# PRINCIPAL COMPONENT ALIGNMENT FOR TRAINING A GENERAL FULL-BAND STEREO AUDIO VAE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Latent alignment to a frozen foundation model is now a standard practice for audio VAEs. The two choices it involves, which teacher and how its features are projected, turn out to govern diffusability: how well a diffusion model trained on the resulting latent space actually generates. We revisit both. Where prior works learn the projection relating teacher features to the latent, we align to principal components of those features. As the teacher we use the audio encoder of a perceptual quality predictor rather than a semantic encoder, and for general audio this proves the better choice. Side-by-side listening tests confirm improved prompt following and audio quality in the generations. Reconstruction quality, meanwhile, turns out to be an unreliable guide: the configuration that reconstructs better is not the one that generates better. The improved diffusability then buys room in the latent to carry a second channel at no cost. Our 128-dimensional stereo latent ties its otherwise identical 64-dimensional mono counterpart at 50/50 on prompt following and on downmixed audio quality. The result is a full-band $48\,\mathrm{kHz}$ VAE for general audio (speech, music, other sounds and mixtures of these) at a $50\,\mathrm{Hz}$ latent rate, which wins over $70\%$ of audio-quality comparisons against the audio VAEs of LTX 2.5 and Stable Audio 3 under a fixed diffusion transformer and training budget.

## 1 INTRODUCTION

Latent diffusion models generate audio in a compressed representation space rather than on the waveform (Rombach et al., 2022), which has made high-fidelity generation practical in systems such as AudioLDM (Liu et al., 2023a), Make-An-Audio (Huang et al., 2023), and Stable Audio (Evans et al., 2024a).

The first-stage autoencoder fixes two things at once: how much of the signal survives compression, and how easy the resulting latent sequence is for a diffusion model to learn. The second property is known as *diffusability* (Skorokhodov et al., 2025). The two are not the same, and work on autoencoders for latent diffusion has repeatedly found that they pull in opposite directions: a wider latent reconstructs better but is harder to model (Yao et al., 2025; Skorokhodov et al., 2025). For stereo audio this conflict cannot be avoided. A latent that is wide enough to reconstruct a two-channel, full-band signal is harder to model than a narrow mono latent. A latent narrow enough to model well does not have room for the second channel. Section 3 makes this tension concrete: under a fixed diffusion transformer and training budget, doubling the latent width of a stereo VAE improves every reconstruction metric and, at the same time, loses the listening test against the narrower model.

Aligning the latent to a frozen pretrained audio model is the standard remedy for poor diffusability (Yao et al., 2025; Niu et al., 2025; Cheng et al., 2026). Alignment involves two choices that prior work has largely fixed by convention: which teacher to align to, and how the teacher's features are brought to the latent dimensionality. We find that the second matters more than is usually assumed. Where prior work learns the projection jointly with the VAE, we align to a principal-component basis of the teacher features that is fitted incrementally during training and receives no gradient. The learned projection reconstructs better, but the principal-component projection generates better. As the teacher we use the audio encoder of a perceptual-quality predictor (Tjandra et al., 2025b) rather than a speech-oriented self-supervised encoder. Linear probes and temporal self-similarity diagnostics point to the explanation: a trainable projection can satisfy the alignment loss without transferring the teacher's structure to the latent, whereas a fixed basis cannot.



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.

With diffusability restored, the channel budget can be spent on stereo. Our final model is a waveform VAE that directly encodes and reconstructs full-band $48\,\mathrm{kHz}$ stereo across general audio (speech, music, other sounds, and mixtures of these), with a $128$-dimensional latent at $50\,\mathrm{Hz}$. In side-by-side listening tests it outperforms the audio VAEs of SAME (Parker et al., 2026) and LTX 2.5 (HaCohen et al., 2026; Lightricks Research, 2026) on both prompt following and audio quality.

Our contributions are threefold. First, we provide evidence that, for latent alignment, a fixed principal-component projection of teacher features yields a more diffusable latent than a trainable projection, together with an analysis of why. Second, we present a controlled comparison of alignment teachers for general audio, showing that a perceptual-quality encoder is preferable to a speech-oriented self-supervised encoder. Third, we build a waveform VAE that carries full-band $48\,\mathrm{kHz}$ stereo, and show that it outperforms open stereo audio VAEs under a fixed diffusion transformer and training budget.

## 2 Related Work and Background

**Audio compression and waveform synthesis.** SoundStream, EnCodec, and the Descript Audio Codec (DAC) established residual vector quantization over convolutional waveform autoencoders trained with multi-scale spectral and adversarial losses (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023). Our model reuses DAC's convolutional backbone with Snake activations (Ziyin et al., 2020; Lee et al., 2023) and its multi-period and multi-resolution spectrogram discriminators (Kong et al., 2020; Jang et al., 2021), but replaces the discrete bottleneck with the continuous latent that a diffusion model requires.

**Continuous audio representations.** VAEs regularize an encoded distribution through a Gaussian prior (Kingma & Welling, 2014; Higgins et al., 2017). RAVE demonstrated convolutional waveform VAEs for neural synthesis (Caillon & Esling, 2022), and AudioLDM transferred latent diffusion to audio through a spectrogram representation (Liu et al., 2023a; Rombach et al., 2022). Open stereo VAEs for generation, such as SAME-L, which is also the audio VAE of Stable Audio 3 (Parker et al., 2026; Evans et al., 2026), have so far been built mainly for music and sound effects. LTX 2.5 covers general audio and emits a $48\,\mathrm{kHz}$ stereo waveform, but its VAE encodes mel-spectrograms computed at $16\,\mathrm{kHz}$. A vocoder with a bandwidth-extension stage synthesizes the waveform and the frequencies absent from that analysis representation (HaCohen et al., 2026; Lightricks Research, 2026). Its output is therefore formally full-band, whereas information above the $8\,\mathrm{kHz}$ Nyquist frequency of the VAE input is generated rather than transmitted through the latent. To our knowledge, no open waveform VAE directly encodes and reconstructs full-band stereo across general audio including speech, music, and sound effects. $\varepsilon$ar-VAE introduced the cross-channel phase coherence metric for stereo-image fidelity (Wang et al., 2025). Section 8 compares our model with SAME-L and the LTX 2.5 audio VAE.

**Alignment and latent geometry.** SpeechTokenizer, X-Codec, and Semantic-VAE inject self-supervised speech features into discrete or continuous speech representations (Zhang et al., 2024; Ye et al., 2024; Niu et al., 2025), and AudioLDM 2 generates audio through AudioMAE features (Liu et al., 2024). In vision, REPA aligns diffusion features with a frozen encoder, and VA-VAE aligns the autoencoder latent using cosine and pairwise-distance objectives (Yu et al., 2025; Yao et al., 2025). The iREPA analysis associates spatial self-similarity with generative performance across vision encoders (Singh et al., 2025). Two recent audio works are closest to ours. JMAS-VAE (Cheng et al., 2026) adopts the VA-VAE objective, with the same margins we use, for a speech VAE. It projects the $64$-dimensional latent *up* to the WavLM feature dimension with a trainable linear layer and does not compare projection choices. OmniVAE (Zhan et al., 2026) argues, as we do, that a speech encoder is the wrong teacher for general audio, and distills a multimodal semantic encoder through trainable projectors. We differ from both in projecting the teacher *down* to the latent with a fixed, gradient-free principal-component basis, and in choosing a perceptual-quality encoder rather than a semantic one as the teacher. Separately, layer-wise WavLM matching has been used as a waveform perceptual loss (Parker et al., 2025). These precedents motivate the individual ingredients. The projection comparison across general audio and its consequences for stereo are the focus of this paper.

# 3 Problem Statement: Channel Budget versus Diffusability

Our starting point is a mono VAE with a 64-dimensional, 50 Hz latent, trained on general audio with the recipe of section 4. Keeping the architecture fixed and switching the input and output to two channels degrades reconstruction across all domains (Table 7 in section D.1). Switching to stereo at 64 channels raises MEL from $0.76/0.74/0.55$ to $0.93/0.89/1.66$ for sounds, music, and speech, and lowers SI-SDR by 3.6 and 5.3 dB on sounds and music. Doubling the latent to 128 channels recovers most of the loss, but a diffusion transformer trained on that latent under the same budget generates worse audio. The signal-level metrics on sounds remain low for both stereo models, because this subset contains the most complex and highest-quality material in the test set and phase errors between the two channels are penalized heavily. The stereo-specific objectives of section 4.3, ablated in section 7, address this problem. In a side-by-side listening test between the mono-64 and the stereo-128 model, both with the linear layer projection for alignment, the mono model wins $55/45$ on prompt following and $64/36$ on audio quality. Capacity added for reconstruction is thus not automatically usable by the generator. Section 6 addresses this by revisiting the projection used for alignment, and section 7 returns to the 128-channel stereo latent with the resulting recipe.

# 4 Method

## 4.1 Model

Our model is a waveform VAE based on the Descript Audio Codec architecture (Kumar et al., 2023) (fig. 10 in section F). The encoder takes one or two waveform channels at 48 kHz and downsamples with five residual blocks of strides $(2, 3, 4, 5, 8)$, a total hop of 960 samples. A multi-head self-attention layer (Vaswani et al., 2017) after the last block captures global context before a pointwise projection produces the mean $\boldsymbol{\mu}$ and log-variance of a diagonal Gaussian posterior over a $d_z$-dimensional latent (Kingma & Welling, 2014). The decoder mirrors the encoder and outputs the same number of channels as the input. The latent rate is therefore 50 Hz (20 ms per frame, 500 frames for ten seconds of audio). For stereo, left and right are encoded *jointly*: the latent is not a concatenation of two mono streams, and each frame carries inter-channel information in its feature dimensions. We use $d_z = 64$ for the mono models of section 6 and $d_z = 128$ for the stereo model of section 7. Everything else in the architecture and the training recipe is identical. Section F lists layer widths and the remaining architectural details.

## 4.2 Latent Alignment

**Teacher.** We align the posterior mean to features of a frozen pretrained audio model. We consider two teachers. WavLM-Large (Chen et al., 2022) is the self-supervised speech encoder used by prior alignment work (Niu et al., 2025; Cheng et al., 2026). WavLM-Aesthetics is the audio encoder of the Audiobox Aesthetics quality predictor (Tjandra et al., 2025b): a 12-layer, 768-dimensional WavLM backbone fine-tuned on human quality ratings of speech, music, and sounds. As alignment target we take its softmax-weighted sum of layer outputs with the predictor's own weights, so that the target combines layers according to a perceptual-quality objective rather than a single recognition layer. Teacher audio is resampled to 16 kHz and, for stereo models, processed independently for the left and right channels. Both channel-specific targets supervise the same joint latent, and the loss is averaged over channels. Teacher frames are linearly interpolated to the 50 Hz latent timeline.

**Projections.** The teacher dimensionality $D_F$ exceeds $d_z$, so one side must be projected. We project the teacher features down to the latent, as the public implementation of VA-VAE does (Yao et al., 2025), because in our experiments this direction gave better reconstruction and diffusability than projecting the latent up to the teacher dimension. Using a row-vector convention, let $F \in \mathbb{R}^{T_F \times D_F}$ contain $T_F$ feature vectors. The conventional choice, used by our *Lin* runs, is a trainable linear layer projection,

$$F' = FW^{\mathsf{T}}, \qquad W \in \mathbb{R}^{d_z \times D_F}, \tag{1}$$

learned jointly with the VAE.

Our alternative is a principal-component basis of the teacher features. We fit incremental PCA (Ross et al., 2008) on every training batch, treating each teacher time frame as one sample, so that the basis

tracks the feature distribution of the training data without a separate preprocessing pass, and retain the $d_z$ directions of largest variance:

$$F' = (F - \boldsymbol{\mu}_F)\, U_{d_z}, \qquad U_{d_z} \in \mathbb{R}^{D_F \times d_z}, \tag{2}$$

where $\boldsymbol{\mu}_F$ is the running feature mean. No gradient flows through $U_{d_z}$. The difference between eqs. (1) and (2) is not the linearity of the map—both are linear—but who chooses it. In eq. (1) the VAE objective does, and the map can adapt to whatever the latent already encodes. In eq. (2) the teacher's own statistics do, and the latent has to adapt to them. Section 6 shows that this matters.

**AF loss.** We adapt the foundation-model alignment objective of Yao et al. (2025). After feature normalization, it combines a marginal cosine-similarity term, which penalizes each latent frame whose cosine similarity to its projected teacher frame falls below $1 - m_1$, with a marginal distance-matrix term, which penalizes each pair of frames within a clip whose latent cosine similarity differs from the corresponding teacher cosine similarity by more than $m_2$. Both terms act on $\boldsymbol{\mu}$ only, not on $\boldsymbol{\sigma}$ or the sampled $\mathbf{z}$. By analogy with the vision foundation (VF) loss of VA-VAE, we call the sum of the two terms the audio foundation (AF) loss and denote it $\mathcal{L}_{\mathrm{AF}}$. Section A.4 gives the equations, and section H the margins and internal weights.

