PoDAR: Power-Decoupled Audio Representation for Generative Modeling

Anonymous authors
Paper under double-blind review

Abstract

The performance of audio latent diffusion models is primarily governed by generator expressivity and the modelability of the underlying latent space. While recent research has focused primarily on the former, as well as improving the reconstruction fidelity of audio codecs, we demonstrate that latent modelability can be significantly improved through explicit factor decoupling. We present PoDAR (**Po**wer-**D**ecoupled **A**udio **R**epresentation), a framework that utilizes a randomized power augmentation and latent consistency objective to decouple signal power from invariant semantic content. This factorization makes the latent space easier to model, which both accelerates the convergence of downstream generative models and improves final overall performance. When applied to a Stable Audio 1.0 VAE with an F5-TTS generator, PoDAR achieves about a $2\times$ acceleration in convergence to match baseline performance, while increasing final speaker similarity by 0.055 and UTMOS by 0.22 on the LibriSpeech-PC dataset. Furthermore, isolating power into dedicated channels enables the application of CFG exclusively to power-invariant content, effectively extending the stable guidance regime to higher scales.

1 Introduction

Modern high-fidelity image and audio generators often adopt a two-stage latent modeling pipeline where an autoencoder maps high-dimensional signals to a compact continuous representation and a downstream diffusion or flow model learns the distribution of these latents under text or other conditioning. Besides avoiding the cost of modeling raw signals, latent representations provide smoother and more semantically structured manifolds that are easier to model than the pixel or waveform space (Rombach et al., 2022; Liu et al., 2023a; Evans et al., 2025). Within this framework, the autoencoder functions as more than a simple compressor because it establishes the fundamental geometry and statistics that the downstream generator must ultimately navigate.

This design makes the *modelability* of the latent space a central concern, where a space is considered modelable if a downstream generator can learn its distribution efficiently while maintaining fast convergence and high quality under guided conditioning (Rombach et al., 2022). Early autoencoders utilized in latent generative modeling were optimized primarily for reconstruction fidelity and perceptual quality, or compression ratio (Rombach et al., 2022; Esser et al., 2021). However, a representation that achieves high reconstruction accuracy is not necessarily easy to model (Xu et al., 2026).

Methods such as Leng et al. (2025); Zheng et al. (2026); Défossez et al. (2024); Li et al. (2025) report improved performance of downstream generator modelability by using a frozen pre-trained encoder to infuse semantic information into the latent space. However, the primary drawback of these methods is that selecting an appropriate representation is highly non-trivial, often requiring extensive ablations (Yu et al., 2025) over encoder architectures and intermediate feature layers. Encoder quality does not necessarily correlate with downstream generative performance (Singh et al., 2026), evidenced by the fact that alignment with larger variants within the same model family can result in similarly or worse generative performance (Didi, 2025). Additionally, these approaches increase the memory footprint during training.

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.

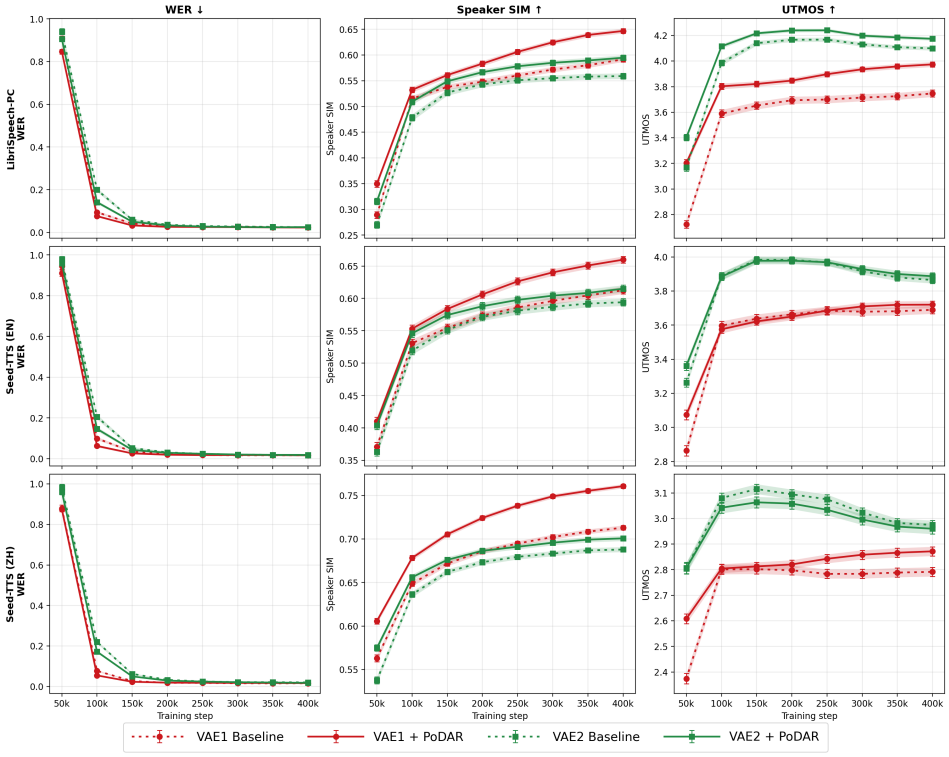


Figure 1: Comparison of WER, speaker SIM, and UTMOS for baseline and PoDAR configurations as a function of training progress for the LibriSpeech and Seed-TTS datasets. We observe that generators trained on codecs with PoDAR consistently outperform their baseline counterparts across speaker similarity (SIM) and speech quality (UTMOS). VAE1 and VAE2 refer to standard codecs from `stable-audio-tools` (Stability AI, 2025), based on DAC (Kumar et al., 2023) and Oobleck respectively.

On the other hand, our experiments suggest that semantic information becomes entangled with independent nuisance variables, such as signal power, which are largely irrelevant to the linguistic content of a speech signal. When such attributes are entangled with semantic content in the latent space, the downstream generator must learn a complex joint distribution over unrelated factors, thereby impairing latent modelability.

Considering this, we take an orthogonal approach to improving the modelability of the latent space, which can be deployed either as a standalone alternative or in conjunction with representation alignment. Power-Decoupled Audio Representation (PoDAR) is a lightweight, representation-agnostic, self-supervised framework that partitions the autoencoder latent space into a power subspace and a power-invariant semantic subspace, without relying on external pre-trained encoders. During the autoencoder training, we apply randomized power perturbations and employ a consistency loss to enforce this decoupling. The primary goal of this framework is not to obtain interpretable factors, but rather to apply these techniques to directly address the modelability bottleneck in two-stage generation by producing a latent distribution that is significantly easier for a downstream diffusion or flow model to learn. Our empirical results show that this representation-level structure improves the final generation quality and speeds up convergence by around $2\times$, as illustrated in Fig. 1.

