AuDiAE: Learning Audio Latents by Coupling Diffusion Priors and Decoders

Anonymous authors
Paper under double-blind review

Abstract

Audio autoencoders determine both the information available for waveform reconstruction and the latent target learned by a downstream generator. Standard Gaussian posterior-KL regularization limits latent information but also penalizes mismatch between the aggregate posterior and a fixed Gaussian. Under a tight KL budget, this additional distribution-matching cost leaves less capacity for reconstruction-relevant detail. Downstream conditional generation additionally depends on how much latent variation is predictable from a condition such as text, while high compression leaves recording-specific acoustics underdetermined for deterministic waveform decoders. We introduce AuDiAE, an **Au**dio **Di**ffusion **A**uto**E**ncoder that couples a learned diffusion rate model with a generative waveform decoder through a common encoder representation. The marginal diffusion prior replaces the fixed reference with a learned density while maintaining a variational information bound. Its conditional counterpart makes condition-predictable information less costly in the training objective while the marginal rate continues to control total latent information. A flow-matching decoder trained to predict the clean waveform models residual variation and values encoded source detail through its reduction in optimal denoising risk. We evaluate AuDiAE at $2048\times$ and its high-compression variant, AuDiAE-H, at $4096\times$ on reconstruction and matched AudioCaps generation. AuDiAE leads every reported reconstruction metric and improves the strongest external AudioCaps FAD by 35% while attaining the highest CLAP. Matched controls support the learned rate reference and correct text–audio pairing; frozen diagnostics isolate tighter prior fit and text use without supporting a causal decoder advantage. Audio samples are available on the demo page.

1 Introduction

Audio autoencoders learn compact latent representations for waveform reconstruction and downstream generative modeling (Evans et al., 2025; Parker et al., 2026). The learned bottleneck determines both the information available to the decoder and the distribution on which a downstream generator is trained. Reconstruction fidelity alone does not establish whether the representation is suitable for generation. Skorokhodov et al. (2025) characterize this suitability as *diffusability*, the ease with which a diffusion model fits a latent distribution. Compactness, source fidelity, and diffusability jointly shape high-compression audio representation design.

Continuous audio variational autoencoders commonly regulate this bottleneck by penalizing each encoder posterior's Kullback–Leibler (KL) divergence from a standard Gaussian (Kingma & Welling, 2014; Evans et al., 2025). Stronger KL regularization can discard reconstruction-relevant detail, whereas weaker regularization can leave latents harder for a downstream generator to model (Seetharaman & Kumar, 2026). Target-KL (Seetharaman & Kumar, 2026) sets a target for the penalty, while STAR-VAE (Liu et al., 2026) redistributes it across channels. However, these refinements still measure the encoder against a fixed Gaussian. The expected posterior KL sums waveform–latent mutual information and aggregate-posterior mismatch to that reference (Hoffman & Johnson, 2016; Chen et al., 2018). Reducing it can therefore discard waveform information, reshape the aggregate distribution, or do both. Under a tight budget, fixed-reference matching consumes rate that could otherwise carry reconstruction-relevant detail (Figure 1(a)).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