### 4.3 Reconstruction, Perceptual, and Adversarial Objectives

**Spectral reconstruction.** Following DAC (Kumar et al., 2023) we use an $\ell_1$ multi-scale log-mel loss $\mathcal{L}_{\mathrm{mel}}$ over seven window sizes from 32 to 2048 samples. For stereo we evaluate it in two channel bases, left/right and mid/side ($M = (L + R)/2$, $S = (L - R)/2$), and sum the resulting losses. The mid/side basis exposes errors in the stereo image that per-channel losses average out (Parker et al., 2026).

**Phase-aware reconstruction.** Magnitude losses discard phase, yet phase coherence matters for transients and for stereo imaging. We add the phase objective of SAME (Parker et al., 2026), $\mathcal{L}_{\mathrm{IFGD}}$, which compares instantaneous-frequency and group-delay increments through normalized complex phasors, without phase unwrapping, plus a self-normalized complex-spectrogram distance, averaged over the same seven STFT resolutions and applied to left and right.

**Perceptual loss.** Following Parker et al. (2025), $\mathcal{L}_{\mathrm{perc}}$ matches the intermediate features of all layers of a frozen WavLM-Aesthetics backbone between input and reconstruction, each layer normalized by the mean magnitude of its target features. Left and right are passed as separate examples, and the loss is not computed in the mid/side basis. In preliminary experiments this term improved the subjective quality of *generated* audio, and we include it in all models.

**Adversarial and feature-matching losses.** Two discriminator branches follow the DAC configuration of multi-period waveform and multi-resolution complex-spectrogram discriminators with least-squares objectives (Mao et al., 2017) and feature matching (Kumar et al., 2019). The *channel-wise* branch treats left and right as separate mono examples with shared parameters. In the *joint-stereo* branch, the waveform discriminators receive the left and right channels as two input channels, whereas the complex-spectrogram discriminators receive the real and imaginary components of both channels as four input feature maps. This allows both types of discriminators to model inter-channel dependencies that are unavailable to the channel-wise branch. We write the generator adversarial and feature-matching terms as $\mathcal{L}_G, \mathcal{L}_{\mathrm{feat}}$ and, for the joint-stereo branch, $\mathcal{L}_G^S, \mathcal{L}_{\mathrm{feat}}^S$.

**KL regularization.** $\mathcal{L}_{\mathrm{KL}}$ is the divergence of the diagonal Gaussian posterior from the standard normal prior, normalized per latent element.

**Total objective.** The final stereo model is trained with the objective

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\mathrm{mel}}\,\mathcal{L}_{\mathrm{mel}} + \lambda_{\mathrm{feat}}\big(\mathcal{L}_{\mathrm{feat}} + \mathcal{L}_{\mathrm{feat}}^S\big) + \lambda_G\big(\mathcal{L}_G + \mathcal{L}_G^S\big) \\ & + \lambda_{\mathrm{perc}}\,\mathcal{L}_{\mathrm{perc}} + \lambda_{\mathrm{KL}}\,\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{AF}}\,\mathcal{L}_{\mathrm{AF}} + \lambda_{\mathrm{IFGD}}\,\mathcal{L}_{\mathrm{IFGD}},\end{aligned} \tag{3}$$

with the weights listed in section H. The mono models of section 6 omit the stereo-specific terms, namely the mid/side part of $\mathcal{L}_{\mathrm{mel}}$, the joint-stereo discriminator branch, and the phase loss $\mathcal{L}_{\mathrm{IFGD}}$, and keep the other terms and weights. Waveform $\ell_1$ and multi-scale STFT distances are monitoring metrics, not additional terms in this objective. Section A gives all definitions.

# 5 EXPERIMENTAL SETUP

## 5.1 DATA

All models are trained on a corpus of approximately $310{,}000$ hours of audio in four domains: speech, music, sounds, and general audio (Table 15). General audio denotes recordings that mix the other domains arbitrarily, with speech, music, sounds, effects, and ambient backgrounds within one clip. Genuinely full-band stereo material is scarce. Many nominally $48\,\mathrm{kHz}$ files are upsampled low-rate recordings, and many two-channel files are duplicated mono, so neither the container sample rate nor the channel count is informative. We therefore measure bandwidth and stereo content from the signal, and the corpus statistics are based on these measurements. Collection, filtering, and sampling are described in section G. To keep the ablation budget manageable, the mono $64$-channel comparisons of section 6 use a subset roughly $30\times$ smaller than the full corpus. The models of sections 3, 7 and 8, including the mono-64 model against which the stereo model is compared, use the full corpus, so that the stereo-versus-mono comparison rests on identical data. Everything else in the recipe is shared.

Reconstruction metrics are reported on a test set of music, speech, and sound effects assembled from open datasets, filtered by stereo status and true sample rate, and disjoint from the training data (about 13 hours in total).

Objective generation metrics are evaluated on AudioCaps test (Kim et al., 2019), Song Describer (Manco et al., 2023), and Seed-TTS Eval (Anastassiou et al., 2024), with the prompt selection and durations given in section B.1.

## 5.2 TRAINING, GENERATION, AND EVALUATION PROTOCOL

**VAE training.** Every VAE is trained with the objective of eq. (3) for $400{,}000$ steps on NVIDIA H100 GPUs on $0.38\,\mathrm{s}$ crops ($19$ latent frames) with a batch size of $128$. Section H lists all hyperparameters.

**Reconstruction metrics.** We report multi-scale mel (MEL) and STFT distances, MFCC distance, waveform $\ell_1$, SI-SDR, SNR, and, for speech, PESQ, on the test set of section 5.1. For stereo models we add cross-channel phase coherence (CCPC) (Wang et al., 2025), the energy-weighted phasor coherence between the inter-channel phase differences of reference and reconstruction, averaged over four STFT resolutions. It measures the stereo image directly, which per-channel metrics do not.

**Generation.** For generation experiments we use an open-source $0.6$-billion-parameter diffusion transformer (Arkhipkin et al., 2025) adapted for audio. For every VAE we train the generator from scratch on an internal general-audio dataset under the same protocol for $60{,}000$ steps on the latents of that VAE, and nothing else changes between comparisons.

**Human preference.** We conduct blind, randomized pairwise listening tests on a fixed pool of $148$ prompts spanning speech, music, and sounds. Listeners judge prompt following and audio quality separately and may mark both outputs as good or both as bad. Win rates give half credit for "both good" and exclude "both bad". Section B gives the protocol and the win-rate definition.

**Objective generation metrics.** Alongside the listening tests, we report six objective generation metrics: LAION CLAP for text–audio similarity, FAD (VGG) and FD (OPEN-L3) for distribution distances, the Content Enjoyment (CE) and Production Quality (PQ) scores of Audiobox Aesthetics, and, for speech prompts, WER and UTMOS. Section B.1 lists the implementations and protocols.

**Linear probes and latent diagnostics.** To measure what a representation retains, we train linear heads on frozen features for speech recognition (CTC on Mini LibriSpeech) and for sound, instrument, and animal-vocalization classification, sweeping the learning rate and repeating each configuration with $10$ seeds (section C). To measure the temporal organization of a representation we adapt the self-similarity diagnostics of iREPA (Singh et al., 2025) from the image lattice to the $50\,\mathrm{Hz}$ timeline: correlation-decay slope (CDS), local–distant similarity gap (LDS), and root-mean-square contrast (RMSC), all computed on the posterior mean of held-out clips (section E).

# 6 IMPROVING DIFFUSABILITY

This section compares four mono 64-channel models trained on the reduced corpus (section 5.1). We evaluate two projectors—the trainable linear layer projection of eq. (1) and the principal-component projection of eq. (2)—with two alignment teachers, WavLM-Large and WavLM-Aesthetics. For each teacher, the models use the same architecture, data, objectives, and training setup, and only the projection method differs, which makes the comparison controlled.

## 6.1 WHY THE PROJECTION MATTERS

**What the projections do to the teacher.** Before training any generator, we can ask how much of the teacher's information each projection keeps. Figures 4 and 5 probe the raw teacher features, their PCA projection, and their trainable projection (taken from the corresponding trained VAE) with linear heads. For both teachers the PCA projection retains most of the raw features' downstream performance, whereas the trainable projection loses a large part of it: the best character error rate of the trainably projected WavLM-Large features is several times that of the raw features, and classification accuracies drop by 10–30 points. The same picture appears in the temporal diagnostics of the projected teacher features (Table 12): the PCA projection keeps CDS within 4–13% of the raw features for both teachers, whereas the trainable projection loses 32–44%, with the same ordering in LDS. A projection learned jointly with the VAE is free to discard whatever the alignment loss does not need, whereas a variance-preserving basis fixed by the teacher's statistics is not.

**What the principal components encode.** A speaker–phone subspace analysis of both teachers (section C.2) shows that the leading principal components are dominated by phone-related directions, with WavLM-Aesthetics additionally retaining prominent speaker-dependent acoustic variation.

**What the projections do to the latent.** The loss of information does not stay in the projection. It propagates to the VAE. Figure 6 probes the frozen encoders of the trained VAEs: PCA-aligned encoders yield lower character error rates and higher classification accuracies than their trainable-projection counterparts for both teachers. Table 11 in section D.5 shows the same ordering in the temporal diagnostics: CDS of the PCA-aligned latents is 0.017–0.033 against 0.002–0.003 for the trainable projection with WavLM-Large, and 0.015–0.035 against 0.006–0.011 with WavLM-Aesthetics, with the same ordering in LDS. In other words, with a trainable projection the AF loss is satisfied partly by moving the projection rather than the latent, and less of the teacher's structure reaches the latent.

**Reconstruction versus generation.** Table 1 compares reconstruction, while fig. 1 (top and middle) and Tables 2 and 3 compare generation for the projection choice under two alignment teachers. For each teacher, we hold the alignment target fixed and compare the trainable linear layer and PCA projectors. We report objective generation metrics on AudioCaps test because it covers the broadest range of general-audio sounds among our evaluation sets. The trainable projection reconstructs better on almost every metric and domain. The PCA projection nevertheless wins the listening test clearly: $77/23$ on prompt following and $74/26$ on audio quality with WavLM-Large, and $60/40$ and $58/42$ with WavLM-Aesthetics. Reconstruction quality is thus an unreliable guide to the quality of a latent for generation, and the latent that a downstream diffusion model prefers is the one whose structure the linear probes and temporal diagnostics identify as closer to the teacher's.

## 6.2 CHOOSING THE TEACHER

WavLM-Large is trained on speech, and its representations are correspondingly weaker for environmental sounds and music. WavLM-Aesthetics is trained to predict human quality ratings of speech, music, and sounds alike, which makes it a natural teacher for general audio. Replacing WavLM-Large with WavLM-Aesthetics as the alignment teacher, with the PCA projection in both cases, improves nearly every reconstruction metric in every domain (Table 1) and every objective generation metric on AudioCaps (Tables 2 and 3), while the listening test between the two teachers is a tie at $52/48$ on prompt following and $49/51$ on audio quality (fig. 1, bottom). The teacher is also cheaper: the Aesthetics backbone has 12 layers of width 768 against 24 layers of width 1024 for WavLM-Large, so its features cost less to compute at every training step.

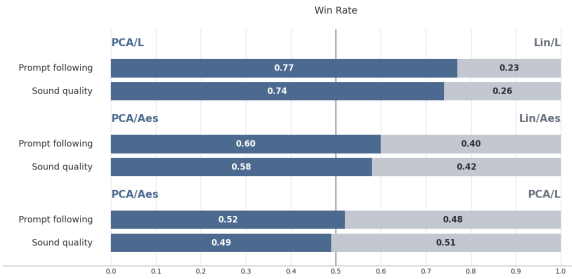


Figure 1: Pairwise human-preference win rates for generators trained on the mono latents, pooled over sounds, speech, and music. Top: PCA versus linear layer projection with WavLM-Large. Middle: the same with WavLM-Aesthetics. Bottom: WavLM-Aesthetics versus WavLM-Large with PCA projection.

| *WavLM-Large alignment teacher* | | | | |
|---|---|---|---|---|
| | Lin | | PCA | |
| Subset | MEL↓ | SI-SDR↑ | MEL↓ | SI-SDR↑ |
| Sounds | **0.6898** | **0.818** | 0.8994 | -1.725 |
| Music | **0.6766** | **11.767** | 0.8462 | 9.897 |
| Speech | **0.4997** | **9.645** | 0.6495 | 7.663 |
| *WavLM-Aesthetics alignment teacher* | | | | |
| | Lin | | PCA | |
| Subset | MEL↓ | SI-SDR↑ | MEL↓ | SI-SDR↑ |
| Sounds | **0.6529** | **1.600** | 0.7591 | 1.168 |
| Music | **0.6755** | **11.777** | 0.7414 | 11.250 |
| Speech | **0.5048** | **9.759** | 0.5481 | 9.077 |

Table 1: Reconstruction of the mono 64-channel models, grouped by alignment teacher. The full metric set is in Table 10. **Bold**: best within each projector pair.

| Metric | Lin | PCA |
|---|---|---|
| LAION CLAP↑ | 0.21 | **0.26** |
| CE↑ | 3.61 | **4.09** |
| PQ↑ | 5.97 | **6.28** |
| FD (OPEN-L3)↓ | **217.58** | 222.32 |
| FAD (VGG)↓ | 6.53 | **4.98** |

Table 2: Objective generation metrics on AudioCaps test for the mono 64-channel models aligned to WavLM-Large. **Bold**: best between the two models.

| Metric | Lin | PCA |
|---|---|---|
| LAION CLAP↑ | 0.27 | **0.29** |
| CE↑ | 4.18 | **4.21** |
| PQ↑ | 6.37 | **6.42** |
| FD (OPEN-L3)↓ | 211.98 | **208.11** |
| FAD (VGG)↓ | 4.03 | **3.69** |

Table 3: Objective generation metrics on AudioCaps test for the mono 64-channel models aligned to WavLM-Aesthetics. **Bold**: best between the two models.

| | Stereo-128 | |
|---|---|---|
| Metric | PCA | Lin |
| LAION CLAP↑ | **0.28** | 0.26 |
| CE↑ | **4.20** | 4.09 |
| PQ↑ | **6.43** | 6.28 |
| FD (OPEN-L3)↓ | 217.81 | **217.00** |
| FAD (VGG)↓ | **3.48** | 5.99 |

Table 4: Objective generation metrics on AudioCaps test for the stereo 128-channel models with PCA and linear layer projection (section 7). All evaluation sets are in Table 8. **Bold**: best.