The same factorization also provides a natural modification for Classifier-Free Guidance (CFG; Ho & Salimans, 2022). Within a conventional entangled space, any attempt to boost conditional adherence may inadvertently amplify nuisance variables, such as signal power, and distort the signal in the process. In a power-decoupled latent space, however, guidance can be applied selectively to the semantic subspace, while leaving the power channels governed strictly by the learned generative dynamics. Improved robustness to high guidance scales thus emerges as an advantageous secondary effect of this method.

Our contributions. The following points summarize the primary technical and empirical contributions of this work:

1. **The PoDAR framework** establishes a lightweight, architecture-agnostic, self-supervised approach for learning power-decoupled audio representations by utilizing a randomized power augmentation and latent consistency objective to partition the latent space into power and power-invariant semantic subspaces without relying on external pre-trained encoders.
2. **Enhanced latent modelability** results from the explicit decoupling of signal power from semantic content, allowing generators to converge about $2\times$ faster compared to conventional entangled baselines, while achieving superior final results.
3. **Partial CFG** allows steering exclusively over the semantic subspace, improving robustness at higher guidance scales.

2 BACKGROUND

Latent generative modeling. Modern high-fidelity generative modeling often utilizes a learned continuous latent representation rather than operating directly on raw pixels or waveforms (Rombach et al., 2022; Liu et al., 2023a; Evans et al., 2025). Let $\boldsymbol{x} \in \mathbb{R}^N$ represent an input signal, such as an image tensor or audio waveform. An autoencoder establishes an encoder-decoder pair through the following transformation:

$$z = E(x), \qquad \widehat{x} = D(z) \tag{1}$$

where L denotes the number of latent dimensions, T indexes spatial or temporal locations and $z \in \mathbb{R}^{L\times T}$ denotes the latent. Following this initial stage, a downstream conditional generator is trained to learn the distribution $p(\boldsymbol{z} \mid \text{cond})$ where cond signifies conditioning information, such as text or class labels. During the generation process, samples are produced in the latent space and subsequently mapped back to the signal domain using the pre-trained decoder $\boldsymbol{D}(\cdot)$.

Diffusion and flow-based models provide robust mechanisms for learning complex distributions $p(\boldsymbol{z} \mid \text{cond})$. Diffusion models generate samples by iteratively denoising an initial noisy variable through learned score, noise or velocity predictions (Ho et al., 2020; Song et al., 2021; Karras et al., 2022), while flow matching and rectified-flow methods instead learn vector fields or ODEs that transport a base distribution to the target data distribution (Lipman et al., 2023; Liu et al., 2023b). Regardless of the specific generative mechanism, the distribution induced by the autoencoder remains the fundamental object that the downstream generator must learn.

Representation alignment. Recent research, particularly in the image domain, has focused on improving modelability of the latent space by aligning it with a pre-trained encoder. REPA (Yu et al., 2025) improves the training of diffusion transformers by aligning internal denoising representations with pretrained visual features, while REPA-E (Leng et al., 2025) extends this concept to the end-to-end tuning of VAE tokenizers. Furthermore, Representation Autoencoders (RAEs; Zheng et al., 2026) replace traditional reconstruction-only VAEs with architectures built around pretrained encoders such as DINO (Caron et al., 2021), SigLIP (Zhai et al., 2023) or MAE (He et al., 2022) to yield semantically rich latents that accelerate the convergence of diffusion models. Although these advancements were primarily developed for the image domain, there has been parallel progress on the audio front. In the development of the Moshi codec (Défossez et al., 2024), the authors align the primary codebook latents with a pre-trained WavLM encoder (Chen et al., 2022) and similarly, the DualCodec framework (Li et al., 2025) aligns its representations with a pre-trained w2v-BERT-2.0 model (Seamless Communication et al., 2023).

Factor decoupling in audio representations. There has been some recent work exploring decomposing neural audio representations to isolate specific acoustic attributes. Shape-gain decomposition (Sadok et al., 2026) has been proposed to improve codec scale-robustness and compression efficiency by manually extracting a scalar gain prior to encoding. Other approaches rely on cross-reconstruction swap losses (Grundhuber et al., 2026; Omran et al., 2023) to separate speech from environmental features, such as reverberation or background noise, thereby enabling applications like acoustic teleportation and speech enhancement. While these methods share an interest in representation decoupling, they are fundamentally optimized for enhancement and compression tasks. In contrast,

the objective of PoDAR is to decouple signal power to eliminate nuisance variables and improve the modelability of the latent space for downstream generative models.

Classifier-free guidance. Classifier-Free Guidance (CFG; Ho & Salimans, 2022) is a fundamental component of modern generative models, since it not only improves conditional adherence, but also generation quality (Karras et al., 2024). It achieves this by extrapolating from an unconditional baseline toward a conditional prediction:

$$\boldsymbol{v}_{\text{cfg}} = \boldsymbol{v}_0 + w(\boldsymbol{v}_{\text{cond}} - \boldsymbol{v}_0), \tag{2}$$

where $\boldsymbol{v}_0$ is the unconditional prediction, $\boldsymbol{v}_{\text{cond}}$ the conditional prediction, and the scalar w controls the strength of the extrapolation towards the conditioning signal.

3 METHOD

Our approach has two components: (i) an augmentation and consistency constraint that encourages power decoupling in latent space, and (ii) a partial CFG rule that applies guidance only to the power-invariant subspace at generation time.

3.1 POWER AUGMENTATION

For an audio waveform $\boldsymbol{x} \in \mathbb{R}^N$, we define a randomized *power augmentation* $\mathcal{A}(\cdot)$

$$\widetilde{\boldsymbol{x}} = \mathcal{A}(\boldsymbol{x}). \tag{3}$$

The augmentation is chosen to vary the power while preserving semantic content, applying a random global gain between -6 and $+6$ dB:

$$\widetilde{\boldsymbol{x}} = \mathcal{A}(\boldsymbol{x}) = g\boldsymbol{x}, \qquad g = 10^{u/20}, \qquad u \sim \text{Uniform}[-6, 6]. \tag{4}$$

We specifically target signal power because it can be perturbed by a purely linear transformation that perfectly preserves the original waveform's temporal alignment. This exact structural correspondence is a strict prerequisite for the consistency objective defined in Section 3.2. Attempting to isolate other acoustic features, such as pitch, voicing, or speaker ID, would typically violate the alignment requirement or introduce artifacts, which would have to be modeled by the network as well.