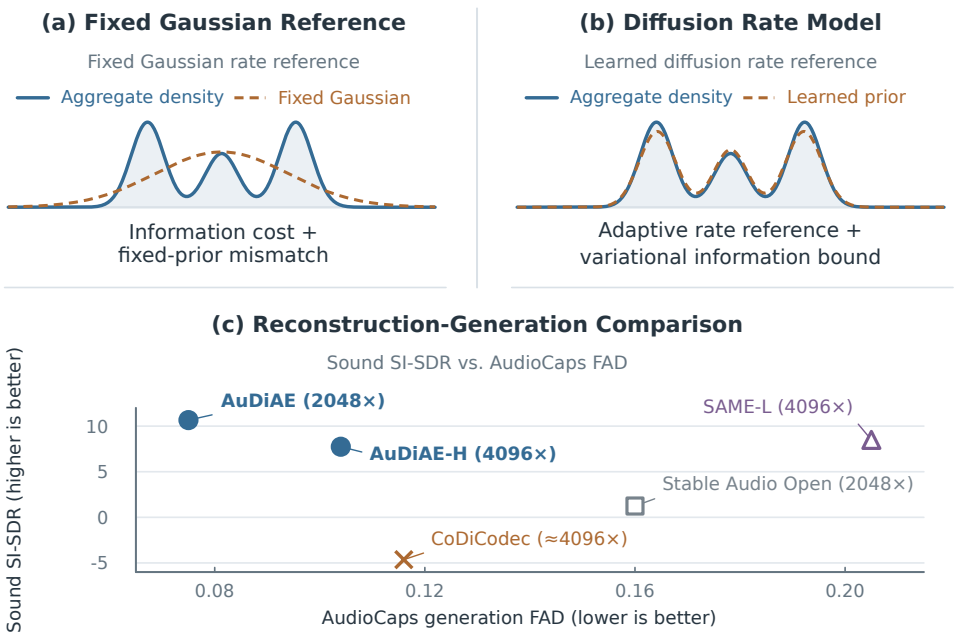


Figure 1: **Adaptive priors and reconstruction–generation quality. (a)** A fixed Gaussian reference charges both transmitted information and aggregate-prior mismatch. **(b)** AuDiAE adapts a diffusion reference to the aggregate latent distribution while maintaining a variational rate bound. **(c)** Every autoencoder reported in Table 2 is shown using AudioCaps generation FAD on the horizontal axis and the corresponding sound-domain reconstruction SI-SDR on the vertical axis; upper-left is better. Parentheses give temporal compression; AuDiAE-H denotes the high-compression $4096\times$ configuration.

For a downstream conditional generator, *diffusability* depends not only on marginal regularity but also on how much latent variation remains after a condition such as a text caption is given. Semantic alignment can reduce this unexplained variation by emphasizing condition-relevant structure; SAME, for example, combines semantic alignment losses with an auxiliary unconditional generative objective (Parker et al., 2026). At high compression, limited capacity devoted to shared semantic content can leave recording-specific acoustics underdetermined. The prevailing spectral–adversarial autoencoder paradigm (Evans et al., 2025; Liu et al., 2026; Parker et al., 2026) nevertheless decodes each latent to a single waveform, collapsing the one-to-many inverse of a compressed representation into a point estimate. A generative decoder (Chen et al., 2017) can instead model the residual waveform distribution, while its reconstruction objective rewards encoded details according to how much they reduce this uncertainty. The central question is *how to learn compact audio latents that are readily modeled by a downstream conditional generator and still support source-faithful waveform synthesis*.

We introduce **AuDiAE**, an **Au**dio **Di**ffusion **A**uto**E**ncoder that couples a learned diffusion rate model with a generative waveform decoder around a common encoder representation. The marginal (unconditional) diffusion prior replaces the fixed Gaussian reference with a learned density while maintaining a variational upper bound on transmitted information (Kingma et al., 2021). For a fixed encoder distribution, adapting this prior can reduce aggregate–prior mismatch without changing the information already carried by the latent (Figure 1(b)); joint optimization can additionally change the encoder's information allocation. The conditional counterpart upper-bounds waveform information left unexplained by text, and mixing the two rates makes text-predictable information less costly while the marginal component controls total information. For reconstruction, a clean-target flow-matching decoder (Lipman et al., 2023) models residual waveform variation and values encoded detail by how much it reduces optimal weighted denoising risk. In effect, the criterion favors source detail when it is predictable from text or useful for waveform denoising, relative to its marginal information cost.