**Summary.** On mono 64-channel models, a principal-component projection and a perceptual-quality teacher together give a latent that reconstructs slightly worse than the conventional choice and generates clearly better. Section 7 uses this recipe to revisit the channel budget of section 3.

## 7 Scaling to Stereo

The stereo 128-channel model is trained on the full corpus with the recipe of section 6 (PCA projection, WavLM-Aesthetics teacher) and the stereo objectives of section 4.3.

**The projection result carries over.** Figure 2 and Table 4 repeat the projection comparison at 128 channels and stereo. PCA wins 68/32 on prompt following and 66/34 on audio quality. The gap is larger than in mono (60/40 and 58/42 with the same teacher), consistent with the wider latent leaving more room for a trainable projection to absorb the AF loss.

**Stereo objectives.** Table 5 isolates the stereo-specific terms of eq. (3) on the final model: the mid/side mel basis, the joint-stereo discriminator, and the phase-aware loss. Mid/side supervision improves MEL and CCPC, while joint-stereo discrimination further improves SI-SDR and CCPC. Adding the phase-aware loss gives the best pooled result on all four reported metrics.

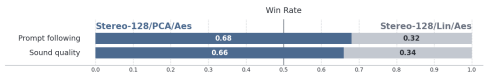


Figure 2: Pairwise human-preference win rates for the stereo 128-channel model, pooled over sounds, speech, and music: PCA versus linear layer projection.

| Configuration | MEL↓ | STFT↓ | SI-SDR↑ | CCPC↑ |
|---|---|---|---|---|
| L/R mel only | 0.6487 | 2.1848 | 7.685 | 0.9246 |
| + mid/side mel | 0.5865 | 2.1469 | 7.322 | 0.9264 |
| + joint-stereo discriminator | 0.5965 | 2.1541 | 7.904 | 0.9405 |
| + phase loss $\mathcal{L}_{\mathrm{IFGD}}$ (final) | **0.5755** | **2.1414** | **8.722** | **0.9443** |

Table 5: Ablation of the stereo objectives on the stereo 128-channel PCA model (reconstruction on the stereo test set, all domains pooled). Each row adds one term to the previous one. **Bold**: best.

Finally, with the PCA projection the stereo 128-channel model reaches parity with the mono 64-channel model of the same recipe in a side-by-side test on mono downmixes, on both prompt following and audio quality. The channel budget that section 3 could not spend is thus spent on the second channel without a penalty in generation.

## 8 Comparison with Open Audio VAEs

To our knowledge, no open audio VAE is at the same time stereo, full-band, and trained for all audio domains. We therefore compare the final model with the closest and most widely used open alternatives, SAME-L (Parker et al., 2026) and the audio VAE of LTX 2.5 (HaCohen et al., 2026), described in section 2, with their interfaces listed in section D.3. For every VAE we train the same 0.6B diffusion transformer from scratch for the same number of steps on its latents (section 5), so the comparison isolates the latent space from the generator.

Figure 3 and Table 6 report the listening tests and the objective generation metrics. Unlike the pooled comparisons of sections 6 and 7, we break the comparison with the open VAEs down by category, since this is where the domain coverage of the models differs most. Our model is preferred over both VAEs in every category on both criteria. Against SAME-L it wins $55/68/64\%$ of the prompt-following and $71/98/93\%$ of the audio-quality comparisons for sounds, speech, and music. Against the LTX 2.5 audio VAE it wins $74/60/67\%$ and $82/78/94\%$. On audio quality it thus wins more than $70\%$ of the comparisons against both VAEs in every category. The objective metrics agree on AudioCaps and Seed-TTS Eval, where our model is best on nearly every metric, while on Song Describer SAME-L and LTX 2.5 score higher on CLAP and FAD (VGG).

Table 13 in section D.7 compares reconstruction on our test set with the two VAEs above and with $\varepsilon$ar-VAE (Wang et al., 2025), a further open stereo autoencoder. Since the models represent different bandwidths ($44.1\,\mathrm{kHz}$ for SAME-L and $\varepsilon$ar-VAE, $48\,\mathrm{kHz}$ for the vocoder output of LTX 2.5 and for ours), each competitor is evaluated at its native sample rate, with the reference and both reconstructions downsampled to that rate. Our model is best on nearly every metric in both rate groups. SAME-L and $\varepsilon$ar-VAE are ahead only on a few sounds and music metrics (STFT, waveform $\ell_1$, and SNR for sounds, STFT and MFCC for music), and LTX 2.5 on none. The waveform-level metrics of LTX 2.5 (waveform $\ell_1$, SI-SDR, and SNR) are far below those of the other models because its vocoder does not reproduce the phase of the input, and we list them for completeness rather than for direct comparison. Table 14 repeats the comparison on Song Describer and LibriSpeech, which are of lower quality and mostly mono but are used for evaluation in other work.

## 9 Discussion and Limitations

**Why a fixed basis beats a learned map.** The AF loss is a cosine objective on the latent and a cosine objective on its within-clip Gram matrix. With a trainable projection, both can be lowered in two ways: by moving the latent toward the teacher, which competes with reconstruction, or by moving the projection toward whatever the latent already encodes, which costs nothing. The probes and diagnostics of section 6 indicate that the optimizer takes the second route, and the near-unaligned temporal structure of the trainable-projection latents (Table 11) is what one would expect if it did. A principal-component basis closes that route. Projecting the latent up into the teacher space (Cheng et al., 2026) does not: in our experiments it was worse than the down-projection in both reconstruction and diffusability (section 4.2).

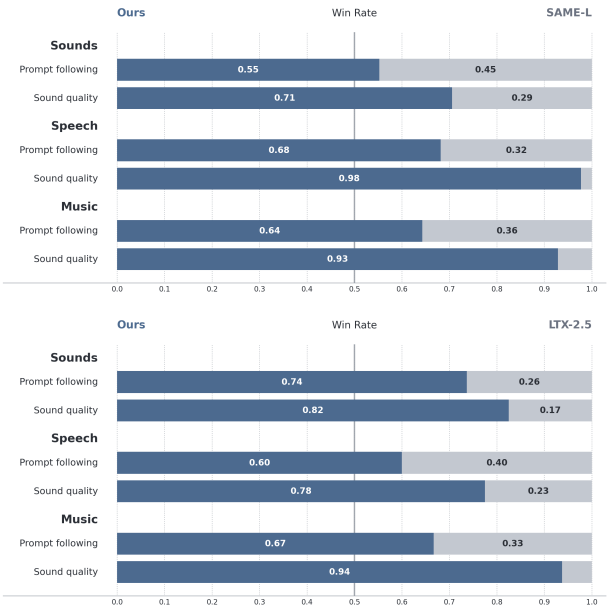


Figure 3: Pairwise human-preference win rates (%) of the final stereo model against SAME-L (top) and the LTX 2.5 audio VAE (bottom) under an identical generator and training budget.

| Metric | Ours | SAME-L | LTX 2.5 |
|---|---|---|---|
| *AudioCaps test* | | | |
| LAION CLAP↑ | **0.29** | 0.25 | <u>0.27</u> |
| CE↑ | **4.33** | 3.60 | <u>4.04</u> |
| PQ↑ | **6.52** | 5.84 | <u>6.36</u> |
| FD (OPEN-L3)↓ | **218.34** | 232.87 | <u>227.02</u> |
| FAD (VGG)↓ | **4.27** | 6.02 | <u>5.13</u> |
| *Song Describer* | | | |
| LAION CLAP↑ | 0.26 | **0.30** | <u>0.29</u> |
| CE↑ | **7.57** | 7.41 | <u>7.56</u> |
| PQ↑ | **8.14** | 7.77 | <u>8.10</u> |
| FD (OPEN-L3)↓ | **114.36** | 127.87 | <u>117.61</u> |
| FAD (VGG)↓ | 2.20 | <u>2.11</u> | **2.04** |
| *Seed-TTS Eval* | | | |
| LAION CLAP↑ | **0.17** | **0.17** | 0.15 |
| CE↑ | **5.93** | 4.20 | <u>5.42</u> |
| PQ↑ | **7.25** | 4.80 | <u>6.32</u> |
| FD (OPEN-L3)↓ | <u>123.29</u> | 145.90 | **119.39** |
| FAD (VGG)↓ | **4.51** | 7.19 | <u>5.80</u> |
| WER (%)↓ | **25.93** | 58.97 | <u>29.62</u> |

Table 6: Objective generation metrics of the generators trained on the three latents. **Bold**: best, <u>underline</u>: second best.

**What the diagnostics can and cannot tell.** Linear-probe accuracy and temporal self-similarity of a latent agree with the listening-test outcome on every pair we tested. CDS in particular separates clearly poor diffusability from good and lets us discard recipes without training a generator, and only when two latents are within about 20% of each other does a trained generator have to decide. Since any self-similarity measure is meaningless on silence or stationary noise, we compute it over the whole test set across domains and report it alongside, not instead of, reconstruction metrics and listening tests.

**Scope.** The teacher operates at 16 kHz and cannot directly supervise the upper octave of a 48 kHz output. That band is shaped by the spectral, phase, and adversarial terms only. A full-band teacher is not an automatic remedy: we also tried the penultimate layer of the LAION CLAP audio encoder (Wu et al., 2023), which operates at 48 kHz, as the alignment teacher, and it gave clearly worse diffusability, as the CDS of its features had predicted before any VAE was trained (0.016 against 0.066 for WavLM-Aesthetics on the sounds subset of our test set). The stereo objectives target channel relationships, but the listening tests are prompt-based and do not evaluate visual grounding or audiovisual synchronization, which require a video-conditioned generator and task-specific tests. Finally, all generation results are obtained with one generator architecture at one scale.

## 10 Conclusion

The obstacle to carrying full-band stereo directly through a waveform VAE for general audio was not reconstruction but diffusability: the channels needed for stereo made the latent harder to model. Two changes to a standard latent-alignment recipe removed it. A principal-component projection of the teacher features, fitted during training and free of gradients, transfers the teacher's structure to the latent instead of letting a trainable projection absorb the AF loss, and a perceptual-quality encoder serves speech, music, and sounds alike as the teacher. With diffusability restored, a 128-channel stereo latent matches a 64-channel mono latent in generation while additionally carrying the stereo image, and the resulting VAE is preferred over open stereo audio VAEs under an identical generator and budget.

### AI USE STATEMENT

In this work, we used generative AI tools for editing and restructuring the manuscript text, checking grammar and LaTeX, and searching the literature. We have not used generative AI tools for generating data, formulating or proving theoretical claims, designing the methodology or experiments, implementing the methods, or analyzing and interpreting the results. All AI-assisted text was reviewed and edited by the authors, and all citations were verified against the original sources. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

### ETHICS STATEMENT

Full-band audio generation can support creative production and accessibility, but downstream systems may also enable misleading soundtracks or voice impersonation. Deployment should address consent, provenance, and appropriate disclosure of synthetic media. Reproduction and release require verification of the licenses and privacy conditions of all audio sources and pretrained models. Aesthetic supervision may also reflect biases in its training ratings and should not be treated as a universal measure of quality.

### REPRODUCIBILITY STATEMENT

Sections 4 and 5 describe the architecture, objectives, data, and generation setup. Sections F and H give layer widths and all hyperparameters, and section A gives every loss in closed form. Sections B, C and E specify the listening-test protocol, the probe protocol, and the temporal diagnostics. We plan to release the code and the model weights as a separate open-source release after the review period. They are not part of this submission.

## REFERENCES

abdulvahap. Music instrument sounds for classification. Kaggle dataset, n.d. URL `https://www.kaggle.com/datasets/abdulvahap/music-instrunment-sounds-for-classification`. Accessed: 2026-09-02.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-TTS: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024. doi: 10.48550/arXiv.2406.02430.

