slot encoding. To distinguish the references from one another and from the target, we associate each reference slot k with a learned embedding $e_k \in \mathbb{R}^D$, broadcast across the N_k tokens of $\mathbf{r}_k$ and added immediately after the linear projection into the transformer hidden dimension:

$$\mathbf{r}_k \leftarrow \mathbf{r}_k + e_k, \qquad k = 1, \ldots, K_{\max}, \tag{2}$$

while the target $\mathbf{z}_t$ receives no additive embedding. This adds only a negligible number of parameters and leaves the rest of the architecture unchanged. We ablate the choice of slot encoding in Appendix A.9, comparing this additive embedding against a RoPE-based alternative and a no-positional baseline.

Generation through natural language. A single text prompt drives the entire generation. It describes the scene holistically (ambient sounds, speaker turns, content, and affect) and refers to each reference voice by an ordinary textual mention such as *reference 1*. For example:

> *"Ocean waves crash gently on the shore. Seagulls call in the distance. The speaker from reference 1 takes a deep breath and says: 'This is exactly what I needed today.' The speaker from reference 2 hums in agreement: 'Yeah, no emails out here.' Another wave rolls in slowly."*

This prompt, together with the concatenated reference latents (Eq. 1) and their slot embeddings (Eq. 2), is the entire input to the model. No special tokens, no identity-preserving adapters (Ye et al., 2023), no per-segment transcripts, and no spatial or temporal supervision (Wang et al., 2024b; Jang et al., 2024) are used. The model produces the entire scene jointly through a single iterative sampling process, including overlapping speech, natural turn-taking, and scene-level ambient sound. Existing multi-speaker systems typically achieve these qualities only through external structure or post-hoc concatenation. In SCENA, binding is delegated entirely to training, which, as we show next, requires careful design of the timestep distribution.

3.2 THE REFERENCE SHORTCUT CHALLENGE

Flow matching background. In flow matching (Lipman et al., 2023; Liu et al., 2023b), training constructs a noised sample $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\boldsymbol{\epsilon}$ at timestep $t \in [0, 1]$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. The model f_θ is trained to predict the velocity field $\mathbf{v} = \boldsymbol{\epsilon} - \mathbf{z}_0$ that transports noise to data. The training objective is:

$$\mathcal{L} = \mathbb{E}_{t\sim p(t),\, \mathbf{z}_0,\, \boldsymbol{\epsilon}} \left[\| f_\theta(\mathbf{z}_t, t, \mathbf{c}) - \mathbf{v} \|^2 \right], \tag{3}$$

where $p(t)$ is the timestep distribution. Standard practice (Esser et al., 2024) uses a logit-normal distribution $t \sim \sigma(\mathcal{N}(\mu, s^2))$, often with a mild shift of μ toward higher noise. In all such variants, the bulk of training mass remains concentrated on intermediate timesteps, where the denoising task is most informative for generation quality.

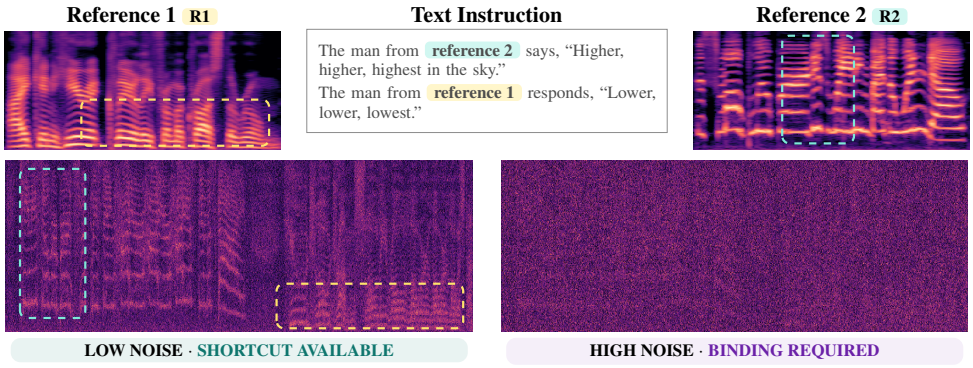

Figure 3: **How noise controls the reference shortcut.** The lower spectrograms show the noised target input to the flow-matching model. At low noise, local target acoustics (dashed boxes) remain matchable to the references, allowing the model to bypass the instruction; high noise obscures these cues and requires text-reference binding. Spectrograms are illustrative, in practice noise is applied to latents.

The shortcut mechanism. We now describe a shortcut available in our setup that, if not addressed, lets the model bypass the text prompt entirely. Binding a reference through the prompt requires the model to compose information across two attention paths. Through cross-attention, it must associate the phrase *"reference k"* in the text with the slot embedding e_k carried by the k-th reference's tokens. Then, through self-attention, the target tokens must locate the reference bearing e_k and route their queries toward it. By contrast, the shortcut requires nothing more than what self-attention already does by default: when the noised target still carries acoustic traces of $\mathbf{z}_0$, its tokens are most similar to the matching reference's, and self-attention routes them there in a single step, with no contribution from the text path. Both routes minimize the training loss; the optimizer favors the simpler one whenever it is available.

Consider a training example with target $\mathbf{z}_0$, references $\{\mathbf{r}_1, \ldots, \mathbf{r}_K\}$, and text prompt $\mathbf{c}$ mentioning "reference k". At timestep t, the noised target is $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\boldsymbol{\epsilon}$, interpolating between the clean target at $t = 0$ and pure noise at $t = 1$. The shortcut is available whenever each speaker's segment of $\mathbf{z}_t$ remains discriminable from non-matching references; if this holds across most of $[0, 1]$, the prompt is rarely needed during training, and the model has little reason to learn the two-path routing. We quantify this with a probe that measures, at each t, whether $\mathbf{z}_t$ still carries enough of $\mathbf{z}_0$ to identify its source reference.

The probe is a binary classifier that mirrors the implicit denoising-time choice: given $\mathbf{z}_t$ and two candidate references (one matching, one from a different sample), predict the match. We use the first 8 of the 48 transformer blocks of our backbone with a small classification head; full setup is in Appendix A.3. The resulting accuracy curve appears in Figure 5. Appendix A.4 further isolates the shortcut in a controlled task where text alone determines the order of the same two references. Accuracy stays at $\geq 98\%$ for $t \leq 0.58$, remains $\geq 90\%$ through $t \approx 0.86$, and is still 75% at $t = 0.96$. Only as t approaches 1 does it collapse to chance. Standard logit-normal distributions and their mildly shifted variants place the majority of training mass squarely in this high-accuracy regime, making the shortcut the easiest strategy for reducing the training loss.