3.2 CONSISTENCY CONSTRAINT

Let $\boldsymbol{z} = \boldsymbol{E}(\boldsymbol{x}) \in \mathbb{R}^{L \times T}$ be the encoder output. We reserve the first k channels for power and treat the remaining channels as power-invariant content:

$$\boldsymbol{z} = \big[\boldsymbol{z}_p;\ \boldsymbol{z}_c\big], \qquad \boldsymbol{z}_p \in \mathbb{R}^{k \times T}, \quad \boldsymbol{z}_c \in \mathbb{R}^{(L-k) \times T}. \tag{5}$$

During training, we enforce invariance of the content channels under power augmentation by penalizing deviations between content latents of $\boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}$:

$$\mathcal{L}_{\text{PoDAR}} = \big\|\phi(\boldsymbol{E}(\boldsymbol{x}))_c - \phi(\boldsymbol{E}(\widetilde{\boldsymbol{x}}))_c\big\|_2^2, \tag{6}$$

where $(\cdot)_c$ denotes channels $k, k{+}1, \ldots, L{-}1$.

For deterministic autoencoders, we use $\phi(\boldsymbol{E}(\boldsymbol{x})) = \boldsymbol{E}(\boldsymbol{x})$. For VAEs with encoder $q_\psi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_\psi(\boldsymbol{x}), \text{diag}(\boldsymbol{\sigma}_\psi^2(\boldsymbol{x})))$, we apply the constraint only to the mean, i.e., $\phi(\boldsymbol{E}(\boldsymbol{x})) = \boldsymbol{\mu}_\psi(\boldsymbol{x})$, justified by the latent variance analysis provided in Appendix D.

The full autoencoder objective becomes:

$$\mathcal{L}_{\text{AE}}^{\text{new}} \;=\; \mathcal{L}_{\text{AE}}^{\text{old}} \;+\; \lambda_{\text{PoDAR}}\, \mathcal{L}_{\text{PoDAR}}. \tag{7}$$

This additional loss encourages the encoder to route augmentation-induced power variation into $\boldsymbol{z}_p$ while keeping only semantic content in $\boldsymbol{z}_c$. The hyperparameter k controls the capacity allocated to power, and λ_{PoDAR} controls the strength of the decoupling.

3.3 PARTIAL CFG

We propose a modified CFG rule, which applies guidance only to the power-invariant channels:

$$\boldsymbol{v} = \big[\boldsymbol{v}_{\text{cond},p};\ \boldsymbol{v}_{0,c} + w(\boldsymbol{v}_{\text{cond},c} - \boldsymbol{v}_{0,c})\big], \tag{8}$$

where $\boldsymbol{v}_{*,p}$ denotes channels $0, 1, \ldots, k{-}1$, and $\boldsymbol{v}_{*,c}$ denotes channels $k, k{+}1, \ldots, L{-}1$ of the prediction $\boldsymbol{v}_*$. By restricting guidance to the power-invariant subspace, partial CFG improves the robustness of the generator at elevated guidance scales.

4 EXPERIMENTAL SETUP

4.1 BASELINES

Autoencoder baseline. We evaluate two continuous audio representations derived from the *Stable Audio 1.0 VAE* (VAE1) and *Stable Audio 2.0 VAE* (VAE2) autoencoders (Evans et al., 2024; 2025) implemented in `stable-audio-tools` (Stability AI, 2025), re-training both VAEs from scratch under the original published recipe.

Both baselines are waveform autoencoders trained at 44.1 kHz with a 64-dimensional continuous Gaussian VAE bottleneck. Training is consistent across the two, with a perceptually weighted multi-resolution STFT reconstruction loss (Yamamoto et al., 2020), an EnCodec-style multi-scale STFT discriminator (Défossez et al., 2023), the same adversarial and feature-matching weights (0.1 / 5.0), and EMA. Table 5 summarizes the key configuration differences between VAE1 and VAE2.

We use the same training mixture as *DAC* (Kumar et al., 2023), spanning speech: DAPS (Mysore, 2015), DNS Challenge 4 clean speech segments (Dubey et al., 2022), Common Voice (Ardila et al., 2020), and VCTK (Yamagishi et al., 2019); music: MUSDB18 (Rafii et al., 2017) and MTG-Jamendo (Bogdanov et al., 2019); and environmental audio: AudioSet balanced and unbalanced train segments (Gemmeke et al., 2017).

Unless otherwise noted, all representations are trained with a single power channel ($k = 1$) and decoupling weights of $\lambda_{\text{PoDAR}} = 0.5$ for VAE1 and $\lambda_{\text{PoDAR}} = 0.1$ for VAE2, as detailed in Section 5.1.

Generator baseline. We use F5-TTS (Chen et al., 2025) as the text-to-speech generator, adapted to operate on the continuous autoencoder latents introduced in Section 4.1. We follow the published F5-TTS v1 Base recipe and modify the input/output projections to match the autoencoder latent dimension. To improve stability and training efficiency, we apply per-channel z-score normalization, following common latent-standardization practice in latent diffusion models (Rombach et al., 2022). The generator operates on normalized latents $\widetilde{\boldsymbol{z}} = (\boldsymbol{z} - \boldsymbol{\mu})/(\boldsymbol{\sigma} + 10^{-6})$, and we invert this transform prior to decoding.

Generators are trained on the bilingual *Emilia ZH-EN* subset of the Emilia dataset (He et al., 2024), as released for the original F5-TTS recipe. We use the official Emilia train split.

Unless stated otherwise, we use $w = 3.0$ for VAE1-based generators and $w = 2.0$ for VAE2-based generators, which we found to be the best CFG operating point for each family in the full CFG sweep.

More details about the exact architecture and inference strategy can be found in Table 6.

4.2 METRICS

Reconstruction metric. For evaluating the reconstruction quality of codecs, we report ViSQOL (audio mode; Hines et al., 2015), a full-reference intrusive metric based on spectro-temporal similarity, computed on the same held-out DAPS speakers (F10 and M10) used by *DAC* (Kumar et al., 2023).

Measuring modelability. To quantify the modelability of the latent space, we initially explored several direct statistical and geometric measures, including the condition number of the latent covariance matrix, the latent smoothness (described in full in Appendix F), and temporal locality. Whil improvement of about 0.05 for VAE1 across e these metrics effectively captured large improvements in modelability, they lacked the sensitivity necessary to consistently reflect more modest gains. Ultimately, the definition of modelability is the difficulty of learning the underlying score or velocity

fields. Therefore, we evaluate it by measuring the convergence speed and final generation metrics of the downstream flow-matching model.

Generation metrics. We evaluate generated speech following the same protocol mentioned in F5-TTS (Chen et al., 2025), which includes WER and speaker similarity (SIM-o). Additionally, to assess the quality of the generated speech in a reference-free manner, we use UTMOS (Saeki et al., 2022).

Results are reported on three public test sets: the LibriSpeech-PC test-clean *cross-sentence* list (Meister et al., 2023) (1,127 utterances), where LibriSpeech-PC is LibriSpeech (Panayotov et al., 2015) with restored punctuation and casing; Seed-TTS test-en (1,088 utterances); and Seed-TTS test-zh (Anastassiou et al., 2024) (2,020 utterances), whose Mandarin material derives from DiDiSpeech (Guo et al., 2021).