Vladimir Arkhipkin, Vladimir Korviakov, Nikolai Gerasimenko, Denis Parkhomenko, Viacheslav Vasilev, Alexey Letunovskiy, Nikolai Vaulin, Maria Kovaleva, Ivan Kirillov, Lev Novitskiy, Denis Koposov, Nikita Kiselev, Alexander Varlamov, Dmitrii Mikhailov, Vladimir Polovnikov, Andrey Shutkin, Julia Agafonova, Ilya Vasiliev, Anastasiia Kargapoltseva, Anna Dmitrienko, Anastasia Maltseva, Anna Averchenkova, Olga Kim, Tatiana Nikulina, and Denis Dimitrov. Kandinsky 5.0: A family of foundation models for image and video generation. *arXiv preprint arXiv:2511.14993*, 2025. doi: 10.48550/arXiv.2511.14993.

ByteDance Speech. Seed-TTS-Eval: Objective evaluation code and test set. GitHub repository, 2024. URL `https://github.com/BytedanceSpeech/seed-tts-eval`. Accessed: 2026-09-25.

Antoine Caillon and Philippe Esling. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. In *International Conference on Learning Representations*, 2022.

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. doi: 10.1109/JSTSP.2022.3188113.

Changhao Cheng, Wei Wang, Wangyou Zhang, Dongya Jia, Jian Wu, Zhuo Chen, and Yanmin Qian. On the distillation loss functions of speech VAE for unified reconstruction, understanding, and generation. *arXiv preprint arXiv:2604.12383*, 2026. doi: 10.48550/arXiv.2604.12383.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander G. Schwing, and Yuki Mitsufuji. av-benchmark: Evaluation code for MMAudio. GitHub repository, 2025. URL https://github.com/hkchengrex/av-benchmark. Accessed: 2026-09-25.

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

Zach Evans, Cj Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12652–12665, 2024a.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*, 2024b. doi: 10.48550/arXiv.2404.10301.

Zach Evans, Julian D. Parker, Matthew Rice, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio 3. *arXiv preprint arXiv:2605.17991*, 2026. doi: 10.48550/arXiv.2605.17991.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022. doi: 10.1109/TASLP.2021.3133208.

Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376, 2006. doi: 10.1145/1143844.1143891.

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. *arXiv preprint arXiv:2601.03233*, 2026. doi: 10.48550/arXiv.2601.03233.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13916–13932, 2023.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. In *Interspeech*, 2021.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019. doi: 10.18653/v1/N19-1011.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2014. doi: 10.48550/arXiv.1312.6114.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, 2020.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. MelGAN: Generative adversarial networks for conditional waveform synthesis. *arXiv preprint arXiv:1910.06711*, 2019. doi: 10.48550/arXiv.1910.06711.

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*, volume 36, pp. 27980–27993, 2023. doi: 10.52202/075280-1214.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. BigVGAN: A universal neural vocoder with large-scale training. In *International Conference on Learning Representations*, 2023.

Lightricks Research. LTX-2.5 model card and audio VAE implementation. Hugging Face model card and official source repository, 2026. URL `https://github.com/Lightricks/LTX-2/tree/main/packages/ltx-core`.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474, 2023a.

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. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024. doi: 10.1109/TASLP.2024.3399607.

Oli Danyi Liu, Hao Tang, and Sharon Goldwater. Self-supervised predictive coding models encode speaker and phonetic information in orthogonal subspaces. In *Interspeech 2023*, pp. 2968–2972, 2023b. doi: 10.21437/Interspeech.2023-871.

Ilaria Manco, Benno Weck, SeungHeon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: A corpus of audio captions for music-and-language evaluation. *arXiv preprint arXiv:2311.10057*, 2023. doi: 10.48550/arXiv.2311.10057.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2794–2802, 2017. doi: 10.1109/ICCV.2017.304.

Zhikang Niu, Shujie Hu, Jeongsoo Choi, Yushen Chen, Peining Chen, Pengcheng Zhu, Yunting Yang, Bowen Zhang, Jian Zhao, Chunhui Wang, and Xie Chen. Semantic-VAE: Semantic-alignment latent representation for better speech synthesis. *arXiv preprint arXiv:2509.22167*, 2025. doi: 10.48550/arXiv.2509.22167.

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

Julian D. Parker, Anton Smirnov, Jordi Pons, CJ Carr, Zack Zukowski, Zach Evans, and Xubo Liu. Scaling transformers for low-bitrate high-quality speech coding. In *Proceedings of the International Conference on Learning Representations*, 2025.

Julian D. Parker, Zach Evans, CJ Carr, Zachary Zukowski, Josiah Taylor, Matthew Rice, and Jordi Pons. SAME: A semantically-aligned music autoencoder. *arXiv preprint arXiv:2605.18613*, 2026.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015. doi: 10.1145/2733373.2806390.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

David A. Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. In *International Journal of Computer Vision*, volume 77, pp. 125–141, 2008.

rushibalajiputthewad. Sound classification of animal voice. Kaggle dataset, n.d. URL https://www.kaggle.com/datasets/rushibalajiputthewad/sound-classification-of-animal-voice. Accessed: 2026-09-02.

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? *arXiv preprint arXiv:2512.10794*, 2025.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *International Conference on Machine Learning*, 2025. doi: 10.48550/arXiv.2502.14831.

Stability AI. Stable audio metrics. GitHub repository, 2024. URL https://github.com/Stability-AI/stable-audio-metrics. Accessed: 2026-09-25.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Audiobox aesthetics: Official inference code. GitHub repository, 2025a. URL https://github.com/facebookresearch/audiobox-aesthetics. Accessed: 2026-09-25.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic assessment for speech, music and sound. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2025b. doi: 10.1109/ASRU65441.2025.11434623.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Kangdi Wang, Zhiyue Wu, Dinghao Zhou, Rui Lin, Junyu Dai, and Tao Jiang. Back to ear: Perceptually driven high fidelity music reconstruction. *arXiv preprint arXiv:2509.14912*, 2025. doi: 10.48550/arXiv.2509.14912.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023. doi: 10.1109/ICASSP49357.2023.10095969.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025. doi: 10.48550/arXiv.2501.01423.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. *arXiv preprint arXiv:2408.17175*, 2024.

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*, 2025.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022.

Jun Zhan, Chen Yang, Yitian Gong, Donghua Yu, Kuangwei Chen, Wenbo Zhang, Kexin Huang, Qi Luo, Zhe Xu, Ying Zhu, Jin Wang, Tengyue Zhang, Qi Chen, Cheng Chang, Songlin Wang, Junqi Dai, Jiasheng Ye, Xiaogui Yang, Tianyi Liang, Xiangyu Peng, Zhaoye Fei, Shimin Li, Qinyuan Cheng, Xie Chen, Xinchi Chen, and Xipeng Qiu. OmniVAE: An audio-video VAE with cross-modal alignment for joint generation. *arXiv preprint arXiv:2607.23855*, 2026. doi: 10.48550/arXiv.2607.23855.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified speech tokenizer for speech language models. In *International Conference on Learning Representations*, 2024.

Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. In *Advances in Neural Information Processing Systems*, 2020.

# A Loss Function Definitions

## A.1 Reconstruction loss

Following the Descript Audio Codec study (Kumar et al., 2023), we use a multi-scale mel-spectrogram reconstruction loss based on the $\ell_1$ distance between the mel-spectrograms of the input and reconstructed audio. Given an input waveform $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$, the loss is defined as

$$\mathcal{L}_{\text{mel}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{1}{N_s} \left\| \mathcal{M}_s(\mathbf{x}) - \mathcal{M}_s(\hat{\mathbf{x}}) \right\|_1 , \tag{4}$$

where $\mathcal{S}$ is the set of analysis scales, $\mathcal{M}_s$ denotes the mel-spectrogram transform at scale $s$, and $N_s$ is the number of elements in the corresponding mel-spectrogram.

## A.2 Discriminator-based loss

For adversarial training, we adopt the discriminator configuration used in the Descript Audio Codec study (Kumar et al., 2023), combining a multi-period discriminator with a multi-scale short-time Fourier transform discriminator. Let $D_k$ denote the $k$-th discriminator, let $K$ be the total number of discriminators, and let $\mathbf{x}$ and $\hat{\mathbf{x}}$ denote real and reconstructed audio, respectively. Using the least-squares generative adversarial network objective (Mao et al., 2017), the discriminator loss is

$$\begin{aligned} \mathcal{L}_D = \frac{1}{K} \sum_{k=1}^{K} \big[ & \mathbb{E}_{\mathbf{x}} (D_k(\mathbf{x}) - 1)^2 \\ & + \mathbb{E}_{\hat{\mathbf{x}}} D_k(\hat{\mathbf{x}})^2 \big], \end{aligned} \tag{5}$$

while the adversarial loss used to train the decoder is

$$\mathcal{L}_G = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\hat{\mathbf{x}}} \left( D_k(\hat{\mathbf{x}}) - 1 \right)^2 . \tag{6}$$

We additionally apply a feature-matching loss (Kumar et al., 2019) to the intermediate discriminator activations:

$$\mathcal{L}_{\text{feat}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{L_k} \sum_{l=1}^{L_k} \frac{1}{N_{k,l}} \left\| D_k^l(\mathbf{x}) - D_k^l(\hat{\mathbf{x}}) \right\|_1 , \tag{7}$$

where $D_k^l$ is the activation of layer $l$, $L_k$ is the number of selected layers, and $N_{k,l}$ is the number of elements in the corresponding feature map. Both terms enter eq. (3) with weights $\lambda_G$ and $\lambda_{\text{feat}}$.

## A.3 Kullback–Leibler divergence

We use the Kullback–Leibler divergence as a regularization term that encourages the approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$ to follow the standard normal prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. For a diagonal Gaussian posterior with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$, the loss is

$$\begin{aligned} \mathcal{L}_{\text{KL}} &= \frac{1}{TD} D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})) \\ &= \frac{1}{2TD} \sum_{t=1}^{T} \sum_{d=1}^{D} \left( \mu_{t,d}^2 + \sigma_{t,d}^2 - \log \sigma_{t,d}^2 - 1 \right), \end{aligned} \tag{8}$$

where $T$ is the number of latent time steps and $D$ is the latent dimensionality. The normalization reports KL divergence per latent element. Matching the prior does not by itself guarantee semantic organization.

## A.4 Alignment loss

We adapt the foundation-model alignment objective of Yao et al. (2025) to a frozen pretrained audio teacher. Alignment is applied exclusively to the posterior mean $\boldsymbol{\mu}$, not to the standard deviation $\boldsymbol{\sigma}$ or

sampled latent $\mathbf{z}$. The projections in Section 4.2 map teacher features to the latent dimensionality. The objective combines a marginal cosine similarity loss,

$$\mathcal{L}_{\text{mcos}} = \frac{1}{N}\sum_{i=1}^{N} \text{ReLU}\left(1 - m_1 - \frac{\mu_i \cdot f_i'}{\|\mu_i\|\|f_i'\|}\right), \tag{9}$$

and a marginal distance-matrix similarity loss,

$$\mathcal{L}_{\text{mdms}} = \frac{1}{N^2}\sum_{i,j} \text{ReLU}\big(\big|\cos(\mu_i, \mu_j) - \cos(f_i', f_j')\big| - m_2\big). \tag{10}$$

Here, $N$ is the number of matched vector pairs used to evaluate the loss, and $\mu_i$ and $f_i'$ denote a posterior-mean vector and its corresponding projected teacher feature. The first term encourages pointwise agreement, and the second encourages agreement between pairwise cosine similarities. The margins $m_1$ and $m_2$ allow discrepancies within a tolerance. The resulting alignment loss is

$$\mathcal{L}_{\text{AF}} = \lambda_{\text{mcos}}\mathcal{L}_{\text{mcos}} + \lambda_{\text{mdms}}\mathcal{L}_{\text{mdms}}, \tag{11}$$

where $\lambda_{\text{mcos}}$ and $\lambda_{\text{mdms}}$ control the contributions of the two alignment terms. We use $\lambda_{\text{mcos}} = \lambda_{\text{mdms}} = 1$, $m_1 = 0.5$, and $m_2 = 0.25$, as in Yao et al. (2025). $\mathcal{L}_{\text{AF}}$ enters eq. (3) with the weight $\lambda_{\text{AF}}$ of section H for both teachers. For stereo models the loss is evaluated against each channel's projected teacher sequence and averaged over channels. Pairwise comparisons are between frames within a clip, not across examples.

### A.5 Perceptual loss

Following Parker et al. (2025), we compute a perceptual loss using a frozen pretrained audio model. Let $C^m(\cdot)$ denote the features extracted from its $m$-th layer, and let $M$ be the number of layers. The loss compares the features of the input waveform $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ across all layers:

$$\mathcal{L}_{\text{perc}}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{M}\sum_{m=1}^{M} \frac{\|C^m(\mathbf{x}) - C^m(\hat{\mathbf{x}})\|_1}{\text{mean}(\|C^m(\mathbf{x})\|_1)}. \tag{12}$$

The denominator normalizes each layer by the mean magnitude of the target features across the batch. This objective supplements spectrogram reconstruction with agreement in pretrained feature space.