At inference, generation starts from $t = 1$ (pure noise) and iteratively denoises. The initial denoising steps, which establish the global structure of the output (including which speaker speaks where), occur at high t, where the shortcut is unavailable. The model, however, has not learned to use text either. The result is a model that fails to bind references to text, producing outputs where speakers are confused, ignored, or applied inconsistently.

3.3 High-Noise-Biased Timestep Distribution

The probe indicates that closing the shortcut requires sufficient training mass near $t = 1$, where the target no longer reveals which reference matches it. At the same time, we want to retain coverage of lower noise levels rather than concentrate all training in a narrow high-noise range. This motivates

a distribution that emphasizes high noise while still sampling across the full timestep range. We therefore use the following Beta+Uniform mixture:

$$p(t) = (1 - \lambda)\,\mathrm{Beta}(t;\,\alpha, 1) + \lambda\,\mathrm{Uniform}(t;\,\epsilon, 1). \tag{4}$$

The $\mathrm{Beta}(\alpha, 1)$ component, with density $\alpha t^{\alpha-1}$, pushes substantial mass into the high-noise tail where the shortcut is least available; the Uniform component retains coverage across the rest of the range.

3.4 Multi-Reference Dataset

We construct a multi-reference training dataset in which each example consists of a target clip, K reference clips, and a caption describing how the full audio scene of the target is related to the references. Each target is a multi-speaker conversational audio clip containing both dialog and non-speech sounds, while each reference is a distinct *single-speaker* clip. This structure enables supervision in which a single natural-language description specifies how multiple reference speakers are realized within a shared conversational scene. Appendix A.1 provides additional corpus information, including scale and reference composition, caption and scene-content distributions.

To construct this dataset, we employ a multi-stage pipeline comprising the following components:

Matching references to target speakers. For each audio clip, we extract speaker embeddings for each segment using a diarization pipeline. We then use embedding similarity to search for a separate reference clip in which each speaker appears. If a reference clip is found for every speaker, the original clip is defined as a target clip. This process ensures alignment between the speakers in the target clip and its references.

Intermediate target caption. We first caption the target clip independently, without conditioning on the references. This intermediate caption provides a reliable description of the scene and reduces hallucinations compared to directly generating a reference-conditioned caption. It serves as a grounding signal for the final caption generation step.

Multi-reference caption generation. Finally, we generate a caption for the target conversational scene conditioned on the reference clips. The caption describes both the dialog and non-speech sounds, while grounding speaker identities in the references. The captioning prompt is constructed by combining the reference clips, the intermediate target caption, and speaker-level timestamps obtained from the diarization process. The intermediate caption and speaker-aligned timestamps are used only to ground the final scene description produced by the offline captioner. During training and inference, SCENA receives only this final free-form description and the raw reference clips, without timestamps or any other structured supervision. Together, these signals improve speaker attribution and help the captioner produce captions that faithfully describe the full conversational scene.

4 Experiments

Our evaluation addresses two complementary questions: whether natural language controls scene content beyond sequential speech, and whether this broader interface remains strong in the conventional multi-speaker dialog task. We then isolate the training mechanism through controlled ablations.

4.1 Experimental Setup

We adapt the LTX-2.3 audio-video diffusion transformer (HaCohen et al., 2026) by retaining only its audio stream and removing video-to-audio cross-attention. We reuse its VAE, text encoder, and prompt-embedding adapter, and fine-tune the model for 20,000 steps at a global batch size of 128 with up to $K_{\max}$=3 references. The model supports audio durations of up to 20 s. Our default training recipe also uses adversarial reference injection, which fills empty slots with unmentioned distractor references, and slot-shuffle augmentation, which permutes the references and their textual indices. Appendix A.2 provides the full implementation and optimization details; Appendix A.10 ablates both augmentations.

Evaluation set. We evaluate on all 1,000 items from the public CoVoMix2 dialog test set (Zhang et al., 2025b), which pairs DailyDialog (Li et al., 2017) two-speaker transcripts with LibriSpeech

Table 1: Comparison on the full CoVoMix2-Dialogue benchmark.

System	cpWER ↓	cpSIM ↑	ACC ↑	WER ↓	SIM-O ↑	UTMOS ↑	SQUIM ↑
MOSS-TTSD (Zhang et al., 2026)	0.191	<u>0.582</u>	0.888	0.066	0.453	**3.81**	4.30
VibeVoice-7B (Peng et al., 2025b)	0.212	0.575	0.871	0.057	<u>0.460</u>	<u>3.64</u>	4.31
VibeVoice-1.5B (Peng et al., 2025b)	0.207	0.547	0.875	0.057	0.437	3.58	4.32
ZipVoice-Dialog (Zhu et al., 2025a)	<u>0.166</u>	0.581	<u>0.895</u>	<u>0.031</u>	0.453	3.58	**4.34**
Dia (Nari Labs) (Nari Labs, 2025)	0.240	0.364	0.807	0.043	0.303	2.55	4.15
ScenA	**0.139**	**0.628**	**0.916**	**0.025**	**0.471**	3.49	<u>4.33</u>

`test-clean` references (Panayotov et al., 2015). Because LTX-2 is limited to 20 s generations, we truncate dialogs only at complete speaker-turn boundaries, leaving 291 unchanged and trimming 709. Primary comparisons use all 1,000 items; controlled ablations use the unchanged 291-item subset, which we call CoVoMix2-Dialogue-20s. The transcripts are rendered in our standard `reference 1`/`reference 2` format; Appendix A.6 evaluates alternative renderings.

Reference robustness set: CoVoMix2-Dialogue-WildRef. We pair 50 CoVoMix2-Dialogue-20s dialogs with 30 in-the-wild English reference clips spanning crowd noise, music, street ambience, wind, and stylized voices, yielding 100 examples. Holding the dialog content fixed isolates robustness to real-world reference conditions.

Metrics. We report three standard metrics: **WER** using Whisper-large-v3 (Radford et al., 2023), **UTMOS** (Saeki et al., 2022), and **SQUIM** (Kumar et al., 2023). For speaker binding, **cpWER** (Watanabe et al., 2020) is permutation-aligned speaker-aware WER, **cpSIM** is per-speaker WavLM-ECAPA similarity (Chen et al., 2022), and **ACC** is the fraction of aligned words assigned to the prompted speaker. **SIM-O** reports overall WavLM-ECAPA similarity without strict speaker assignment. We use pyannote.audio (Bredin, 2023) for diarization and MMS (Pratap et al., 2024) for forced alignment.