Across the evaluated benchmarks, AuDiAE gives the best observed value in every reconstruction column and leads matched AudioCaps generation with 0.075 FAD and 0.328 CLAP, while AuDiAE-

H exposes a compactness–fidelity trade-off (Figure 1(c)). Matched controls support system-level benefits from the learned rate reference and paired captions over unconditional and shuffled-caption controls; frozen diagnostics separately isolate tighter prior fit and text use.

Our contributions are as follows:

- We introduce a learned marginal diffusion rate model that adapts the rate reference to the aggregate encoder distribution while upper-bounding transmitted information.
- We extend it with a text-conditional rate whose mixture with the marginal rate favors text-predictable information while preserving total-information control.
- We pair this bottleneck with clean-target waveform flow matching; controls support learned and paired-caption rate objectives, while frozen diagnostics isolate tighter prior fit and text use.

2 Background and Problem Formulation

2.1 Gaussian-Regularized Autoencoding and Prior Mismatch

Let $X \sim p_{\mathrm{data}}$ denote an audio waveform. An encoder with parameters ϕ defines a continuous stochastic latent $Z \sim q_\phi(\cdot \mid X)$, and a decoder models the waveform through $p_\theta(x \mid z)$; a deterministic decoder is a special case. Uppercase symbols denote random variables and lowercase symbols their realizations. We write $\mathbf{I}$ for an identity covariance and $I(\cdot;\cdot)$ for mutual information. A conventional audio variational autoencoder (VAE) uses a diagonal Gaussian posterior $q_\phi(z \mid x) = \mathcal{N}(\mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x)))$ (Kingma & Welling, 2014). Define its aggregate posterior $q_\phi(z) = \int q_\phi(z \mid x) p_{\mathrm{data}}(x)\, dx$. For any reference density p, the expected posterior KL, or rate, decomposes whenever the terms are finite (Hoffman & Johnson, 2016):

$$R_p := \mathbb{E}_X \,\mathrm{KL}\left(q_\phi(\cdot \mid X) \| p\right) = I(X;Z) + \mathrm{KL}\left(q_\phi \| p\right). \tag{1}$$

The two terms are transmitted information and aggregate–reference mismatch, so R_p upper-bounds $I(X;Z)$. Conventional training minimizes $\mathcal{L}_{\mathrm{rec}} + \beta R_0$, where R_0 uses the fixed standard Gaussian $p_0(z) = \mathcal{N}(0, \mathbf{I})$. Equation 1 gives $R_0 = I(X;Z) + M_0$, with $M_0 = \mathrm{KL}(q_\phi \| p_0)$. Holding q_ϕ fixed, choosing $p = q_\phi$ makes the bound tight, whereas p_0 incurs M_0. Two encoders can therefore transmit the same input information yet incur different R_0 because their aggregate distributions differ. Lowering R_0 can discard input information, reshape the aggregate posterior toward p_0, or do both; under a tight budget, mismatch consumes rate that could carry reconstruction-relevant detail. The Gaussian posterior KL remains a valid information upper bound, but does not isolate information quantity from distributional shape. Nor does this decomposition determine whether Gaussian-shaped latents help a particular generator. Fitting a flexible prior to a frozen encoder can reduce mismatch without changing $I(X;Z)$ or the downstream generation target.

2.2 Conditional Learnability

For the waveform X above, let C denote its paired generation condition, so $(X, C) \sim p_{\mathrm{data}}$. Let E_ϕ denote our deterministic encoder map and $U = E_\phi(X)$ its clean output; unlike the conventional VAE above, E_ϕ does not predict a posterior variance. Instead, the prescribed channel $Z \sim q_{\mathrm{ch}}(\cdot \mid U)$ introduces all encoder-side stochasticity. This gives the Markov chain $C \to X \to U \to Z$. A downstream generator models the encoder-induced law $q_\phi(u \mid c)$. This conditional law, rather than the marginal $q_\phi(u)$ alone, determines the target seen by the generator: identical marginals can still induce different conditional learning problems. For the stochastic representation, the Markov structure gives

$$I(X;Z \mid C) = I(X;Z) - I(C;Z), \qquad I(C;Z) \le I(C;U). \tag{2}$$

Here $I(X;Z \mid C)$ is encoded input variation left unexplained by the condition, and the inequality follows from data processing. It does not by itself determine the downstream law on U.