4.3 SWAP TEST

To evaluate the localization of power within the designated latent channels, we conduct a swap test, as detailed in Algorithm 1. In this procedure, the first k channels of a latent representation derived from an augmented signal (+6 dB) are replaced with the corresponding channels from the original signal. The resulting power ratio, R_{dB}, of the decoded waveform serves as a measure of decoupling.

Algorithm 1 Swap test for power localization in first k latent channels

Require: $x \in \mathbb{R}^N$; $\mathrm{Enc} : \mathbb{R}^N \to \mathbb{R}^{L\times T}$; $\mathrm{Dec} : \mathbb{R}^{L\times T} \to \mathbb{R}^N$

$Z \in \mathbb{R}^{L\times T} \leftarrow \mathrm{Enc}(x)$

$Z' \in \mathbb{R}^{L\times T} \leftarrow \mathrm{Enc}(2x)$ ▷ +6 dB

$Z'[:k,:] \leftarrow Z[:k,:]$

$x' \in \mathbb{R}^N \leftarrow \mathrm{Dec}(Z')$

$R_{\mathrm{dB}} \leftarrow 10\log_{10} \frac{\sum_{n=1}^{N} x'[n]^2}{\sum_{n=1}^{N} x[n]^2}$

return R_{dB}

An ideal result of $R_{\mathrm{dB}} \approx 0$ indicates that the power variation is effectively isolated within the first k channels, as the swap successfully neutralizes the added gain. Conversely, $R_{\mathrm{dB}} \approx 6$ suggests that power information remains entangled across the remaining channels. In the latter case, the decoupling loss coefficient λ_{PoDAR} must be increased to reduce the coupling between semantic and power information.

5 RESULTS

5.1 CODEC RECONSTRUCTION AND SWAP TEST

Table 1 summarizes the reconstruction fidelity and swap-test results for the autoencoder representations described in Section 4.1.

Table 1: ViSQOL and swap-test gain per λ_{PoDAR}.

VAE	λ_{PoDAR}	ViSQOL	Swap Gain (dB)
VAE1 (baseline)	0.0	4.01 ± 0.05	-
VAE1	0.1	4.06 ± 0.05	$+2.21 \pm 0.04$
VAE1	0.5	3.70 ± 0.04	$+0.99 \pm 0.02$
VAE1	0.75	3.65 ± 0.05	$+0.97 \pm 0.02$
VAE2 (baseline)	0.0	3.96 ± 0.06	-
VAE2	0.1	3.96 ± 0.06	$+0.83 \pm 0.02$

A 1 dB shift in signal amplitude is widely recognized as the just noticeable difference (JND) for human perception of audio loudness. Accordingly, the decoupling weight λ_{PoDAR} is tuned for both

representations to ensure that the measured swap gain remains below this perceptual threshold. Empirical evaluation demonstrates that the required λ_{PoDAR} varies by architecture. For VAE2, a loss weight of 0.1 already achieves a swap gain below the 1 dB threshold while simultaneously preserving reconstruction quality. In contrast, this weight proves inadequate for the requirements of VAE1. While a coefficient of 0.5 or above reduced the swap gain below the 1 dB JND threshold, the reconstruction fidelity exhibits a clear monotonic decline as these weights are scaled.

The ViSQOL scores for VAE1 degrade significantly as λ_{PoDAR} is scaled to 0.5. This reduction is anticipated, as the objective imposes explicit structural constraints on the latent space to enforce factor decoupling. However, as discussed in Section 1, superior reconstruction fidelity does not consistently translate to improved downstream generative performance (Xu et al., 2026). Consequently, the reduction in reconstruction quality is justified insofar as the enhanced modelability of the latent space translates to superior performance in the final generative model.

While the matched swap test in Table 1 demonstrates that global gain variations are softly decoupled, achieving perfect disentanglement is not the objective of this framework. Phonetic content inherently dictates relative loudness (e.g., vowels are naturally louder than unvoiced fricatives), creating an intrinsic acoustic coupling. This is observable in non-matched swap tests, where swapping power channels between unrelated utterances reveals residual energy correlations tied to the original phonetic structure. Furthermore, as detailed in Appendix E, linear probing experiments confirm that some power information remains in the content channels. Ultimately, PoDAR achieves a soft decoupling that routes sufficient variance away from the semantic channels to make the distribution significantly easier for the downstream generative model to learn.

5.2 Generator results

We evaluate the generative performance of the PoDAR representation using the LibriSpeech and Seed-TTS benchmarks. Fig. 1 illustrates the comparative trajectories for WER, speaker similarity, and UTMOS as defined in Section 4.2 for both baseline and PoDAR configurations across VAE1 and VAE2 as a function of training steps. The final performance for these models is summarized in Table 2.

Table 2: Final WER, speaker similarity, and UTMOS performance across baseline and PoDAR configurations for VAE1 and VAE2 on the LibriSpeech and Seed-TTS benchmarks. **Bold** indicates the best within each dataset, VAE family, and metric.

Model	WER ↓	speaker SIM ↑	UTMOS ↑
LibriSpeech-PC			
VAE1-Baseline	0.023 ± 0.003	0.592 ± 0.005	3.75 ± 0.03
VAE1-PoDAR	0.023 ± 0.003	$\mathbf{0.647 \pm 0.005}$	$\mathbf{3.97 \pm 0.02}$
VAE2-Baseline	0.024 ± 0.003	0.559 ± 0.005	4.10 ± 0.02
VAE2-PoDAR	0.025 ± 0.003	$\mathbf{0.595 \pm 0.005}$	$\mathbf{4.17 \pm 0.01}$
Seed-TTS (EN)			
VAE1-Baseline	0.017 ± 0.003	0.613 ± 0.006	3.69 ± 0.02
VAE1-PoDAR	0.017 ± 0.003	$\mathbf{0.660 \pm 0.005}$	$\mathbf{3.72 \pm 0.02}$
VAE2-Baseline	0.017 ± 0.003	0.594 ± 0.006	3.87 ± 0.02
VAE2-PoDAR	0.019 ± 0.003	$\mathbf{0.615 \pm 0.006}$	3.89 ± 0.02
Seed-TTS (ZH)			
VAE1-Baseline	0.016 ± 0.002	0.713 ± 0.003	2.79 ± 0.02
VAE1-PoDAR	0.017 ± 0.002	$\mathbf{0.761 \pm 0.002}$	$\mathbf{2.87 \pm 0.02}$
VAE2-Baseline	0.020 ± 0.002	0.688 ± 0.003	2.98 ± 0.02
VAE2-PoDAR	0.018 ± 0.002	$\mathbf{0.701 \pm 0.003}$	2.96 ± 0.02

The PoDAR representation facilitates accelerated convergence across all evaluated datasets, requiring approximately $2\times$ fewer training steps to reach speaker similarity performance comparable to the baseline. Speaker similarity is notably higher for the generator trained on PoDAR, with a consistent improvement of 0.055 for VAE1 across both benchmarks and a slightly smaller gain for VAE2.