### A.6 Stereo-specific terms

#### A.6.1 Multi-scale mel reconstruction

The seven analysis-window lengths are $\{32, 64, 128, 256, 512, 1024, 2048\}$ samples, with corresponding mel-bin counts $\{5, 10, 20, 40, 80, 160, 320\}$. Each scale compares log-mel magnitudes with an $\ell_1$ distance. Raw-magnitude weighting is zero in this term. For a stereo pair $x = (L, R)$, the total spectral reconstruction is the sum of the losses on flattened $(L, R)$ and flattened $((L + R)/2, (L - R)/2)$. This sum is weighted by 15 in eq. (3). The monitoring STFT and waveform distances do not alter the generator objective.

#### A.6.2 Phase-derivative and complex-distance terms

Following Parker et al. (2026), let $X$ and $Y$ denote complex STFTs of the reconstruction and reference at one resolution. For $Z \in \{X, Y\}$, define normalized products along time and frequency:

$$U_t^Z(f,t) = \frac{Z(f,t)\overline{Z(f,t-1)}}{|Z(f,t)|\,|Z(f,t-1)| + \epsilon}, \tag{13}$$

$$U_f^Z(f,t) = \frac{Z(f,t)\overline{Z(f-1,t)}}{|Z(f,t)|\,|Z(f-1,t)| + \epsilon}. \tag{14}$$

The phase-increment losses are

$$\mathcal{L}_{\text{IF}} = \mathbb{E}\left[w_t\left(1 - \text{Re}(U_t^X\overline{U_t^Y})\right)\right], \tag{15}$$

$$\mathcal{L}_{\text{GD}} = \mathbb{E}\left[w_f\left(1 - \text{Re}(U_f^X\overline{U_f^Y})\right)\right], \tag{16}$$

where $w_t$ and $w_f$ are detached, mean-normalized geometric means of adjacent magnitudes along the respective axis. A self-normalized complex-distance term is

$$\mathcal{L}_{\mathrm{cd}} = \mathbb{E} \log \left( 1 + \frac{|X - Y|^2}{\sigma_{\mathrm{sg}} + \epsilon} \right), \qquad \sigma_{\mathrm{sg}} = \mathrm{sg}\left( \mathrm{std}(|X - Y|^2) \right). \tag{17}$$

At each resolution, the three terms are summed. Their average across the seven FFT sizes used for mel reconstruction gives $\mathcal{L}_{\mathrm{IFGD}}$. The STFT uses hop $n_{\mathrm{fft}}/4$, no centering, and a power-sine tight window. There is no separate magnitude term in this phase objective. Left and right are evaluated as separate examples, without a mid/side duplicate, and the resulting loss enters eq. (3) with weight $\lambda_{\mathrm{IFGD}}$.

### A.6.3 Adversarial and feature-matching terms

Both discriminator branches use waveform periods $\{2, 3, 5, 7, 11\}$ and complex-spectrogram FFT sizes $\{2048, 1024, 512\}$ (and 1/4 hop sizes), with five frequency bands per resolution. The joint-stereo branch doubles the internal widths of the channel-wise branch, and its first spectrogram layer has width $64$ rather than 32. No additional multi-scale waveform or constant-Q discriminator is used.

Both branches use the least-squares targets and reference feature maps in section A.2. The generator adversarial and feature-matching sums are $\mathcal{L}_G$ and $\mathcal{L}_{\mathrm{feat}}$ of section A.2, with superscript $S$ for the joint-stereo branch. Generator adversarial losses have weight $\lambda_G$ and feature matching has weight $\lambda_{\mathrm{feat}}$ in each branch (section H). Discriminator parameters and detached reference features are held fixed during generator updates, while reconstruction features retain gradients to the generator.

## B Human Preference Evaluation

Each comparison is a blind, randomized side-by-side (SBS) test on the fixed pool of $148$ prompts of section 5.2. For every prompt, annotators listen to the two generations in random order and rate the pair separately on prompt following and audio quality, choosing either one of the two outputs or one of two tie responses, "both good" or "both bad." Let $n_A$ and $n_B$ count responses selecting only model $A$ or model $B$, respectively, and let $n_{\mathrm{gg}}$ count "both good" responses. The win rate for model $A$ is

$$w_A = \frac{n_A + 0.5\, n_{\mathrm{gg}}}{n_A + n_B + n_{\mathrm{gg}}}. \tag{18}$$

The rate for model $B$ is defined symmetrically. Thus, "both good" gives half credit to each model, while "both bad" is excluded from both the numerator and denominator. For a nonzero denominator, the two rates sum to one and are reported as percentages. These scores summarize relative preference among responses other than "both bad" and do not measure the absolute frequency of acceptable outputs. For the stereo-versus-mono test (section 7), stereo generations are downmixed to mono as $(L + R)/2$ before presentation, without loudness matching, since the target loudness is part of the prompt.

### B.1 Objective evaluation sets and metrics

Objective generation metrics are evaluated on three sets. For AudioCaps test (Kim et al., 2019), we use the 964 prompts selected by the MMAudio evaluation protocol (Cheng et al., 2025). Both references and generations are $10\,\mathrm{s}$ long (approximately $2.68$ hours for each corpus). For Song Describer (Manco et al., 2023), we use the same prompt table as Stable Audio 2 (Evans et al., 2024b), remove duplicate prompts by retaining their first occurrence in table order, and obtain 446 prompts. References and generations are again $10\,\mathrm{s}$ long (approximately $1.24$ hours for each corpus). For Seed-TTS Eval, we use the 1,088 speech prompts released with the original Seed-TTS work (Anastassiou et al., 2024; ByteDance Speech, 2024) and the corresponding references for those prompts. Each generation is matched to the duration of its reference, with durations from $4.33$ to $9.17\,\mathrm{s}$, totaling approximately $1.31$ hours for each corpus.

Alongside the listening tests, we report six objective generation metrics. LAION CLAP measures text–audio similarity, while FAD (VGG) compares generated and reference distributions in VGGish feature space. For both metrics, we use the implementation and evaluation protocol from the MMAudio

evaluation code (Cheng et al., 2025), downmixing stereo audio to mono before evaluation. CE and PQ denote the Content Enjoyment and Production Quality scores from Audiobox Aesthetics, respectively, computed with the authors' original implementation (Tjandra et al., 2025b;a). We compute FD (OPEN-L3) on stereo audio using the implementation and protocol from the Stable Audio 2/3 evaluation code (Stability AI, 2024). For speech prompts, WER is reported as a percentage using the Seed-TTS Eval implementation and protocol (ByteDance Speech, 2024), after downmixing stereo audio to mono.

# C Linear-Probe Protocol and Results

## C.1 Downstream probes of features

To assess information retained by the projections, we use linear probing: we train lightweight linear prediction heads on representations from a frozen pretrained encoder. We consider automatic speech recognition and audio classification. We train each head using the same set of learning rates, repeating each configuration with 10 different random seeds. In the corresponding plots, the curves show the mean score for each evaluation metric, and the shaded regions indicate 95% confidence intervals.

For automatic speech recognition, the linear head predicts token probabilities at each time step and is trained using the connectionist temporal classification objective (Graves et al., 2006). We conduct this experiment on the Mini LibriSpeech corpus (Panayotov et al., 2015), a compact subset of LibriSpeech intended for automatic speech recognition evaluation.

For audio classification, we use three datasets covering complementary acoustic domains. ESC-50 (Piczak, 2015) contains environmental sounds, the Music Instrument Sounds for Classification dataset (abdulvahap, n.d.) contains recordings of musical instruments, and the Sound Classification of Animal Voice dataset (rushibalajiputthewad, n.d.) contains vocalizations from different animal categories. For ESC-50, we use folds 1–4 for training and fold 5 for evaluation. To limit the size of the instrument and animal subsets, we randomly sample 40 recordings from each class. For these two datasets, the recordings in each class are randomly divided into training and test sets using a 3:1 ratio.

## C.2 Speaker–phone principal-component analysis

Following Liu et al. (2023b), we analyze how speaker and phone variation is distributed across the principal directions of each teacher. On Mini LibriSpeech, we average frame-level features separately by speaker, by aligned phone, and by each observed speaker–phone pair. This produces speaker, phone, and joint speaker–phone matrices, to which we apply PCA independently. We then compare phone and speaker directions with the joint directions using absolute cosine similarity. The heatmaps in fig. 7 show the correspondence between phone and joint directions. In fig. 8, joint directions identified by this comparison as phone- or speaker-aligned are overlaid on the variance spectrum, and directions not assigned to either group are gray.

For WavLM-Large, phone alignment is concentrated in the highest-variance joint components, whereas most speaker-aligned components occur farther down the spectrum and explain less variance. WavLM-Aesthetics has more speaker-aligned structure among its leading components, but its high-variance region remains dominated by phone-aligned directions. Thus, retaining the leading PCA directions preserves primarily phone-related structure for both teachers, while WavLM-Aesthetics appears to preserve more prominent speaker-dependent acoustic variation as well.

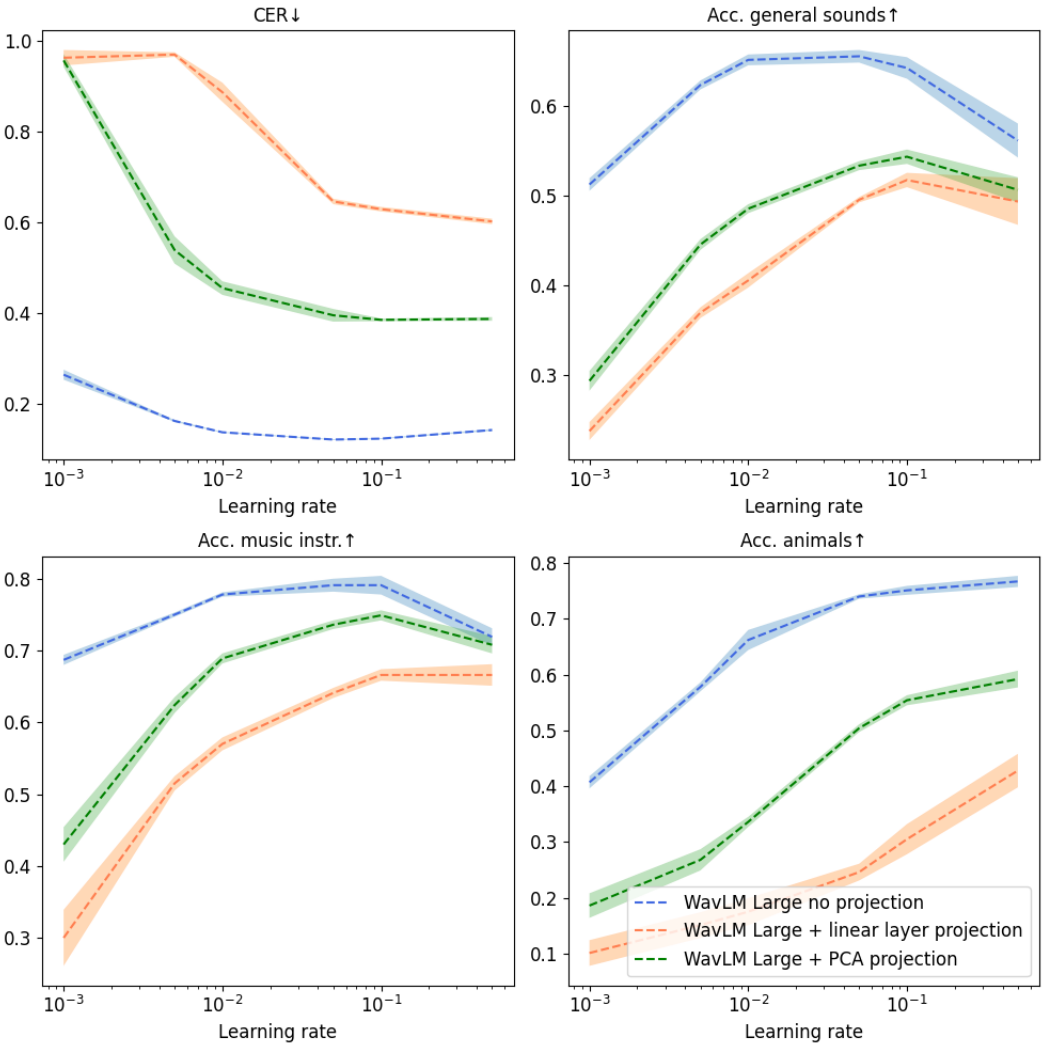


Figure 4: Downstream-task performance of WavLM-Large features without projection and after pretrained linear or PCA projections across different learning rates.