3 AuDiAE

AuDiAE learns its audio bottleneck through marginal and text-conditional diffusion rate models around a common encoder output (Figure 2). The marginal model replaces the fixed Gaussian

reference while upper-bounding total information (Section 3.1), and the conditional extension favors condition-predictable information without removing marginal rate control (Section 3.2). We pair this bottleneck with a generative waveform decoder for reconstruction (Section 3.3) and train the components jointly (Section 3.4). Given a paired caption Y, $C = T_{\text{text}}(Y)$ denotes its frozen feature; $\varnothing$ denotes the learned null condition used by the rate model. Appendix B derives these objectives and makes their scope explicit: mismatch and slack remain, conditional rate defines a preference rather than a guaranteed gain, and MRSTFT lies outside the variational analysis.

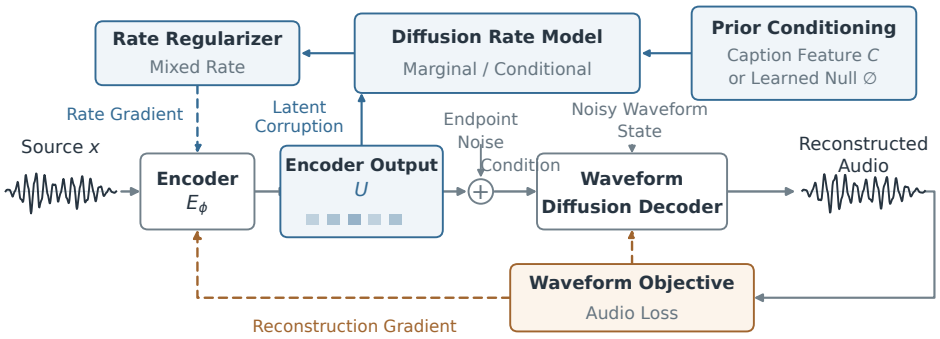


Figure 2: **High-level overview of the AuDiAE framework.**

3.1 Diffusion-Prior Audio Bottleneck

Following the rate decomposition in Section 2.1, we seek an adaptive reference density with tractable information control. The learned prior models the endpoint distribution used to condition the decoder. The deterministic encoder output U introduced in Section 2.2 lies in $\mathbb{R}^d$. Here T_z is the number of latent time steps, d_z is the number of channels per step, and $d = T_z d_z$. Let latent log-SNR $\lambda \in [a, b]$ run from high-noise reverse-process start a to decoder endpoint b, with $\alpha_\lambda^2 = \text{sigmoid}(\lambda)$ and $\sigma_\lambda^2 = \text{sigmoid}(-\lambda)$, and draw endpoint noise $\eta_b \sim \mathcal{N}(0, \mathbf{I})$. The decoder condition is

$$Z_b = \alpha_b U + \sigma_b \eta_b, \tag{3}$$

so that $\lambda = \log(\alpha_\lambda^2/\sigma_\lambda^2)$. Thus $q_{\text{ch}}(z_b \mid u) = \mathcal{N}(\alpha_b u, \sigma_b^2 \mathbf{I})$, and the induced encoder posterior and its aggregate density are

$$q_\phi^b(z \mid x) = q_{\text{ch}}(z \mid E_\phi(x)), \qquad q_\phi^b(z) = \int q_\phi^b(z \mid x) p_{\text{data}}(x)\, dx. \tag{4}$$