Regarding UTMOS, VAE1 demonstrates a 0.22 increase on LibriSpeech, while the improvement for VAE2 is more moderate, which we hypothesize stems from baseline regularization. As can be seen in Table 5, VAE2 applies a KL penalty of 10^{-4}, two orders of magnitude higher than VAE1 at 10^{-6}, restricting the latent manifold for additional structure. Both architectures show only slight changes on the Seed-TTS dataset. These findings demonstrate that although the PoDAR codec exhibits lower reconstruction fidelity as described in Section 5.1, the resulting representation is more modelable and yields superior UTMOS scores compared to the baseline. Crucially, the contrasting results between the two architectures highlight that PoDAR can independently induce a highly modelable representation without relying on severe KL penalties (VAE1), while still providing consistent benefits atop already highly regularized spaces (VAE2).

Finally, the WER remains largely unchanged because the results for both configurations fall within the confidence intervals of the ground truth WER reported in Chen et al. (2025).

6 Ablations

All ablation experiments are conducted utilizing the VAE1 representation, with results reported on the LibriSpeech-PC dataset.

6.1 Impact of partial CFG

This ablation evaluates the impact of the partial CFG strategy within the PoDAR framework. We compare a configuration where CFG is applied across the full latent vector, including the power channels, against a selective approach where guidance is restricted to the power-invariant channels as described in Section 3.3. The performance metrics for both configurations are illustrated in Fig. 2.

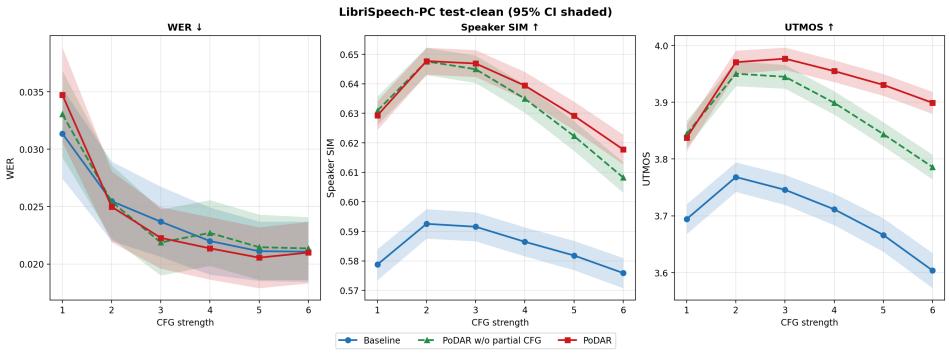


Figure 2: Impact of partial CFG on WER, speaker SIM and UTMOS performance across CFG scales.

The PoDAR implementation without partial CFG already yields substantial performance gains over the baseline, in both speaker similarity and UTMOS metrics. While adding partial CFG achieves marginally higher peak values for both, the primary utility is its robustness at elevated guidance scales. Whereas the performance of the PoDAR implementation without partial CFG degrades rapidly as the CFG scale increases, the partial CFG strategy sustains high metric values across a broader range of guidance values.

6.2 Impact of λ_{PoDAR}

This ablation evaluates generators trained on VAE1 representations with varying λ_{PoDAR} coefficients detailed in Table 1. We investigate how the strength of the decoupling objective during autoencoder training influences the performance of the downstream generator. The resulting comparisons across these configurations are illustrated in Fig. 3.

Within the evaluated hyperparameter range, increasing λ_{PoDAR} results in measurable improvements in speaker similarity and UTMOS metrics. A distinct divergence appears between the swap test results and the downstream generator performance. Although the configurations for $\lambda_{\text{PoDAR}} = 0.5$ and $\lambda_{\text{PoDAR}} = 0.75$ yielded comparable swap gain results in Table 1, the latter achieves higher scores

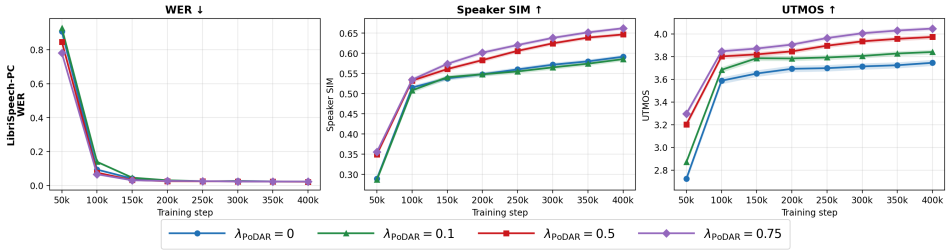


Figure 3: Impact of λ_{PoDAR} on WER, speaker SIM, and UTMOS.

in speaker similarity and UTMOS. This finding suggests that the strength of the decoupling objective can influence generative performance beyond the attribution captured by the swap test.

6.3 Impact of k

A central design choice of PoDAR is the capacity allocated to the power subspace, determined by the number of power channels, k. Throughout the paper we utilize $k = 1$, which, as shown in Section 5.1, empirically provided sufficient capacity to reduce the swap test error below the JND threshold of 1 dB.

To verify this capacity requirement, we conducted an ablation using the VAE1 representation with $k = 2$ power channels. The downstream model trained on the $k = 2$ representation yielded UTMOS and speaker similarity metrics that were statistically indistinguishable from the $k = 1$ configuration. These findings confirm that a single power channel provides sufficient representational capacity to capture the energy variance and allocating additional channels to the power subspace does not yield further generative improvements.

7 Conclusion

We introduced PoDAR (Power-Decoupled Audio Representation), a framework designed to enhance the modelability of continuous audio latent spaces by utilizing a randomized power augmentation and latent consistency objective to decouple signal power from invariant semantic content. Empirical evaluation on the LibriSpeech-PC benchmark demonstrates that PoDAR achieves about a $2\times$ acceleration in convergence to match baseline performance while concurrently providing gains of 0.055 in speaker similarity and 0.22 in UTMOS. Moreover, this decoupling enables a partial CFG strategy that restricts semantic steering to the power-invariant subspace, effectively improving the robustness of the generator at elevated guidance scales.

Limitations. The primary limitations of this work involve the increased computational overhead during training of the autoencoders, since the consistency objective requires dual encoder passes for both the original sample x and the augmented input $\mathcal{A}(x)$. Furthermore, the efficacy of the framework has been demonstrated exclusively within the speech domain and specifically targets power decoupling rather than broader acoustic factors or alternative modalities.