4.2 Comparison with Multi-Speaker Dialog Baselines

We compare ScenA against MOSS-TTSD (Zhang et al., 2026), VibeVoice-1.5B and VibeVoice-7B (Peng et al., 2025b), ZipVoice-Dialog (Zhu et al., 2025a), and Dia (Nari Labs, 2025) on all 1,000 CoVoMix2-Dialogue items. Each baseline is run with its public default settings and receives the same DailyDialog content in its native structured format. As shown in Table 1, ScenA is best on all three binding-aware metrics as well as WER and SIM-O. SQUIM places ScenA within 0.01 of the leader (4.33 versus 4.34), while UTMOS favors MOSS-TTSD. We attribute the lower UTMOS score to the LTX-2.3 backbone, which is trained on in-the-wild video soundtracks rather than studio speech and inherits an acoustic profile that UTMOS scores more conservatively than SQUIM does. This analysis shows that although we train ScenA for scene generation, it also performs well for dialog generation, beating all baselines on every binding-aware metric, transcription accuracy, and similarity to the references, while remaining competitive on the naturalness estimators.

Robustness to in-the-wild references. Table 2 repeats the comparison on CoVoMix2-Dialogue-WildRef, where the studio-clean LibriSpeech prompts are replaced by 30 in-the-wild reference clips (50 dialogs, see §4.1). ScenA retains the best cpSIM, SIM-O, WER, and SQUIM, and remains close to the leader on cpWER and ACC, where MOSS-TTSD edges ahead. On the wild references, all systems degrade. Baseline cpSIM drops by 0.086–0.192, while ScenA remains highest at 0.424. The smaller open-source baselines (VibeVoice, Dia) fall off under the harder reference distribution.

Human evaluation. In a side-by-side A/B test spanning both evaluation sets, ScenA is significantly preferred over every baseline, with non-tie win rates from 59.8% to 84.6%. Appendix A.11 provides the full results and protocol.

4.3 Noise Schedule Ablation

The probe illustrates that the shortcut is available; this experiment tests whether closing it at training time is what actually unlocks binding.

Table 2: Comparison on CoVoMix2-Dialogue-WildRef (50 dialogs paired with 30 in-the-wild reference clips).

System	cpWER ↓	cpSIM ↑	ACC ↑	WER ↓	SIM-O ↑	UTMOS ↑	SQUIM ↑
MOSS-TTSD (Zhang et al., 2026)	**0.156**	0.390	**0.844**	0.059	0.295	**3.45**	4.21
VibeVoice-7B (Peng et al., 2025b)	0.172	0.386	0.841	0.045	0.317	2.56	2.91
VibeVoice-1.5B (Peng et al., 2025b)	0.202	0.365	0.826	0.089	0.293	2.33	2.85
ZipVoice-Dialog (Zhu et al., 2025a)	0.173	0.396	0.825	0.038	0.315	3.20	4.19
Dia (Nari Labs) (Nari Labs, 2025)	0.272	0.278	0.752	0.086	0.256	2.45	3.92
ScenA	0.167	**0.424**	0.819	**0.022**	**0.348**	3.30	**4.28**

We compare our Beta+Uniform sampler (§3.3) against three standard logit-normal schedules and one very-high-noise logit-normal control. The first row (μ=0.17, σ=0.75) is closest to typical audio flow-matching training. The next two progressively shift training mass toward higher noise. The final logit-normal schedule is chosen to match the fraction of training mass that our Beta+Uniform sampler places above t=0.9. Figure 4 shows the density of each timestep distribution and its corresponding metrics. Across the first three logit-normal schedules, all three binding-aware metrics improve as training mass shifts toward higher noise. The matched very-high-noise control achieves the best cpSIM and SIM-O, but is worse than ScenA on cpWER, ACC, and WER. ScenA achieves the best cpWER and ACC, together with the second-best cpSIM and SIM-O.

	Binding			Standard			
System	cpWER ↓	cpSIM ↑	ACC ↑	WER ↓	SIM-O ↑	UTMOS ↑	SQUIM ↑
LogitNormal μ=0.17, σ=0.75	0.167	0.503	0.830	0.019	0.402	3.64	4.29
LogitNormal μ=0.62, σ=0.75	0.158	0.500	0.850	0.019	0.402	**3.65**	4.30
LogitNormal μ=0.77, σ=1.0	0.154	0.549	0.859	**0.018**	0.438	3.54	4.30
LogitNormal μ=1.89, σ=0.75	0.170	**0.576**	0.853	0.028	**0.461**	3.61	**4.34**
ScenA (Beta+Uniform)	**0.145**	0.567	**0.866**	0.020	0.451	3.44	4.32

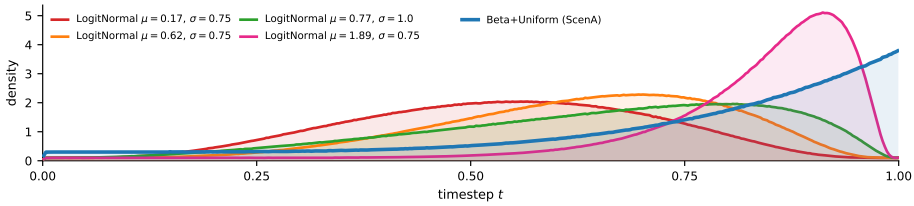

Figure 4: Noise-schedule ablation on CoVoMix2-Dialogue-20s. Top: results, with row tints matching the densities below. Bottom: density of each timestep distribution.

4.4 Measured Scene-Generation Capabilities

We separately evaluate ScenA's ability to generate audio scenes from free-form instructions. These instructions include speaker-specific vocalizations, discrete sound events, emotional delivery, ambience, and simultaneous speech. Existing multi-speaker dialogue systems do not support these controls through their structured speech-only interfaces, so a direct comparison with them is not possible. We instead compare ScenA generations with added scene descriptions against matched generations of the same conversations without the added scene content.

For vocalizations and discrete events, we first generate a plain conversation and then generate the same conversation with an additional instruction, for example, "Then, a dog barks" or "The speaker from reference 1 then laughs." The transcript, references, setting, duration, and random seed are held fixed. We score each clip with the BEATs audio-event classifier (Chen et al., 2023). Table 3 reports the maximum frame-level classifier confidence assigned to the requested sound, averaged across clips. Appendix A.7 provides the full implementation details.