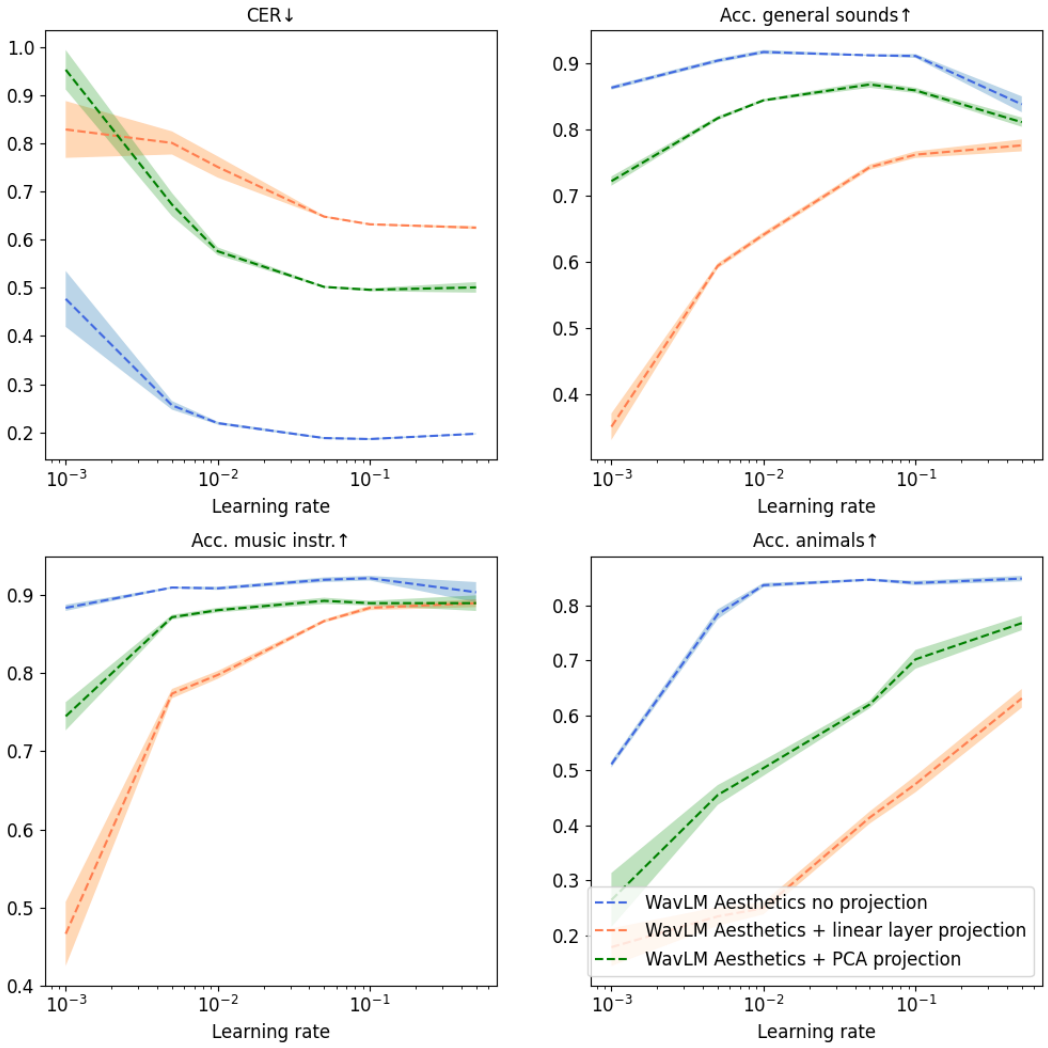


Figure 5: Downstream-task performance of WavLM-Aesthetics features without projection and after pretrained linear or PCA projections across different learning rates.

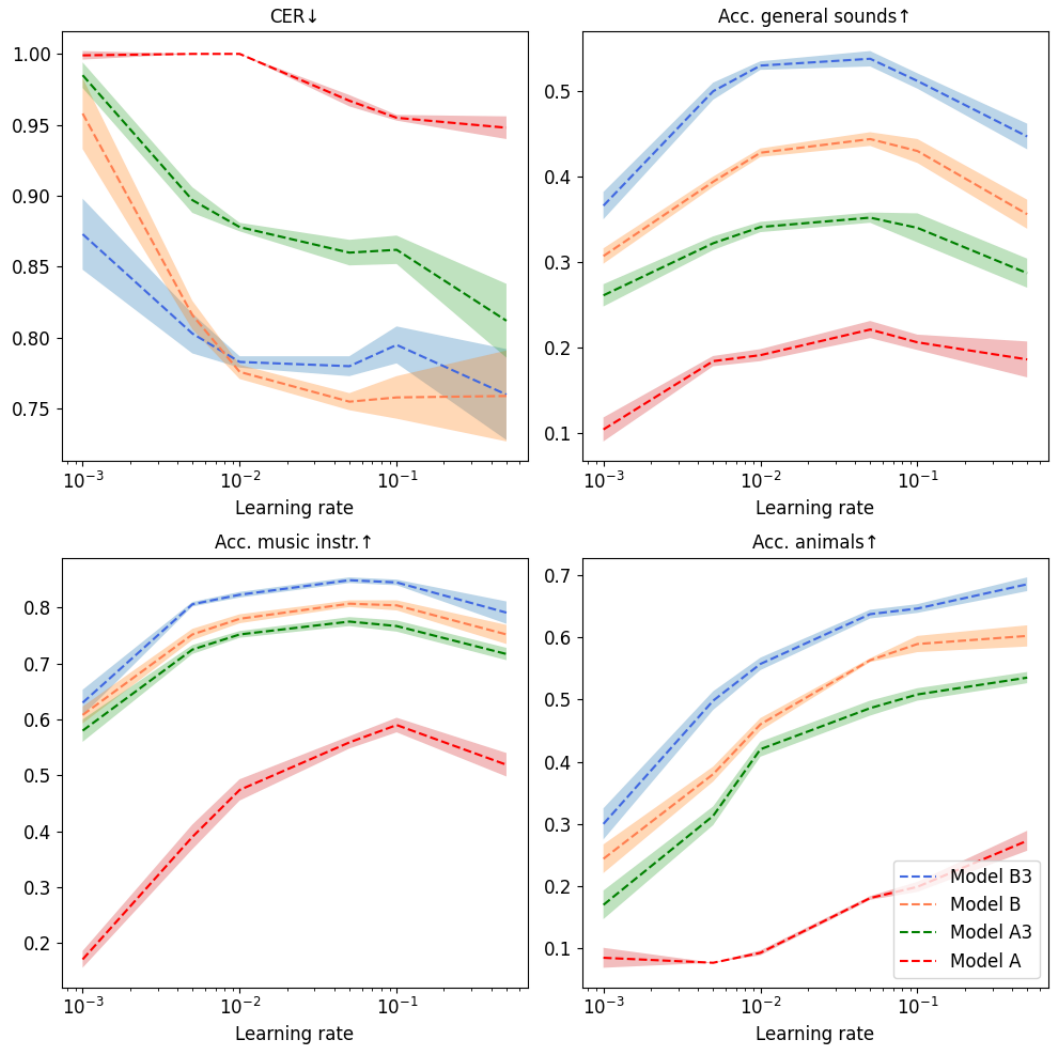


Figure 6: Linear probing of frozen VAE encoder representations.

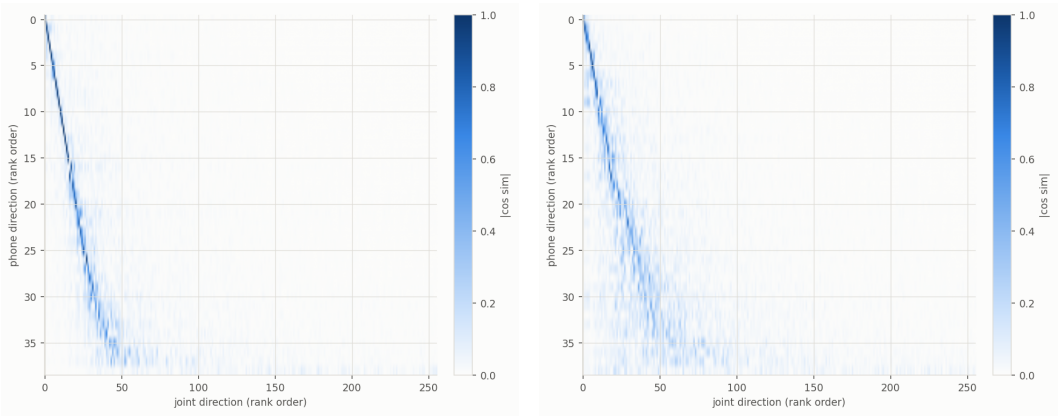


Figure 7: Absolute cosine similarity between phone principal directions and joint speaker–phone principal directions on Mini LibriSpeech. Left: WavLM-Large. Right: WavLM-Aesthetics. Both teachers place substantial phone-related structure in leading joint directions.

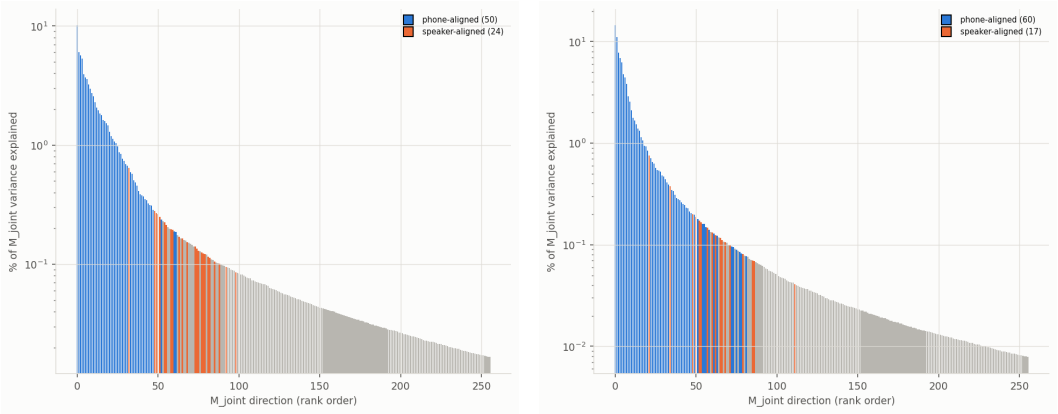


Figure 8: Variance explained by the joint speaker–phone principal directions, with phone-aligned directions in blue and speaker-aligned directions in orange. Left: WavLM-Large. Right: WavLM-Aesthetics. Phone-aligned directions dominate the high-variance region for both teachers, while speaker-aligned directions appear earlier for WavLM-Aesthetics.

# D Additional Tables

## D.1 Reconstruction of the problem-statement models

Table 7: Reconstruction on the stereo test set for the three configurations of the problem statement. The models differ only in the latent width and in mono versus stereo input and output.

| | Mono-64 | | | Stereo-64 | | | Stereo-128 | | |
|---|---|---|---|---|---|---|---|---|---|
| Subset | MEL↓ | STFT↓ | SI-SDR↑ | MEL↓ | STFT↓ | SI-SDR↑ | MEL↓ | STFT↓ | SI-SDR↑ |
| Sounds | 0.7591 | 2.7083 | 1.168 | 0.9291 | 2.8247 | -2.411 | 0.8277 | 2.8103 | -5.996 |
| Music | 0.7413 | 2.9303 | 11.250 | 0.8862 | 2.9894 | 5.974 | 0.7936 | 2.9818 | 8.606 |
| Speech | 0.5481 | 1.6684 | 9.077 | 1.6555 | 4.0507 | 9.339 | 0.441 | 1.400 | 10.679 |

### D.2 OBJECTIVE GENERATION METRICS OF THE STEREO MODELS ON ALL EVALUATION SETS

Table 8: Objective generation metrics of the stereo 128-channel models with PCA and linear layer projection on all three evaluation sets. **Bold**: best.

| | Stereo-128 | |
|---|---|---|
| Metric | PCA | Lin |
| *AudioCaps test* | | |
| LAION CLAP↑ | **0.28** | 0.26 |
| CE↑ | **4.20** | 4.09 |
| PQ↑ | **6.43** | 6.28 |
| FD (OPEN-L3)↓ | 217.81 | **217.00** |
| FAD (VGG)↓ | **3.48** | 5.99 |
| *Song Describer* | | |
| LAION CLAP↑ | **0.31** | 0.29 |
| CE↑ | **7.59** | 7.50 |
| PQ↑ | **8.12** | 8.04 |
| FD (OPEN-L3)↓ | **118.91** | 119.99 |
| FAD (VGG)↓ | 1.95 | **1.89** |
| *Seed-TTS Eval* | | |
| LAION CLAP↑ | **0.17** | 0.15 |
| CE↑ | 5.79 | **5.85** |
| PQ↑ | 7.00 | **7.48** |
| FD (OPEN-L3)↓ | **115.36** | 127.96 |
| FAD (VGG)↓ | **4.53** | 6.36 |
| WER (%)↓ | **16.95** | 44.91 |

### D.3 INTERFACES OF THE COMPARED AUDIO VAES

Table 9: Interfaces of the compared audio VAEs. Latent rate is frames per second of audio.