The diffusion prior reverses the variance-preserving corruption $Z_\lambda = \alpha_\lambda U + \sigma_\lambda \eta_\lambda$, where $\eta_\lambda \sim \mathcal{N}(0, \mathbf{I})$. A denoiser with parameters ψ predicts the clean latent $\widehat{U}_\psi(z, \lambda, c)$, equivalently

$$\epsilon_\psi(z, \lambda, c) = \frac{z - \alpha_\lambda \widehat{U}_\psi(z, \lambda, c)}{\sigma_\lambda}. \tag{5}$$

The marginal prior p_u uses the null condition $c = \varnothing$. For a normalized log-SNR density π_z on $[a, b]$, define

$$J_u = \mathbb{E}_{\lambda \sim \pi_z, X, \eta_\lambda} \frac{\|\eta_\lambda - \epsilon_\psi(Z_\lambda, \lambda, \varnothing)\|^2}{2\pi_z(\lambda)}. \tag{6}$$

The draws η_λ used for prior training and η_b used for the decoder condition are independent. The two paths share U and the endpoint law at b, rather than a particular noise realization or noise level. Because a is finite, the forward law $Z_a \mid X \sim \mathcal{N}(\alpha_a E_\phi(X), \sigma_a^2 \mathbf{I})$ generally differs from the standard Gaussian used to initialize the reverse process. Its expected KL to that initialization distribution gives the boundary term

$$B_a = \frac{1}{2}\left[\alpha_a^2 \mathbb{E}\|U\|^2 + d(\sigma_a^2 - 1 - \log \sigma_a^2)\right]. \tag{7}$$

Diffusion variational inference (Kingma et al., 2021; Heek et al., 2026) then gives

$$I(X; Z_b) \le R_u \le \overline{R}_u = B_a + J_u, \qquad R_u = \mathbb{E}_X \,\text{KL}\left(q_\phi^b(\cdot \mid X) \| p_u\right). \tag{8}$$

Here R_u is the exact expected rate under p_u, and $\overline{R}_u$ is its tractable diffusion variational bound. The reverse process and bound are detailed in Appendix B.1.

The exact rate decomposes as $R_u = I(X; Z_b) + M_u$, where $M_u = \mathrm{KL}(q_\phi^b \| p_u)$ is the aggregate–prior mismatch. Holding q_ϕ^b fixed, learning the prior can reduce M_u and the variational slack $G_u = \overline{R}_u - R_u$, but cannot change $I(X; Z_b)$. It is important to distinguish two superficially similar operations: rescaling U *before* the channel changes its precision and potentially $I(X; Z_b)$, whereas an invertible normalization *after* forming Z_b rescales signal and noise together and preserves mutual information. Specifically, for a centered coordinate U_j, $j \in \{1, \dots, d\}$, the channel SNR is $\alpha_b^2 \operatorname{Var}(U_j)/\sigma_b^2 = e^b \operatorname{Var}(U_j)$; rescaling U therefore changes the SNR even at fixed b, and Equation 7 records this dependence through $\mathbb{E}\|U\|^2$. Post-channel normalization, by contrast, is only a coordinate change. Conflating these operations could incorrectly attribute a rate change to an information-preserving normalization (Appendix B.2).

3.2 Text-Conditional Rate Regularization

The conditional-learning requirement in Section 2.2 concerns the law of the deterministic output U given text, whereas the rate objective acts on its noisy counterpart Z_b. Equation 2 shows that conditioning subtracts the latent–text dependence term from the total input information, motivating a text-conditioned rate alongside the marginal rate. Using the frozen caption feature C, the same diffusion prior receives C or the null condition, defining $p_c(z \mid c)$ or $p_u(z)$. The conditional aggregate is $q_\phi^b(z \mid c) = \int q_\phi^b(z \mid x) p_{\mathrm{data}}(x \mid c)\, dx$. During autoencoder training, the encoder observes only audio and the waveform decoder receives only Z_b, so $Z_b \perp C \mid X$.