(a) Sound events

Requested sound	Plain	Instructed
Cough	0.001	0.550
Laughter	0.001	0.523
Sigh	0.002	0.318
Doorbell	0.000	0.331
Footsteps	0.013	0.208
Impact	0.058	0.221
Music	0.136	0.322
Wind	0.099	0.187

(b) Emotion

Requested	Plain	Instructed
Angry	0.395	0.665
Happy	0.355	0.459

(c) Overlapping speech

Setting	Overlap
Plain sequential	0.0%
Overlap request	39.6%
Interruption-friendly	70.8%

Table 3: Controlled scene-generation measurements: mean maximum BEATs confidence for sounds, mean RAVDESS-classifier confidence for emotion, and percentage of clips with detected overlapping speech.

Table 3 shows higher classifier confidence for every requested sound. For music and wind, both generations receive the same third reference containing the ambience, but only one prompt asks the model to continue it into the scene.

We use the same paired design to evaluate emotional delivery with a wav2vec2 speech-emotion classifier fine-tuned on RAVDESS. Anger and happiness produce clear increases relative to matched neutral-delivery controls. We also measure whether a speaker-activity detector finds simultaneous speech. An explicit overlap instruction substantially raises the fraction of positive clips, and dialogue written to invite interruptions raises it further, indicating that overlap control works best when the dialogue content supports it. More generally, qualitative inspection suggests that scene instructions are followed more reliably when they fit the surrounding context. Expressive delivery is more convincing for emotionally charged dialogue than for a neutral exchange, just as overlap is more reliable when the dialogue naturally invites interruption.

The requested vocalizations and discrete events largely preserve the speech metrics, while sustained ambience reduces speaker similarity and attribution accuracy. Appendix A.7 reports the full protocols, sample counts, and preservation results. Appendix A.8 presents qualitative spectrogram examples of more complex and unstructured scene behaviors. The accompanying demo page provides a broader audio gallery.

5 Limitations

We inherit two practical limits from the backbone: a 20 s generation cap and $K_{\max} = 3$ reference speakers, beyond which the self-attention sequence grows linearly with K. Both are softer than they appear: in LTX-2, audio tokens are roughly an order of magnitude fewer than video tokens, so an audio-only configuration has substantial headroom, enough to extend duration with modest fine-tuning and to fit additional references (or to add a learned reference-side compressor) without redesigning the model. A third limit is inherent to the flow-matching paradigm: generation duration must be set before sampling, so FM systems resort to either heuristics or user input.

6 Discussion

Contemporary multi-speaker systems typically address binding through structured supervision: per-turn tags, multi-stream transcripts, identity encoders, or spatial layouts. Our results show that these mechanisms are not necessary. A standard flow-matching transformer with concatenated reference latents and a single learned slot vector per index can learn the binding from text, provided training places enough mass in the high-noise regime to defeat the reference shortcut. This simpler interface enables the generation capabilities in §4.4, which cannot be expressed through the structured interfaces of existing systems.

The shortcut argument itself is not specific to audio: it requires only (i) one or more clean references concatenated with a noised target, (ii) the noised target retaining enough information for similarity-based selection at moderate noise, and (iii) training mass concentrated outside the high-noise tail. These conditions are satisfied by any reference-conditioned flow-matching model. We expect the diagnosis and the high-noise-biased fix to transfer to image and video reference conditioning, though direct cross-modality verification is future work.

AI USE STATEMENT

AI assistants supported literature search, code and evaluation-script development, data analysis, manuscript editing, and figure layout. The authors reviewed and verified all implementations, references, experimental results, figures, and text, and remain responsible for the content of the paper.

ETHICS STATEMENT

Reference-conditioned speech generation can be misused for impersonation, fraud, or non-consensual synthetic media. We do not believe our method introduces qualitatively new risks beyond single-speaker voice cloning; standard mitigations (audio watermarking, AI-content disclosures, voice-consent verification, synthetic-speech detection) apply here and should be combined rather than relied on alone.

REPRODUCIBILITY STATEMENT

The anonymous supplementary artifact provides the trained SCENA checkpoint, inference code, baseline configurations, the complete 1,000-item public evaluation set, and the evaluation code for the reported metrics. Sections 3.4 and 4.1 document the corpus-construction and training pipelines, and Appendix A.1 reports the corpus composition, and caption distribution.

Because the underlying in-house audio cannot be redistributed, exact retraining from the original corpus is outside the scope of the release. The checkpoint and public evaluation artifacts nevertheless permit independent regeneration and end-to-end verification of the reported benchmark results.

REFERENCES

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-TTS: A family of high-quality versatile speech generation models, 2024.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, pp. 1983–1987, 2023.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. XTTS: A massively multilingual zero-shot text-to-speech model. In *Proc. INTERSPEECH*, 2024.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193, 2023.

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

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024b.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked guidance for any-reference video generation with subject disentanglement, 2025.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. CosyVoice 2: Scalable streaming speech synthesis with large language models, 2024.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training, 2025.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot TTS. In *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English,

Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model, 2026.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via Min-SNR weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning (ICML)*, 2023.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context LoRA for diffusion transformers, 2024.

Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models, 2024.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Jionghao Bai, Xiaoda Yang, Jialong Zuo, Yu Zhang, Rui Liu, Xiang Yin, and Zhou Zhao. MegaTTS 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*, 2025.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *International Conference on Machine Learning (ICML)*, 2024.

Zeqian Ju, Dongchao Yang, Jianwei Yu, Kai Shen, Yichong Leng, Zhengtao Wang, Xu Tan, Xinyu Zhou, Tao Qin, and Xiangyang Li. MoonCast: High-quality zero-shot podcast generation. *arXiv preprint arXiv:2503.14345*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. TorchAudio-Squim: Reference-less speech quality and intelligibility measures in TorchAudio. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*, 2017.

Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing realistic human photos via stacked ID embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining, 2023a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023b.

Nari Labs. Dia: A TTS model capable of generating ultra-realistic dialogue in one pass. GitHub repository, 2025. Open-weights release. `https://github.com/nari-labs/dia`.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Puyuan Peng, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. VoiceStar: Robust zero-shot autoregressive TTS with duration control and extrapolation, 2025a.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, and Furu Wei. VibeVoice technical report, 2025b.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Proc. INTERSPEECH 2022*, pp. 4521–4525, 2022.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. Audiobox: Unified audio generation with natural language prompts, 2023.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. InstantID: Zero-shot identity-preserving generation in seconds, 2024a.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-Diffusion: Multi-subject zero-shot image personalization with layout guidance, 2024b.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-TTS: An efficient LLM-based text-to-speech model with single-stream decoupled speech tokens, 2025.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, and Neville Ryant. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *Proceedings of the 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation, 2025.