| Model | Sample rate | Channels | Latent rate | $d_z$ | Params | Representation |
|---|---|---|---|---|---|---|
| SAME-L (Parker et al., 2026) | 44.1 kHz | 2 | $\sim$ 10.8 Hz | 256 | 852M | waveform |
| LTX 2.5 audio VAE (HaCohen et al., 2026) | 48 kHz | 2 | 25 Hz | 128 | 182M | mel + vocoder |
| Ours | 48 kHz | 2 | 50 Hz | 128 | 167M | waveform |

### D.4 RECONSTRUCTION OF THE MONO MODELS

Table 10: Reconstruction of the mono 64-channel models of section 6 on the test set, grouped by alignment teacher and domain. **Bold**: best within each projector pair.

| | | *WavLM-Large* | | *WavLM-Aesthetics* | |
|---|---|---|---|---|---|
| Subset | Metric | Lin | PCA | Lin | PCA |
| Sounds | MEL↓ | **0.6898** | 0.8994 | **0.6529** | 0.7591 |
| | STFT↓ | **2.5350** | 2.9147 | **2.4514** | 2.7076 |
| | MFCC↓ | **0.1645** | 0.2894 | **0.1630** | 0.2505 |
| | Wave.↓ | **0.0508** | 0.0587 | **0.0475** | 0.0496 |
| | SI-SDR↑ | **0.818** | -1.725 | **1.600** | 1.168 |
| | SNR↑ | **3.312** | 1.843 | **3.810** | 3.514 |
| Music | MEL↓ | **0.6766** | 0.8462 | **0.6755** | 0.7414 |
| | STFT↓ | **2.7929** | 3.0399 | **2.7825** | 2.9304 |
| | MFCC↓ | **0.2447** | 0.3020 | **0.2435** | 0.2874 |
| | Wave.↓ | **0.0305** | 0.0361 | **0.0304** | 0.0320 |
| | SI-SDR↑ | **11.767** | 9.897 | **11.777** | 11.250 |
| | SNR↑ | **11.998** | 10.247 | **12.014** | 11.501 |
| Speech | MEL↓ | **0.4997** | 0.6495 | **0.5048** | 0.5481 |
| | STFT↓ | **1.6711** | 1.7573 | 1.6815 | **1.6684** |
| | MFCC↓ | **0.1130** | 0.1562 | **0.1131** | 0.1265 |
| | Wave.↓ | **0.0068** | 0.0084 | **0.0068** | 0.0072 |
| | SI-SDR↑ | **9.645** | 7.663 | **9.759** | 9.077 |
| | SNR↑ | **10.096** | 8.326 | **10.213** | 9.589 |
| | PESQ↑ | **4.194** | 3.571 | **4.292** | 4.187 |

### D.5 TEMPORAL DIAGNOSTICS OF THE VAE LATENTS

Table 11: Temporal self-similarity diagnostics of the VAE latents (posterior mean, held-out clips), grouped by alignment teacher. PCA-aligned latents show consistently stronger temporal structure than their trainable-projection counterparts. **Bold**: best within each projector pair.

| Domain | Model | CDS↑ | LDS↑ | RMSC↑ |
|---|---|---|---|---|
| *WavLM-Large alignment teacher* | | | | |
| Sounds | Lin | 0.0022 | 0.0358 | 0.3411 |
| | PCA | **0.0166** | **0.2052** | **0.5122** |
| Music | Lin | 0.0030 | 0.0415 | 0.4831 |
| | PCA | **0.0332** | **0.3149** | **0.7237** |
| Speech | Lin | 0.0032 | 0.0315 | 0.4371 |
| | PCA | **0.0297** | **0.2496** | **0.6530** |
| *WavLM-Aesthetics alignment teacher* | | | | |
| Sounds | Lin | 0.0061 | 0.1045 | **0.6335** |
| | PCA | **0.0151** | **0.1846** | 0.6315 |
| Music | Lin | 0.0082 | 0.1465 | 0.7985 |
| | PCA | **0.0351** | **0.3475** | **0.8641** |
| Speech | Lin | 0.0110 | 0.1032 | 0.7668 |
| | PCA | **0.0334** | **0.2693** | **0.8118** |

### D.6 TEMPORAL DIAGNOSTICS OF THE PROJECTED TEACHER FEATURES

Table 12: Temporal self-similarity diagnostics of the projected teacher features on held-out clips: raw features (identity), PCA projection, and the trainable linear layer projection taken from the corresponding trained VAE. Scores are unweighted means over sounds, music, and speech. **Bold**: best, underline: second best within each teacher.

| Teacher | Projection | CDS↑ | LDS↑ | RMSC↑ |
|---|---|---|---|---|
| WavLM-Aesthetics | raw | **0.0500** | **0.5122** | 0.8154 |
| | PCA | 0.0482 | 0.4947 | 0.7907 |
| | trainable | 0.0338 | 0.3860 | **0.8590** |
| WavLM-Large | raw | **0.0551** | **0.6065** | **0.8163** |
| | PCA | 0.0477 | 0.5548 | 0.7674 |
| | trainable | 0.0310 | 0.3981 | 0.7134 |

### D.7 Reconstruction against open audio VAEs

Table 13: Reconstruction on the stereo test set, grouped by the sample rate at which the metrics are computed. Within each group the reference and all reconstructions are downsampled to that rate. LTX 2.5 analyzes a 16 kHz mel representation and is evaluated on its 48 kHz vocoder output, so its metrics characterize the complete analysis–VAE–vocoder pipeline rather than direct full-band transmission through the latent. **Bold**: best in group.

| | | 44.1 kHz | | | 48 kHz | |
|---|---|---|---|---|---|---|
| Subset | Metric | SAME-L | $\varepsilon$ar-VAE | Ours | LTX 2.5 | Ours |
| Sounds | MEL↓ | 0.9617 | 0.8849 | **0.6779** | 1.4009 | **0.6817** |
| | STFT↓ | 2.3570 | **2.0898** | 2.3126 | 4.3718 | **2.3331** |
| | MFCC↓ | 0.2161 | 0.1942 | **0.1588** | 0.3899 | **0.1631** |
| | Wave.↓ | 0.0449 | **0.0440** | 0.0475 | 0.0888 | **0.0475** |
| | SI-SDR↑ | 2.658 | 1.609 | **2.727** | -48.295 | **2.741** |
| | SNR↑ | **4.795** | 4.294 | 4.749 | -2.416 | **4.754** |
| | CCPC↑ | 0.8436 | 0.8470 | **0.8575** | 0.7311 | **0.8576** |
| Music | MEL↓ | 0.8754 | 0.9399 | **0.7085** | 1.3151 | **0.6892** |
| | STFT↓ | **2.6449** | 2.8116 | 2.8001 | 4.0485 | **2.6679** |
| | MFCC↓ | **0.2434** | 0.2722 | 0.2499 | 0.3593 | **0.2283** |
| | Wave.↓ | 0.0380 | 0.0363 | **0.0313** | 0.1700 | **0.0311** |
| | SI-SDR↑ | 9.721 | 9.985 | **11.568** | -46.243 | **11.649** |
| | SNR↑ | 10.189 | 10.475 | **11.813** | -2.847 | **11.864** |
| | CCPC↑ | 0.9563 | 0.9638 | **0.9765** | 0.7771 | **0.9765** |
| Speech | MEL↓ | 0.8101 | 0.6782 | **0.3355** | 1.1846 | **0.3555** |
| | STFT↓ | 1.8485 | 1.4548 | **1.2274** | 3.3933 | **1.4231** |
| | MFCC↓ | 0.1719 | 0.1337 | **0.0610** | 0.3027 | **0.0710** |
| | Wave.↓ | 0.0076 | 0.0073 | **0.0054** | 0.0318 | **0.0054** |
| | SI-SDR↑ | 9.489 | 9.406 | **11.781** | -36.506 | **11.777** |
| | SNR↑ | 9.970 | 9.955 | **12.070** | -2.935 | **12.066** |
| | PESQ↑ | 2.433 | 3.192 | **4.473** | 2.995 | **4.473** |
| | CCPC↑ | 0.9981 | 0.9971 | **0.9987** | 0.9881 | **0.9987** |

Table 14: Reconstruction comparison on the open sets. **Bold**: best in group.

| | | 44.1 kHz | | | 48 kHz | |
|---|---|---|---|---|---|---|
| Subset | Metric | SAME-L | $\varepsilon$ar-VAE | Ours | LTX 2.5 | Ours |
| Song Describer | MEL↓ | 1.3729 | 1.0172 | **0.7953** | 1.2841 | **0.7802** |
| | STFT↓ | 4.8432 | 3.3991 | **3.3646** | 4.0267 | **3.2060** |
| | MFCC↓ | 0.4132 | 0.3253 | **0.3171** | 0.3419 | **0.2973** |
| | Wave.↓ | 0.1133 | 0.0401 | **0.0346** | 0.1999 | **0.0346** |
| | SI-SDR↑ | -16.737 | 11.382 | **13.097** | -42.790 | **13.098** |
| | SNR↑ | 4.777 | 11.708 | **13.264** | -2.773 | **13.265** |
| | CCPC↑ | 0.8486 | 0.9701 | **0.9805** | 0.7741 | **0.9806** |
| LibriSpeech | MEL↓ | 1.2538 | 0.8368 | **0.3237** | 2.0286 | **0.3162** |
| | STFT↓ | 5.2836 | 3.8835 | **3.1338** | 6.2693 | **3.4890** |
| | MFCC↓ | 0.4487 | 0.3038 | **0.1848** | 0.5784 | **0.1840** |
| | Wave.↓ | 0.0067 | 0.0065 | **0.0044** | 0.0477 | **0.0044** |
| | SI-SDR↑ | 13.758 | 13.511 | **16.396** | -31.975 | **16.395** |
| | SNR↑ | 13.889 | 13.747 | **16.492** | -3.124 | **16.492** |
| | PESQ↑ | 2.914 | 3.360 | **4.460** | 3.712 | **4.460** |
| AudioCaps | MEL↓ | 1.0521 | 0.7638 | **0.5276** | 1.1980 | **0.5327** |
| | STFT↓ | 2.8198 | 1.8738 | **1.8246** | 3.3200 | **1.9408** |
| | MFCC↓ | 0.2888 | 0.1484 | **0.1064** | 0.2795 | **0.1138** |
| | Wave.↓ | 0.0275 | 0.0269 | **0.0238** | 0.0932 | **0.0238** |
| | SI-SDR↑ | 7.285 | 6.974 | **10.123** | -45.388 | **10.122** |
| | SNR↑ | 8.285 | 8.129 | **9.777** | -2.839 | **9.777** |
| | CCPC↑ | 0.9685 | 0.9697 | **0.9771** | 0.9356 | **0.9772** |

# E From Spatial Metrics to Temporal Audio Diagnostics

The diagnostic formulation is a change in the domain over which latent self-similarity is measured. In a vision representation, a token is indexed by position on an $H \times W$ patch lattice. In our representation, it is indexed by time on a 50 Hz sequence. We apply the spatial metric family of Singh et al. (2025) to that timeline. To the best of our knowledge, this is the first application of this metric family as temporal diagnostics for continuous audio tokenizers, not a new general theory of temporal correlation. The definitions below specify the evaluation protocol used for Tables 11 and 12.

## E.1 Temporal kernel and correlation decay

For a held-out clip, let $u_t = \mu_t \in \mathbb{R}^{d_z}$ and normalize each nonzero frame as $\hat{u}_t = u_t / \|u_t\|_2$, with numerical stabilization for small norms. Define

$$K(t, t') = \langle \hat{u}_t, \hat{u}_{t'} \rangle, \qquad d(t, t') = |t - t'|, \qquad \tau(t, t') = d(t, t')/50\,\text{s}. \tag{19}$$

This is equivalent to a $T \times 1$ lattice, but no image axis or frequency axis is introduced. Feature channels are not treated as spatial coordinates, and the two waveform channels are already represented jointly in $u_t$. Thus these are time-domain *latent* diagnostics, not raw-waveform autocorrelation or measures of acoustic source location.

For lag $\delta$, the correlogram is

$$g(\delta) = \frac{1}{T - \delta} \sum_{t=1}^{T-\delta} K(t, t + \delta). \tag{20}$$

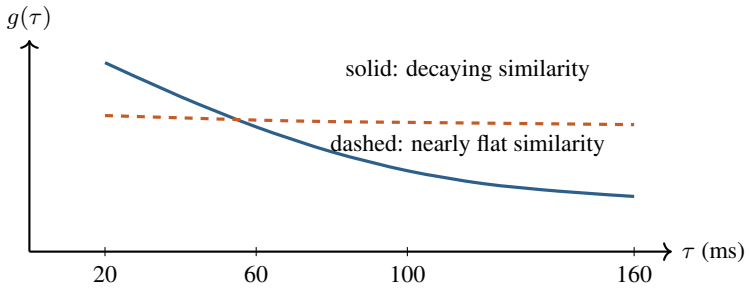


Figure 9: Schematic temporal correlograms, not experimental measurements. CDS measures the negative fitted slope over a specified lag range. The solid curve has greater local decay than the dashed curve, but neither curve alone establishes reconstruction fidelity or downstream generation quality.