Future work. While our downstream generative evaluation is strictly on TTS, the underlying PoDAR autoencoders are trained on a mixture of speech, music, and environmental audio. A natural extension of this research is to evaluate it on downstream text-to-audio (TTA) and text-to-music (TTM) tasks, where signals often exhibit even wider dynamic ranges. Future research may extend the augmentation-consistency principle to alternative modalities such as images and incorporate additional factors like voicing, pitch, saturation, or contrast by defining suitable augmentations and allocating dedicated latent subspaces for multi-factor decoupling. It remains an open question whether this augmentation-consistency framework can be applied to non-linear augmentations that introduce artifacts.

AI USE STATEMENT

In this work, we used generative AI tools to assist in implementing methods, using AI coding assistants to write and edit parts of our augmentation, loss and partial CFG code. We have not used generative AI tools for any other task in the required-disclosure category of the ICLR 2027 AI policy. Additionally, we used generative AI tools to draft parts of the paper, to edit the text for readability, to search for information, and to format references. We have reviewed and edited all AI-assisted work, including AI-generated code and AI-drafted or AI-edited text. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work involves no human subjects, releases no data, models, or code, and raises no privacy or data-security issues. All baselines, both the Stable Audio autoencoders and the F5-TTS generator, and all training and evaluation datasets are open source or publicly available and are used in accordance with their licenses, listed in Appendix A (Table 3). The main consideration is that by roughly halving the training needed to reach a given generation quality, PoDAR reduces the training compute (GPU-hours in Appendix B) and with it the energy consumption and carbon footprint of training. However, the same efficiency lowers the barrier to misuse of speech generation, such as unauthorized voice cloning or synthetic-speech disinformation. Finally, our evaluation covers only English and Mandarin, so we make no claims about performance or fairness across other languages.

REPRODUCIBILITY STATEMENT

We do not release code, but the paper contains everything needed to reproduce our results. The method consists of three self-contained modifications, each fully specified in Section 3: the power augmentation, the consistency objective in (6), and the partial CFG rule in (8). Both autoencoder baselines and the generator are re-trained from open-source implementations, `stable-audio-tools` (Stability AI, 2025) and F5-TTS (Chen et al., 2025), following their published recipes (Section 4.1), and all training and evaluation data are public datasets listed with their licenses in Appendix A. Section 4 gives the training mixtures, data splits, test sets, and evaluation protocol, and Appendix C lists the training and inference hyperparameters of the baseline and PoDAR-modified autoencoders and of the latent generator, including the sampler configuration and random seed (Tables 5 and 6).

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. *arXiv preprint arXiv:2406.02430*, 2024.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. In *International Conference on Language Resources and Evaluation (LREC)*, pp. 4218–4222, 2020.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop at the International Conference on Machine Learning (ICML)*, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

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.

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. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6255–6271, 2025.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech*, pp. 3830–3834, 2020.

Kieran Didi. The unification of representation learning and generative modelling. Blog post, 2025. URL `https://kdidi.netlify.app/blog/ml/2025-12-31-r4g/`.

Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, and Robert Aichner. ICASSP 2022 deep noise suppression challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9271–9275, 2022.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, 2021.

Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *International Conference on Machine Learning (ICML)*, pp. 12652–12665, 2024.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable Audio Open. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. FunASR: A fundamental end-to-end speech recognition toolkit. In *Interspeech*, pp. 1593–1597, 2023.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

Philipp Grundhuber, Mhd Modar Halimeh, and Emanuël A. P. Habets. Acoustic teleportation via disentangled neural audio codec representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, and Xiangang Li. DiDiSpeech: A large scale Mandarin speech corpus. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6968–6972, 2021.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 885–890, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Andrew Hines, Jan Skoglund, Anil C. Kokaram, and Naomi Harte. ViSQOL: An objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, 2015.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272, 2025.

Jiaqi Li, Xiaolong Lin, Zhekai Li, Shixi Huang, Yuancheng Wang, Chaoren Wang, Zhenpeng Zhan, and Zhizheng Wu. DualCodec: A low-frame-rate, semantically-enhanced neural audio codec for speech generation. In *Interspeech*, pp. 4883–4887, 2025.

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

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning (ICML)*, pp. 21450–21474, 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.

Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. LibriSpeech-PC: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end ASR models. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2023.

Gautham J. Mysore. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—A dataset, insights, and challenges. *IEEE Signal Processing Letters*, 22(8):1006–1010, 2015.

Ahmed Omran, Neil Zeghidour, Zalán Borsos, Félix de Chaumont Quitry, Malcolm Slaney, and Marco Tagliasacchi. Disentangling speech from surroundings with neural embeddings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pp. 28492–28518, 2023.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation. Zenodo, 2017. URL `https://doi.org/10.5281/zenodo.1117372`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Samir Sadok, Laurent Girin, and Xavier Alameda-Pineda. The equalizer: Introducing shape-gain decomposition in neural audio codecs. *arXiv preprint arXiv:2602.15491*, 2026.

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

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *International Conference on Learning Representations (ICLR)*, 2026.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Stability AI. stable-audio-tools. GitHub repository, 2025. URL https://github.com/Stability-AI/stable-audio-tools.

Tongda Xu, Mingwei He, Shady Abu-Hussein, José Miguel Hernández-Lobato, Chunhang Zheng, Kai Zhao, Chao Zhou, Ya-Qin Zhang, and Yan Wang. Making reconstruction FID predictive of diffusion generation FID. *arXiv preprint arXiv:2603.05630*, 2026.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). University of Edinburgh, The Centre for Speech Technology Research (CSTR), 2019. URL https://datashare.ed.ac.uk/handle/10283/3443.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, 2020.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations (ICLR)*, 2026.

A ASSET LICENSES

Table 3 lists every code repository, dataset, and pretrained model used in this paper, together with its license and primary citation. All assets are publicly available; no proprietary data or weights are used.

Table 3: Licenses for all third-party assets used in this paper.