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-free multi-subject image generation with localized attention, 2023.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation, 2024.

Hanke Xie, Dake Guo, Chengyou Wang, Yue Li, Wenjie Tian, Xinfa Zhu, Xinsheng Wang, Xiulin Li, Guanqiong Miao, Bo Liu, and Lei Xie. DialoSpeech: Dual-speaker dialogue generation with LLM and flow matching, 2025a.

Hanke Xie, Haopeng Lin, Wenxiao Cao, Dake Guo, Wenjie Tian, Jun Wu, Hanlin Wen, Ruixuan Shang, Hongmei Liu, Zhiqi Jiang, Yuepeng Jiang, Wenxi Chen, Ruiqi Yan, Jiale Qian, Yichao Yan, Shunshun Yin, Ming Tao, Xie Chen, Lei Xie, and Xinsheng Wang. SoulX-Podcast: Towards realistic long-form podcasts with dialectal and paralinguistic diversity, 2025b.

Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. FireRedTTS-2: Towards long conversational speech generation for podcast and chatbot, 2025c.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023.

Fan Yu, Tao Wang, You Wu, Lin Zhu, Wei Deng, Weisheng Han, Wenchao Wang, Lin Hu, Xiangyu Liang, Xiaodong He, Yankun Huang, Yu Gu, Yuan Liu, Yuxuan Wang, Zhangyu Xiao, Ziteng Wang, Boya Dong, Feng Dang, Jinming Chen, Jingdong Li, Jun Wang, Yechen Jin, Yuan Zhang, Zhengyan Sheng, and Xin Wang. JoyVoice: Long-context conditioning for anthropomorphic multi-speaker conversational synthesis, 2025.

Bowen Zhang, Congchao Guo, Geng Yang, Hang Yu, Haozhe Zhang, Heidi Lei, Jialong Mai, Junjie Yan, Kaiyue Yang, Mingqi Yang, Peikai Huang, Ruiyang Jin, Sitan Jiang, Weihua Cheng, Yawei Li, Yichen Xiao, Yiying Zhou, Yongmao Zhang, Yuan Lu, and Yucen He. MiniMax-Speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder. *arXiv preprint arXiv:2505.07916*, 2025a.

Leying Zhang, Yao Qian, Long Zhou, Shujie Liu, Dongmei Wang, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Lei He, Sheng Zhao, and Michael Zeng. CoVoMix: Advancing zero-shot speech generation for human-like multi-talker conversations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Leying Zhang, Yao Qian, Xiaofei Wang, Manthan Thakker, Dongmei Wang, Jianwei Yu, Haibin Wu, Yuxuan Hu, Jinyu Li, Yanmin Qian, and Sheng Zhao. CoVoMix2: Advancing zero-shot dialogue generation with fully non-autoregressive flow matching, 2025b.

Yuqian Zhang, Donghua Yu, Zhengyuan Lin, Botian Jiang, Mingshu Chen, Yaozhou Jiang, Yiwei Zhao, Yiyang Zhang, Yucheng Yuan, Hanfu Chen, Kexin Huang, Jun Zhan, Cheng Chang, Zhaoye Fei, Shimin Li, Xiaogui Yang, Qinyuan Cheng, and Xipeng Qiu. MOSS-TTSD: Text to spoken dialogue generation, 2026.

Han Zhu, Wei Kang, Liyong Guo, Zengwei Yao, Fangjun Kuang, Weiji Zhuang, Zhaoqing Li, Zhifeng Han, Dong Zhang, Xin Zhang, Xingchen Song, Lingxuan Ye, Long Lin, and Daniel Povey. ZipVoice-Dialog: Non-autoregressive spoken dialogue generation with flow matching, 2025a.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. ZipVoice: Fast and high-quality zero-shot text-to-speech with flow matching. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025b.

A SUPPLEMENTARY

A.1 TRAINING CORPUS STATISTICS AND AVAILABILITY

Table 4: Training-corpus datasheet. Scene-content rates are conservative keyword-based measurements over captions.

Property	Measurement
Number of samples	215,536
Reference count	31.0% with one reference, 66.2% with two, 2.7% with three
Distinct audio	≈1,528 h: 621 h of targets and ≈907 h of unique references
Reference reuse	303,966 distinct clips in 370,053 slots, 85.2% used once
Caption length	mean 70.1 words, median 68, 5th–95th percentile 32–115
Scene content	85.6% ambient/non-speech, 44.0% emotion/delivery, 21.9% paralinguistic, 6.3% explicit overlap

The corpus is derived from an in-house collection of in-the-wild video soundtracks and its audio cannot be redistributed.

A.2 IMPLEMENTATION AND TRAINING DETAILS

Audio-only backbone. LTX-2 is a dual-stream audio-video diffusion transformer with separate audio and video streams coupled by bidirectional cross-modal attention. We use the LTX-2.3 release, retain only the audio stream, and remove all video-to-audio cross-attention layers, yielding a standalone audio-only diffusion transformer. The VAE, text encoder, and prompt-embedding adapter are reused without modification, and references are encoded by the same VAE as the target.

Optimization. We fine-tune the backbone to consume up to $K_{\max}=3$ speaker references per training example. Reference clips can be of any length up to the model's 20 s maximum, with only the target noised during training. Training proceeds for 20,000 steps at a global batch size of 128 on $16\times$ NVIDIA GB200 GPUs, taking approximately 24 hours. We optimize with AdamW ($\beta_1=0.9$, $\beta_2=0.95$, $\epsilon=10^{-8}$, weight decay 0.01) at a peak learning rate of 1×10^{-4}, reached via a 1,000-step linear warmup and held constant thereafter. We maintain an exponential moving average (EMA) of the parameters with decay 0.9999. The training objective is rectified-flow velocity prediction (Liu et al., 2023b; Lipman et al., 2023), and evaluations use the EMA weights.

Auxiliary training-time augmentations. Two augmentations further tighten reference-to-text binding. *Adversarial reference injection* fills empty reference slots with extra "distractor" references not mentioned in the prompt. The only way to satisfy the prompt is then to bind by text rather than copy from whatever happens to be in the sequence. This changes only the input and introduces no new loss term. *Slot-shuffle augmentation* reduces positional bias by permuting the references in the self-attention sequence and rewriting the prompt's `reference k` indices accordingly. Our default introduces shuffling as a curriculum: the slot order remains fixed for the first 10,000 steps, after which references are shuffled.