We fit $g(\delta) \approx \alpha + \beta\delta$ by ordinary least squares over $\delta = 1, \ldots, \Delta$, using $\Delta = \min(8, \lfloor T/2 \rfloor)$, and report

$$\mathrm{CDS} = -\hat{\beta} = -\frac{\sum_{\delta=1}^{\Delta}(\delta - \bar{\delta})(g(\delta) - \bar{g})}{\sum_{\delta=1}^{\Delta}(\delta - \bar{\delta})^2}. \tag{21}$$

At $50\,\mathrm{Hz}$, the default fit spans 20–160 ms. Positive CDS indicates decreasing similarity over this interval, whereas near-zero CDS indicates a flat local correlogram, not necessarily a constant entire clip. At least two valid lags are needed to estimate a slope. When comparing tokenizers at different frame rates, lags must cover matched durations: fitting against seconds yields a slope $50$ times the per-frame value for this model.

### E.2 Complementary metrics and evaluation protocol

LDS contrasts adjacent-frame similarity with similarity beyond a declared far-lag threshold $r_{\mathrm{far}}$:

$$\mathrm{LDS} = \mathbb{E}_{|t-t'|=1} K(t, t') - \mathbb{E}_{|t-t'|\geq r_{\mathrm{far}}} K(t, t'). \tag{22}$$

The threshold must leave a nonempty distant-pair set and be reported in frames and seconds. RMSC measures variation of normalized frames around their temporal mean:

$$\mathrm{RMSC} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \|\hat{u}_t - \bar{u}\|_2^2}, \qquad \bar{u} = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_t. \tag{23}$$

RMSC detects directional collapse but, unlike CDS and LDS, does not depend on frame order. It is therefore a complementary diversity statistic, not a temporal-ordering score in isolation.

The encoder runs in evaluation mode on held-out clips truncated to at most ten seconds. Metrics are computed from the posterior mean separately for each valid clip before aggregation, rather than forming pairs across clip boundaries. Durations, lag settings, normalization, and the treatment of silence must be identical across the systems being compared.

These diagnostics connect directly to the pairwise term of the AF loss: the correlogram averages entries of the same within-clip similarity matrix that alignment constrains during training. They nevertheless remain distinct from a training-loss value, since the AF margin permits residual differences and a teacher's local structure need not be optimal for a particular downstream generator.

Clips with too few valid frames for the chosen lag range are excluded from slope estimation rather than assigned zero CDS. All diagnostics are computed on the same test set as the reconstruction metrics (section 5.1).

## F Model Architecture

The encoder and decoder follow the Descript Audio Codec (Kumar et al., 2023) with one additional block on each side. Mono and stereo models differ only in the number of input/output channels (1 or 2) and in the latent width $d_z$ (64 or 128).

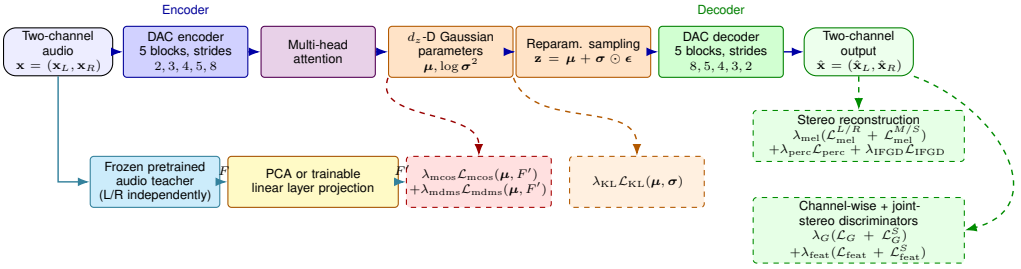


Figure 10: Audio VAE and training objectives. The DAC-based encoder, attention layer, and variational bottleneck produce $d_z$-dimensional latents. Frozen teacher features are mapped to the latent dimensionality either by a PCA basis fitted incrementally during VAE training or by a trainable linear layer, and the AF loss aligns the posterior mean $\boldsymbol{\mu}$ to them. The KL term regularizes both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. The spectral, phase, perceptual, and adversarial losses compare the decoded waveform with the input.

The encoder starts with an input convolution of width $64$ over the waveform channels, followed by five downsampling stages with strides $(2, 3, 4, 5, 8)$. Each stage contains three residual units with kernel size $7$ and dilations $(1, 3, 9)$, followed by a strided convolution that doubles the width. Weight normalization and Snake activations are used throughout the residual backbone. For a scalar activation $a$, Snake is $a + \alpha^{-1} \sin^2(\alpha a)$ with learned channel-wise $\alpha$. The final encoder width is $2048$.

An eight-head self-attention layer mixes the encoder frames before a pointwise projection maps $2048$ channels to $2d_z$. Splitting these channels gives the mean $\mu$ and log-variance $\ell = \log \sigma^2$. Training uses reparameterized samples,

$$z = \mu + \exp(\ell/2) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{24}$$

The decoder projects $z$ back to $2048$ channels. Its convolutional trunk starts at width $1536$ and upsamples with strides $(8, 5, 4, 3, 2)$, halving the channel width at each stage before an output convolution to the waveform channels and $\tanh$. The attention layer has zero dropout. The model is not constrained to causal or streaming inference.

The AF loss and the temporal diagnostics operate on $\mu$, avoiding stochastic variation introduced by posterior sampling.

The $50\,\text{Hz}$ interface gives a $20\,\text{ms}$ latent spacing and a $500$-frame sequence for ten seconds of audio. This is the sequence budget seen by a downstream generator, not a limit on the frequencies synthesized by the waveform decoder. Likewise, a joint latent does not imply mono output: each frame carries channel information in its feature dimensions, while the decoder predicts two waveforms.

# G Data Collection, Packaging, Loading, and Sampling

## G.1 Data Collection and Filtering

Training a stereo $48\,\text{kHz}$ VAE is difficult for three reasons, all of which concern the scarcity of genuinely full-band stereo material rather than the scarcity of audio as such. First, the corpus must cover music, speech, sound effects, and general audio in comparable proportions. Second, the declared sample rate does not imply full-band content: crowdsourcing platforms such as Freesound, the source of the FSD50K dataset (Fonseca et al., 2022), contain nominally $48\,\text{kHz}$ files that are recordings made at $8$, $16$, or $24\,\text{kHz}$ and subsequently upsampled. The container sample rate therefore carries little information about the true signal bandwidth. Third, channel count does not imply stereo: many two-channel files are duplicated mono, whose side channel is identically zero and which provide no stereo-imaging signal to the model. We therefore measure bandwidth and stereo content from the signal rather than reading them from the container. The corpus statistics in Table 15 are based on these measurements.

We crawled approximately $100$ million raw audio files of varying quality. To obtain high-quality stereo data, we applied the following blocking filters in sequence:

1. **Declared-metadata filtering:** duration at least $0.5\,\mathrm{s}$, two channels, and declared sample rate of at least $32\,\mathrm{kHz}$.
2. **Domain separation:** assignment to music, speech, sound effects, or general audio according to acquisition source.
3. **Segmentation and silence filtering:** segmentation into $10\,\mathrm{s}$ crops and removal of any crop whose silent fraction exceeds $60\%$.

This procedure yields approximately $100{,}000$ hours each for speech and music, $50{,}000$ hours for sounds, and $60{,}000$ hours for general audio. These per-domain estimates sum to approximately $310{,}000$ hours, or about $10^8$ ten-second crops (Table 15). The training and test sets contain no overlapping audio samples.

Table 15: Training-corpus statistics by domain. Hours are approximate. Stereo status and true sample rate are estimated from the waveform rather than inferred from file metadata.

| Domain | Stereo (%) | True SR (kHz) | Hours |
|---|---|---|---|
| Speech | $\sim 78$ | $\leq 44.1$ | $\sim 100{,}000$ |
| Music | $\sim 95$ | $\leq 44.1$ | $\sim 100{,}000$ |
| Sounds | $\sim 89$ | $\leq 192$ | $\sim 50{,}000$ |
| General audio | $\sim 76$ | $\leq 48$ | $\sim 60{,}000$ |

### G.2 Packaging and Loading

Training data are drawn from the four domain partitions in Table 15. The combined corpus is of order $10^8$ ten-second crops and resides on a shared parallel filesystem, which rules out addressing it as individual files. At this scale, a directory traversal costs tens of minutes per rank. The resulting path list occupies several gigabytes that copy-on-write does not protect across forked data-loader workers, because CPython stores reference counts in object headers. Random-excerpt extraction with one file open per attempt imposes several thousand metadata operations per second on the shared filesystem.

We therefore store each partition as a set of uncompressed tar archives in the WebDataset-style index layout, with an index manifest recording each archive's location and sample count. At construction, the loader parses only these manifests, so the resident index is proportional to the number of archives rather than the number of clips. Archives are memory-mapped in place, without copying them to node-local scratch, and their member-offset tables are parsed once when opened, so locating a member requires pointer arithmetic rather than an additional system call.

An initial version packed crops of varying duration into the archives. Repacking them to a uniform duration of $10\,\mathrm{s}$ reduced data-access time by approximately $10\%$ under an otherwise identical configuration because fixed member sizes make the sequential read pattern within an archive predictable.

### G.3 Training-Data Sampling

Domain-sampling weights are uniform: $w = 0.25$ for each of speech, music, sounds, and general audio. Each draw selects a domain according to this distribution and reads the next member sequentially from that domain's currently open archive. When an archive is exhausted, the next archive is taken from a per-domain queue that is reshuffled on each pass. Each worker therefore holds one open archive per domain, which keeps reads sequential and bounds the working set, at the cost of delegating sample-level decorrelation to the packing stage.

Each training example is a $0.38\,\mathrm{s}$ excerpt. At $48\,\mathrm{kHz}$, it spans exactly $18{,}240 = 19 \times 960$ samples, where $960 = 2 \cdot 3 \cdot 4 \cdot 5 \cdot 8$ is the total encoder stride. The excerpt therefore maps to an integral sequence of $19$ latent frames without a ragged boundary.

Offset selection is loudness-guided. We draw up to eight random offsets and evaluate each using an ITU–R BS.1770 integrated-loudness meter against a threshold of $-40$ LUFS. The first excerpt above the threshold is accepted, and if none passes, the final candidate is retained. The threshold thus biases sampling away from silence without excluding any archive member.

## H TRAINING CONFIGURATION

Table 16: Training configuration of the final stereo model.

| Setting | Value |
|---|---|
| Waveform / latent rate | $48\,\mathrm{kHz}$ stereo / $50\,\mathrm{Hz}$ |
| Encoder strides / base width | $(2, 3, 4, 5, 8)$ / $64$ |
| Encoder output / latent width | $2048$ / $d_z = 128$ |
| Decoder width / strides | $1536$ / $(8, 5, 4, 3, 2)$ |
| Attention | One $8$-head layer on the encoder side of the bottleneck, no dropout |
| Alignment | Frozen WavLM-Aesthetics teacher, PCA projection, $\lambda_{\mathrm{AF}} = 10$ |
| AF margins $m_1$ / $m_2$, internal weights $\lambda_{\mathrm{mcos}}$ / $\lambda_{\mathrm{mdms}}$ | $0.5$ / $0.25$, $1$ / $1$ |
| $\lambda_{\mathrm{mel}}$ / $\lambda_{\mathrm{feat}}$ / $\lambda_G$ | $15$ / $2$ / $1$ (both discriminator branches) |
| $\lambda_{\mathrm{KL}}$ / $\lambda_{\mathrm{perc}}$ / $\lambda_{\mathrm{IFGD}}$ | $10^{-3}$ / $2$ / $20$ |
| Training batch / crop / steps | $128$ / $0.38\,\mathrm{s}$ / $400{,}000$ |
| Optimizer | AdamW, learning rate $10^{-4}$, $\beta = (0.8, 0.99)$ |
| Learning-rate schedule | Exponential factor $0.999996$ per step |
| Gradient clipping | Generator $100$, discriminators $10$ |
| Precision / augmentation | Full precision / no waveform augmentation |
| Loudness | $-16$ LUFS target, $-40$ dB silence cutoff |
| Data-group probabilities | Speech, music, sounds, general: $0.25$ each |
| Validation | Every $5{,}000$ steps, $5\,\mathrm{s}$ clips, batch $8$ |

Table 16 lists the configuration of the final stereo model. Training audio is supplied in tar shards and sampled with replacement (section G). Nonfinite gradient values are replaced before norm clipping in both generator and discriminator updates. The recipe does not use rotary attention, a parallel teacher-feature skip into the latent, channel masking, or input-noise regularization. Mid/side is an auxiliary reconstruction basis, not the encoded waveform format.

Before teacher feature extraction, resampled $16\,\mathrm{kHz}$ audio is padded by $64$ samples on each side. Incremental PCA uses partial singular-value updates over teacher features flattened across examples and time. The current basis projects the features without a gradient, and reverse-projection reconstruction is not part of the AF loss. The PCA state at the end of training is saved with the checkpoint, since freezing the teacher does not by itself fix the projection.