Asset	Type	License	Reference
Code / pretrained models			
`stable-audio-tools`	VAE training code	MIT	Stability AI, 2025
F5-TTS	Generator code	MIT	Chen et al., 2025
faster-whisper (large-v3)	ASR (en WER)	MIT	Radford et al., 2023
FunASR `paraformer-zh`	ASR (zh WER)	MIT	Gao et al., 2023
WavLM-Large + ECAPA-TDNN	SIM-o backbone	MIT/Apache 2.0	Chen et al., 2022; Desplanques et al., 2020
SpeechMOS `utmos22_strong`	UTMOS predictor	MIT	Saeki et al., 2022
Autoencoder training datasets (DAC mixture)			
DAPS	Speech	CC BY-NC 4.0	Mysore, 2015
DNS Challenge 4 (clean speech)	Speech	CC BY 4.0	Dubey et al., 2022
Common Voice	Speech	CC0 1.0	Ardila et al., 2020
VCTK	Speech	ODC-By 1.0	Yamagishi et al., 2019
MUSDB18	Music	CC BY-NC-SA 4.0	Rafii et al., 2017
MTG-Jamendo	Music	CC BY-NC-SA 4.0	Bogdanov et al., 2019
AudioSet (bal. + unbal. train)	Environmental	CC BY 4.0	Gemmeke et al., 2017
Generator training dataset			
Emilia ZH-EN	Speech (bilingual)	CC BY-NC 4.0	He et al., 2024
Evaluation datasets			
LibriSpeech (test-clean)	Speech (en)	CC BY 4.0	Panayotov et al., 2015
LibriSpeech-PC (cross-sentence)	Speech (en)	CC BY 4.0	Meister et al., 2023
Seed-TTS test-en	Speech (en)	CC BY-NC 4.0	Anastassiou et al., 2024
Seed-TTS test-zh / DiDiSpeech	Speech (zh)	CC BY-NC 4.0	Anastassiou et al., 2024; Guo et al., 2021

B COMPUTE RESOURCES

All training experiments were conducted using 4 NVIDIA H200 GPUs and the same hardware was used for inference and metric evaluation. Table 4 summarizes the per run and aggregate compute footprint for the training results reported in this paper, totaling approximately 4,608 GPU hours.

Table 4: Training-compute footprint reported in this paper. *Codecs* covers the 4 autoencoders used for the main results (VAE1 / VAE2 $\times$ baseline / PoDAR) plus the 2 PoDAR-weight ablations on VAE1 ($\lambda_{\text{PoDAR}} \in \{0.1, 0.75\}$). *Generators* covers one F5-TTS v1 Base latent generator per codec (400,000 optimizer updates each).

Phase	Per-run wall-clock	Runs	GPUs/run	Subtotal (GPU-h)
Codec training (1 epoch on DAC mixture)	$\sim$ 1 day	6	4 $\times$ H200	576
Generator training (400,000 updates)	$\sim$ 7 days	6	4 $\times$ H200	4,032
Training total	—	—	—	**4,608**

We estimate that the total computational budget for the full research cycle including inference runs and unsuccessful iterations was below 10,000 GPU hours.

C FULL TRAINING HYPERPARAMETERS

Table 5 summarizes the structural differences and optimization strategies for the baseline Stable Audio autoencoders utilized throughout this work.

Table 5: Stable Audio autoencoder baselines from `stable-audio-tools`.

Aspect	Stable Audio 1.0 VAE	Stable Audio 2.0 VAE
Backbone	**DAC**	**Oobleck**
Encoder strides	`[4, 4, 8, 8]`	`[2, 4, 4, 8, 8]`
Output nonlinearity	**Tanh**	**None**
Discriminator filters	`32`	`64`
KL weight	`1e-6`	`1e-4`
AE optimizer	LR `1e-4`, no weight decay	LR `1.5e-4`, weight decay `1e-3`
LR schedule	ExponentialLR (`gamma=0.999996`)	InverseLR (`inv_gamma=200000`, `power=0.5`, `warmup=0.999`)
Discriminator	LR `1e-4`	LR `3e-4`

Additionally, we provide the training configurations necessary for the full reproduction of our experimental results in this section. Table 6 details the specific hyperparameters utilized for both the latent generator and the PoDAR modified autoencoder architectures including the specialized power augmentation settings.

D LATENT VARIANCE ANALYSIS

To quantify the impact of the L_2 consistency objective ($\mathcal{L}_{\text{PoDAR}}$) on the latent space, we measured both the scale of the latent manifold (the standard deviation of the predicted means, μ) and the posterior uncertainty (the mean of the predicted standard deviations, σ) on the DAPS clean test set. The results are summarized in Table 7.

The $\mathcal{L}_{\text{PoDAR}}$ objective inherently regularizes the overall latent space, but the magnitude of this effect is highly dependent on the baseline KL penalty. For VAE1, which employs a low KL penalty (10^{-6}), the initially unconstrained baseline manifold shrinks significantly, with the signal scale dropping from 1.482 to 0.395. Conversely, for VAE2, which utilizes a much higher KL penalty (10^{-4}), the manifold is already tightly constrained, so the additional L_2 loss exerts a much smaller relative impact on the signal scale, from 0.993 to 0.863.

Crucially, these statistics justify our choice in Section 3 to apply $\mathcal{L}_{\text{PoDAR}}$ exclusively to the means (μ) rather than the standard deviations (σ). For VAE1, the posterior σ is 0.006, practically zero, rendering an L_2 penalty applied to it meaningless. For VAE2, the high KL penalty independently anchors the posterior σ, keeping it stable at approximately 0.12 even with the addition of PoDAR. Therefore, applying the consistency penalty to the variance parameter is unnecessary across both operating regimes.

E STRICTNESS OF REPRESENTATION DECOUPLING

As discussed in Section 5.1, our goal is soft decoupling rather than perfect disentanglement, given that acoustic semantic content and energy are intrinsically coupled. To evaluate the strictness of the separation between the power and semantic subspaces, we trained a linear probe to predict the power channel using solely the remaining semantic channels as input. The performance metrics for this evaluation are summarized in Table 8.

Table 6: Generator and autoencoder training hyperparameters. Generator configuration mirrors the published F5-TTS v1 Base recipe with input/output projections resized to the autoencoder latent dimension.

Component	Hyperparameter	Value
F5-TTS v1 Base latent generator		
Backbone	Architecture	DiT
	Hidden dim / depth / heads	1024 / 22 / 16
	FF mult / text dim	2 / 512
	ConvNeXt-V2 text layers	4
Conditioning	$p_{\text{audio drop}}$	0.3
	p_{uncond}	0.2
Optim	Optimizer	fused AdamW
	Learning rate (peak)	7.5×10^{-5}
	Weight decay	0
	Grad clip (ℓ_2)	1.0
	Warmup updates	20,000
	Schedule	linear warmup, linear decay
	Mixed precision	fp16
	EMA	0.9999
Sampling	ODE solver	Euler
	NFE	32
	Sway sampling	-1
	Seed	0
	CFG scale	3.0 (VAE1), 2.0 (VAE2)
Stable Audio 1.0 / 2.0 VAE (PoDAR-modified)		
Data	Segment length	65,536 samples (1.49 s)
	Sample rate	44,100 Hz
	Mixture	DAC training mix (Section 4.1)
PoDAR	Power channels k	1
	λ_{PoDAR} (VAE1)	0.5 (main); 0.1, 0.75 (ablation)
	λ_{PoDAR} (VAE2)	0.1
	Augmentation	uniform gain in $[-6, +6]$ dB

Table 7: Impact of the $\mathcal{L}_{\text{PoDAR}}$ objective on latent manifold statistics measured on the DAPS clean test set.