A.3 REFERENCE SHORTCUT PROBE SETUP

The probe is trained on a binary classification task: given the noised target $\mathbf{z}_t$ and two candidate references (one drawn from the same sample as the target, the other from a different sample), predict which of the two matches the target. We take the first 8 of the 48 transformer blocks of our backbone (discarding the rest) and attach a small two-layer MLP classification head on the pooled output. Inputs are routed exactly as in the full model. The target caption enters through the text cross-attention. The noised target $\mathbf{z}_t$ and the two references form the self-attention sequence, with the references shuffled per example. The timestep t is supplied through the backbone's standard conditioning. The head outputs a single logit predicting which of the two reference slots belongs to the target. We train the probe for 10,000 steps at batch size 128, sampling t uniformly in $[0, 1]$, and evaluate accuracy on 256 held-out examples at each of 50 evenly spaced timesteps in $[0, 1]$.

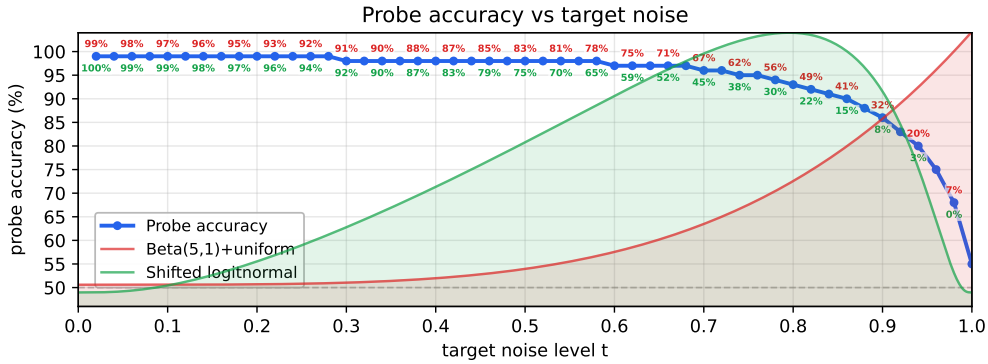

Figure 5: Reference-shortcut probe (§A.3). Blue shows probe binary-classification accuracy on the held-out set, and the gray dashed line marks chance (50%). Light curves show two training-time timestep distributions on the same t axis, our Beta+Uniform mixture (red) and a logit-normal distribution centered around $t = 0.8$ (green), each independently rescaled to a fixed peak height. The small numbers above (red) and below (green) each blue marker give the fraction of training samples drawn at or above that noise level under each distribution, i.e. $\mathbb{P}[T \geq t]$.

A.4 Synthetic-Halves Shortcut Diagnostic

We construct a controlled ordering task from two stationary harmonic references, r_A and r_B. Every target contains both references in two equal halves, while the caption alone selects their order: *In-Order* requests $[r_A; r_B]$ and *Flipped* requests $[r_B; r_A]$ (Figure 6). Because the acoustic content is identical across the two targets, correct generation requires binding the caption to the references rather than merely copying both references.

With the standard logit-normal timestep distribution, validation loss approaches zero while ordering accuracy remains at chance (Figure 7). Under high-noise Beta sampling, accuracy instead rises above 95% very early in training. For a fixed *Flipped* input, changing only the initial noise can reverse the order or introduce several switches, whereas training with a high-noise Beta distribution consistently produces B then A across the four shown draws (Figure 8). This separation between reconstruction loss and prompt adherence is the reference shortcut in isolation: at informative noise levels, the model can reconstruct both components by matching target content to the references without learning their text-specified order; emphasizing high noise removes that matching cue and forces text-reference binding.

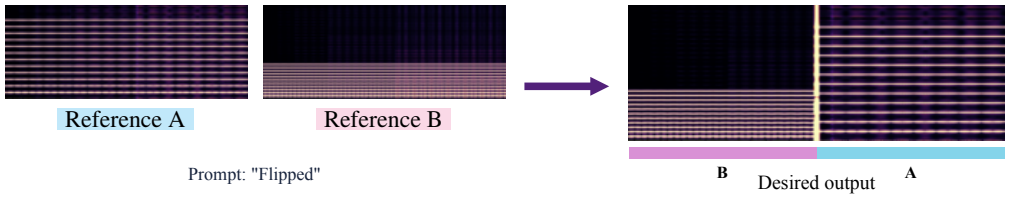

Figure 6: Synthetic-halves task. Both targets contain the same two references; the *Flipped* prompt requires reference B in the first half and reference A in the second.

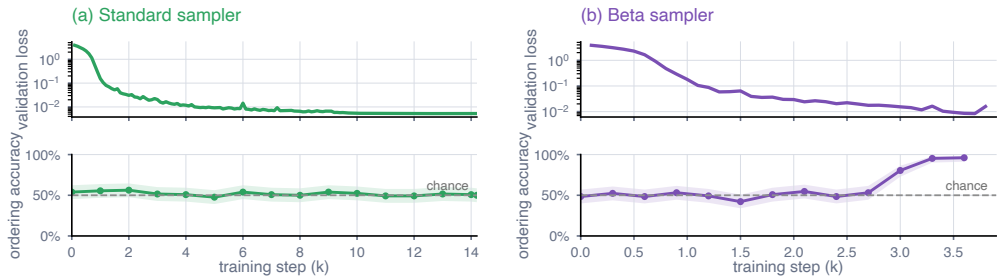

Figure 7: Validation loss and ordering accuracy under the standard logit-normal and high-noise Beta timestep distributions. With standard sampling, loss approaches zero while accuracy remains near chance. With Beta sampling, accuracy rises above 95% in just 3.5k steps. Each checkpoint is evaluated on n=128 samples, with shaded regions showing 95% confidence intervals.

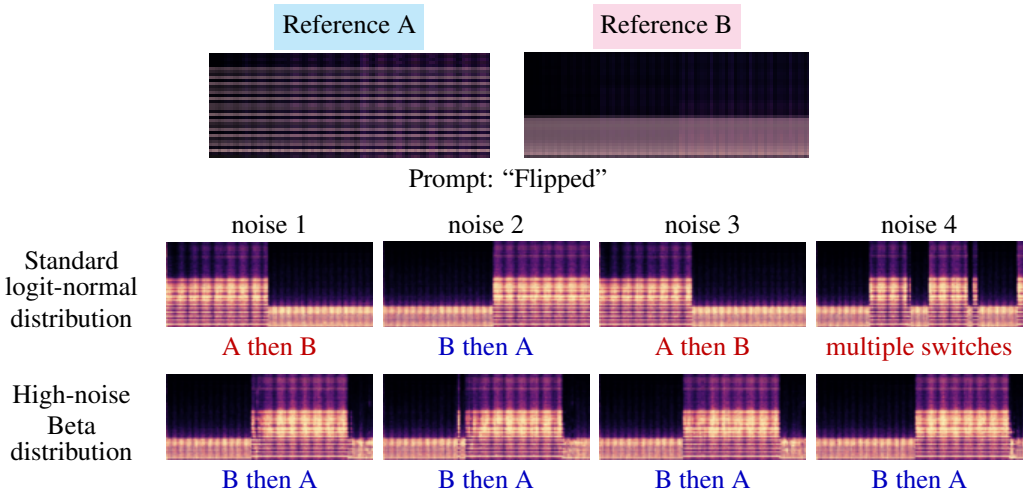

Figure 8: Four generations for one fixed *Flipped* input, varying only the initial noise. Standard logit-normal outputs reverse the requested order or switch repeatedly, whereas the high-noise Beta distribution consistently produces B then A.

A.5 Uniform-Mixture-Weight Ablation

We sweep the Uniform-component weight λ while holding the data, architecture, and remaining optimization settings fixed. Table 5 shows that attribution is stable across the tested range: ACC varies by 0.015 and cpWER by 0.035. On this ablation, WER rises from 0.020 to 0.030–0.032 at the lower weights. We use λ=0.3, as it strikes a balance between speaker similarity and binding metrics, and WER.

Table 5: Uniform-weight sweep for Beta$(5, 1)$+Uniform on CoVoMix2-Dialogue-20s

λ	cpWER ↓	cpSIM ↑	ACC ↑	WER ↓	SIM-O ↑
0.0	0.171	0.585	0.856	0.030	0.476
0.1	0.180	**0.590**	0.851	0.032	**0.480**
0.3	**0.145**	0.567	**0.866**	**0.020**	0.451

A.6 Prompt-Format Robustness

To test robustness to prompt format, we re-render 128 matched items in the evaluation template and in three alternative forms: descriptive prose ("The speaker from reference 1 says, 'I missed you.',"), direct instructions ("Have the speaker from reference 1 say, 'I missed you.',"), and instructions with

timestamps ("Have the speaker from reference 1 say, 'I missed you,' from 00:03 to 00:07."). Dialog, references, and random seed are held fixed. None of the three rewrites changes any reported metric significantly; cpWER moves by at most 0.013 and speaker similarity by at most 0.011.

A.7 Scene-Capability Evaluation Protocol

Paired scene-content study. Each condition contains 64 matched requested/control pairs from 32 source dialogs and two random seeds. For music and wind, both members of each pair receive the same third reference containing the ambience. A separate probe without this reference found essentially no text-only effect for ambience, but qualitative results indicate that ambience can be generated when it has high correlation with the rest of the scene description and audio references.

Request group	Pairs	ΔWER	ΔcpWER	ΔSIM-O	ΔcpSIM	ΔACC	ΔUTMOS
Vocalizations	192	+0.000	+0.029	−0.008	−0.010	−0.021	−0.228
Discrete events	192	+0.004	+0.017	−0.009	−0.018	−0.009	−0.237
Ambience	128	+0.004	−0.007	−0.095	−0.111	−0.136	−0.061
Emotion	192	−0.003	+0.023	−0.032	−0.036	−0.025	−0.068

Table 6: Matched change in speech metrics after adding a scene request to the same prompt. Positive is better for SIM-O, cpSIM, ACC, and UTMOS, but worse for WER and cpWER. The UTMOS decrease for audible events should be interpreted cautiously because a speech-naturalness estimator can penalize intentionally added non-speech content.

Overlap study. The overlap evaluation contains 1,776 outputs across two studies. The first uses 24 ordinary CoVoMix dialogs composed of alternating, complete speaker turns. For each dialog and two random seeds, we test five prompting strategies for eliciting overlapping speech at five requested output lengths, set to 100%, 90%, 80%, 70%, or 60% of the estimated time needed to speak the full transcript. This produces 1,200 clips, with 48 clips for each prompt and length setting. Requesting a shorter output tests whether pressure to fit the same words into less time induces overlap, while WER reveals whether the model instead omits speech.

The second study tests whether the dialog itself makes overlap more likely. We write 24 dialogs containing natural interruption cues and 24 neutral controls matched in setting, turn count, and word count. For example, one speaker begins, "So the deadline has moved to Thursday, which means that we", and the other cuts in with, "Thursday? No, no, it is Wednesday now." We generate each dialog with three prompt variants, two requested output lengths, and two random seeds, producing 576 clips. A frame-level speaker-activity detector marks a clip as containing overlap when at least two speakers remain active simultaneously for 0.30 s. We also verify the ordering of conditions using overlap derived from speaker-diarization segments. At a milder setting, the interruption-cued dialogs reach 43.8% overlap while changing WER by +0.007 and UTMOS by −0.011 relative to matched full-duration generations without overlap instructions.

A.8 Qualitative Spectrograms

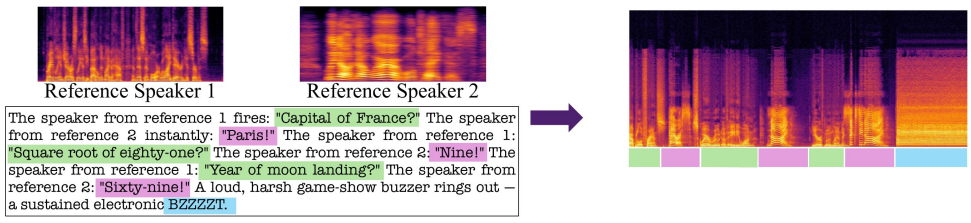

Figure 9: Rapid-fire quiz closed by a non-speech buzzer.

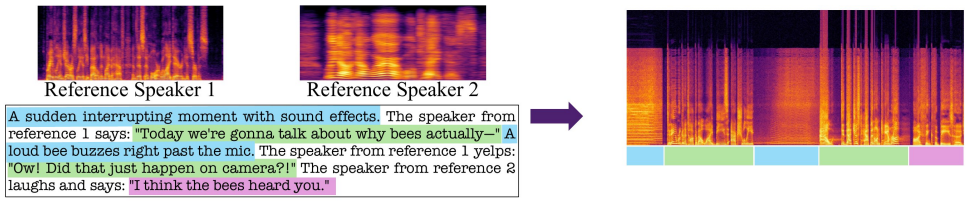

Figure 10: Scene-level prompt with a mid-utterance ambient interruption and paralinguistic events.