Codec	Posterior std (mean of σ)	Signal scale (std of μ)
VAE1-Baseline	0.006	1.482
VAE1-PoDAR	0.002	0.395
VAE2-Baseline	0.120	0.993
VAE2-PoDAR	0.124	0.863

Table 8: Linear probe performance predicting the power channel from the semantic channels.

Representation	R^2	Pearson r
VAE1-PoDAR	0.67	0.83
VAE2-PoDAR	0.66	0.81

These results show that, while PoDAR successfully routes enough energy variance into the power subspace to improve latent modelability, the semantic channels retain phonetically-linked power information.

F LATENT-SPACE SMOOTHNESS

As noted in Section 4.2, we also probe each codec by perturbing its latents and measuring how much the decoded audio degrades. The intuition behind this is that a latent space is more modelable if it is smooth, i.e. if nearby latents decode to perceptually similar audio. The more tolerant it is to local perturbations, the easier the task of the downstream generative model. These geometric probes capture large differences in latent geometry, but they are not sensitive enough to serve as a standalone modelability metric, so we report them here as a diagnostic.

Data and segmentation. We evaluate on the clean subset of the DAPS corpus. Each recording is converted to mono, resampled to the codec sample rate (44.1 kHz), and power-normalised to -16 dB. From every file we take four non-overlapping 10 s segments starting 5 s into the recording, giving 40 segments in total; each segment is zero-padded to a multiple of the encoder downsampling ratio. All four codecs see the same segments, and all randomness is seeded per segment.

Covariance-shaped perturbations. Let $\boldsymbol{Z} \in \mathbb{R}^{C\times T}$ be the latent of one segment, with C channels and T frames, and let $\Sigma = \mathrm{Cov}(\boldsymbol{Z}) \in \mathbb{R}^{C\times C}$ be its empirical channel covariance across time, with eigendecomposition $\Sigma = V\,\mathrm{diag}(\lambda)V^\top$. We define the square-root factor $A = V\,\mathrm{diag}(\lambda^{1/2})$ and perturb the latent as

$$\widetilde{\boldsymbol{Z}}(\sigma) = \boldsymbol{Z} + \sigma\, A\boldsymbol{E}, \qquad E_{c,t} \sim \mathcal{N}(0,1) \text{ i.i.d.} \tag{9}$$

so that the injected noise has covariance $\sigma^2\Sigma$ per frame. Shaping the noise by the codec's own latent covariance makes a perturbation of magnitude σ correspond to the same *relative* displacement along every principal direction of the latent distribution, independent of the absolute scale or anisotropy of a particular codec.

Distortion metric. Degradation is measured by mel-cepstral distortion (MCD). Mel spectrograms use 96 mel bands with hop length equal to the encoder downsampling ratio and window length twice the hop (50% overlap), so that one mel frame still corresponds to exactly one latent frame. We use 13 cepstral coefficients $c_1, \ldots, c_{13}$, excluding c_0 so that a constant loudness offset does not penalise every frame, and flooring the mel magnitudes at 10^{-5} so that silent frames do not dominate:

$$\mathrm{MCD} = \frac{10}{\ln 10} \cdot \frac{1}{T} \sum_{t=1}^{T} \sqrt{2 \sum_{k=1}^{13} \big(c_{k,t}^{(1)} - c_{k,t}^{(2)}\big)^2}. \tag{10}$$

Smoothness (robustness) curve. For each segment we sweep σ over 10 logarithmically spaced values in $[10^{-3}, 10^{0}]$, decode $\widetilde{\boldsymbol{Z}}(\sigma)$ from Eq. (9), and report the MCD between the decoded audio and the original waveform. The dashed horizontal lines in Fig. 4 give the unperturbed reconstruction MCD of each codec, i.e. the floor that the curve approaches as $\sigma \to 0$. These floors are relatively high because codec reconstruction is imperfect, particularly on some DAPS segments, so even $\sigma = 0$ already incurs substantial MCD relative to the original waveform. A curve that stays flat over a wider range of σ before rising indicates a smoother latent space: nearby latents decode to perceptually similar audio, so a downstream generator operating in that space is less sensitive to prediction error. Shaded bands are 95% confidence intervals of the mean over the 40 segments. The ranking in Fig. 4 is VAE2-PoDAR slightly more robust than VAE2-Baseline, which is more robust than VAE1-PoDAR, which is in turn substantially more robust than VAE1-Baseline. This ordering matches the UTMOS trend (and the gaps between the performance of the different codecs) in Table 2 for LibriSpeech-PC but it is not very conclusive for the other datasets.

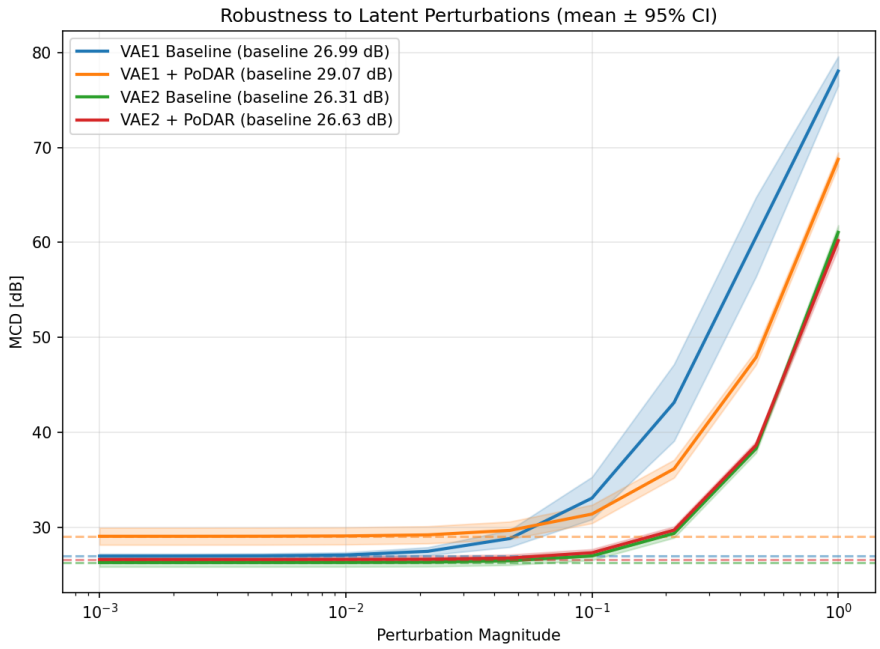


Figure 4: Robustness to covariance-shaped latent perturbations. MCD is measured between the original waveform and the decode of the perturbed latent, as a function of perturbation magnitude σ. Dashed lines are unperturbed reconstruction MCD. Mean $\pm$ 95% CI over 40 DAPS segments.