A.9 Reference Slot Encoding

We compare three slot-encoding choices. The first is our additive slot embedding (Eq. 2). The second is a RoPE-based alternative that augments the rotary positional encoding (Su et al., 2021) with an additional dimension. The new dimension takes value k for tokens belonging to reference $\mathbf{r}_k$ and value 0 for tokens of the target $\mathbf{z}_t$. The original temporal RoPE continues to index time within each segment, while the new dimension distinguishes the target from each reference:

$$\mathrm{RoPE}(\mathbf{x}) = \mathrm{RoPE}_{\text{temporal}}(\tau) \otimes \mathrm{RoPE}_{\text{ref}}(k), \tag{5}$$

where $\otimes$ denotes concatenation along the per-dimension rotations, τ is the within-segment time of token $\mathbf{x}$, and k is its slot index. The third is a *no-positional* baseline that concatenates references in the self-attention sequence without any additional encoding to distinguish them from the target or from each other.

Adding a new RoPE dimension changes the positional statistics seen by the pretrained backbone. For a fair comparison, we precede RoPE fine-tuning with a 5,000-step warmup on our general-audio corpus, with no references but the extra RoPE dimension already in place. This lets the backbone adapt to the modified positional encoding before reference conditioning is introduced. The additive and no-positional variants require no such adaptation and start fine-tuning directly from the backbone.

Table 7 reports results on CoVoMix2-Dialogue-20s. The additive embedding leads on every binding-aware metric (cpWER, cpSIM, ACC); the RoPE variant follows closely. The no-positional baseline collapses on binding: ACC falls to 0.513 (essentially chance for two speakers) and cpSIM drops by 0.16. WER and naturalness are largely unchanged, since the model can still produce coherent speech when freed from the binding constraint. This indicates that some explicit slot signal is necessary, but binding does not hinge on its precise form.

Table 7: Reference slot encoding ablation on CoVoMix2-Dialogue-20s. The RoPE variant uses an additional 5,000-step warmup on general audio with the extra RoPE dimension in place.

System	cpWER ↓	cpSIM ↑	ACC ↑	WER ↓	SIM-O ↑	UTMOS ↑	SQUIM ↑
SCENA (no-pos)	0.232	0.403	0.513	**0.018**	0.333	**3.60**	4.28
SCENA (RoPE)	0.181	0.547	0.835	0.020	0.449	3.58	**4.32**
SCENA (additive)	**0.145**	**0.567**	**0.866**	0.020	**0.451**	3.44	**4.32**

A.10 Training Recipe Ablations

Table 8 ablates the two auxiliary augmentations described in §4.1 (adversarial reference injection and slot-shuffle augmentation) on CoVoMix2-Dialogue-20s. Removing adversarial references (*no adversarial*) drops cpSIM by ∼0.10 and SIM-O by ∼0.08, with WER and naturalness essentially unchanged, confirming that the augmentation specifically tightens reference-speaker fidelity rather than generation quality. Reference-shuffle behaviour is more nuanced: *always-shuffle* from step 0 collapses (ACC 0.50, near chance), because the model never gets a chance to anchor a stable reference-to-slot mapping before the augmentation starts moving slots around. *No-shuffle* at all is competitive on cpWER and ACC, but our curriculum retains a clear lead on the strict speaker-fidelity metrics (cpSIM, SIM-O, SQUIM) where positional bias would surface in adversarial / wild-reference conditions.

Table 8: Training-recipe ablation on CoVoMix2-Dialogue-20s. *no adversarial*: same recipe as ours but without random distractor references. *always-shuffle*: reference-slot shuffle augmentation from step 0. *no-shuffle*: never shuffle. Our default uses both adversarial distractors and a shuffle curriculum (no shuffle for the first 10 k steps, shuffle thereafter).

System	cpWER ↓	cpSIM ↑	ACC ↑	WER ↓	SIM-O ↑	UTMOS ↑	SQUIM ↑
SCENA (no adversarial)	0.157	0.467	0.859	**0.018**	0.368	3.52	4.29
SCENA (always-shuffle)	0.232	0.402	0.502	0.019	0.334	**3.68**	4.26
SCENA (no-shuffle)	**0.131**	0.491	**0.886**	**0.018**	0.380	**3.68**	4.28
SCENA	0.145	**0.567**	0.866	0.020	**0.451**	3.44	**4.32**

A.11 Subjective Evaluation

We ran a side-by-side A/B preference test against MOSS-TTSD, VibeVoice-7B, ZipVoice-Dialog, and Dia, on items drawn from both CoVoMix2-Dialogue-20s and CoVoMix2-Dialogue-WildRef. SCENA is significantly preferred over every baseline (Table 9).

Table 9: Side-by-side A/B preference results on a mix of CoVoMix2-Dialogue-20s and CoVoMix2-Dialogue-WildRef items. SCENA *preferred* is SCENA's win rate among non-tie ratings. Significance is from a two-sided binomial test (* $p<0.05$, ** $p<0.01$, *** $p<0.001$).

Opponent	SCENA preferred
ZipVoice-Dialog (Zhu et al., 2025a)	84.6%***
Dia (Nari Labs, 2025)	74.2%***
VibeVoice-7B (Peng et al., 2025b)	68.3%**
MOSS-TTSD (Zhang et al., 2026)	59.8%*

Table 10 shows the instructions presented to evaluators for the side-by-side preference test. Each evaluator is given two reference recordings of the original speakers (natural, noisy recordings) followed by two synthesized versions of a dialog between them, denoted (a) and (b). For every comparison the evaluator selects (a), (b), or "About the same". Within a session, our system is paired against one randomly chosen competitor per question, with the (a)/(b) order randomized per item; evaluators are blind to system identity.

The study contains 385 judgments from 53 unique raters over 34 unique dialogue items, with approximately 11 ratings per item. Raters read the transcript while listening to the two original references and the two candidate dialogues. Ties accounted for 9.6% of judgments overall, with per-opponent rates ranging from 5.8% to 11.4%, and were excluded when computing preference rates. We use the same SCENA checkpoint reported in Tables 1 and 2. Inference settings and prompt rendering match the quantitative evaluation, with every baseline run at its public defaults and given the dialogue in its native structured format.

Table 10: Side-by-side A/B Preference Evaluation Instructions.

Instruction
You'll hear recordings of two original speakers (natural recordings with background noise; ignore the noise). Then you'll hear 2 synthesized versions of a dialog between them. Pick the version where the voices sound more like the original speakers and the conversation feels more natural. **Which one is better?** (a): Dialog (a) is better (b): Dialog (b) is better About the same: Cannot tell which is better