Alongside the marginal rate R_u defined in Equation 8, define the conditional aggregate–prior mismatch as $M_c := \mathbb{E}_C \,\mathrm{KL}(q_\phi^b(\cdot \mid C) \| p_c(\cdot \mid C))$. The conditional rate is

$$R_c = \mathbb{E}_{X,C} \,\mathrm{KL}\left(q_\phi^b(\cdot \mid X) \| p_c(\cdot \mid C)\right) = I(X; Z_b \mid C) + M_c. \tag{9}$$

Replacing the null condition in Equation 6 by C gives J_c and $\overline{R}_c = B_a + J_c$, with conditional variational slack $G_c := \overline{R}_c - R_c \geq 0$.

For a mixture weight $\rho \in [0, 1)$, define the mixed mismatch-and-slack term $\mathcal{E}_\rho = (1 - \rho)(M_u + G_u) + \rho(M_c + G_c)$. We regularize the encoder with

$$\begin{aligned} \overline{R}_\rho &= (1 - \rho)\overline{R}_u + \rho \overline{R}_c \\ &= I(X; Z_b) - \rho I(Z_b; C) + \mathcal{E}_\rho. \end{aligned} \tag{10}$$

For fixed $\mathcal{E}_\rho$, the explicit $-\rho I(Z_b; C)$ term in Equation 10 lowers the rate cost of information predictable from text. This realizes the information bottleneck preference for condition-relevant information (Tishby et al., 1999; Alemi et al., 2017) without aligning latent coordinates to a text embedding or prescribing semantic channels (Appendix B.3). The marginal component preserves total-information control,

$$I(X; Z_b) \leq \frac{\overline{R}_\rho}{1 - \rho}. \tag{11}$$

Because $\overline{R}_u - \overline{R}_c$ also contains mismatch and slack, it is not a direct measure of text-predictable information; $\overline{R}_\rho$ is also a regularizer rather than an operational bitrate.

3.3 Waveform Diffusion Decoding

At high compression, a rate-constrained latent can remain compatible with multiple waveforms, leaving residual acoustic variation in the conditional law of X given Z_b. We therefore formulate decoding as conditional waveform generation: the latent supplies source structure, while the decoder models residual acoustic variation (Chen et al., 2017). Generative decoding cannot recover source information absent from Z_b; source fidelity therefore depends on both what the encoder transmits and how the decoder uses it. Let A be a fixed preprocessing map and $S = A(X) \in \mathbb{R}^{d_x}$ the waveform target. Given Z_b from Equation 3, we independently sample $t \in [0, 1)$ and $\xi \sim \mathcal{N}(0, \mathbf{I})$. A

clean-target predictor defines the corresponding conditional velocity (Lipman et al., 2023):

$$S_t = tS + (1-t)\xi, \qquad \widehat{S} = F_\theta(S_t, t, Z_b), \tag{12}$$
$$\widehat{v}_\theta(S_t, t, Z_b) = \frac{F_\theta(S_t, t, Z_b) - S_t}{1-t}.$$

Let $\ell = 2\log[t/(1-t)]$ denote the nominal log-SNR of the waveform path and $t(\ell) = \mathrm{sigmoid}(\ell/2)$. On $[\ell_-, \ell_+]$, a positive weight $w(\ell)$ defines the denoising risk

$$\mathcal{D}_w(F, Z_b) = \frac{1}{2d_x} \int_{\ell_-}^{\ell_+} w(\ell) e^{\ell} \mathbb{E}\|S - F(S_{t(\ell)}, t(\ell), Z_b)\|_2^2 \, d\ell. \tag{13}$$

This form separates prediction space from noise-level weighting (Kingma & Gao, 2023); Appendix B.5 gives the sampled estimator and the exact conversion among parameterizations and effective weights.

For a fixed encoder, the Bayes predictor is $F^*(S_t, t, Z_b) = \mathbb{E}[S \mid S_t, t, Z_b]$, with risk $\mathcal{D}_w^*(Z_b) := \mathcal{D}_w(F^*, Z_b)$. The latent's synthesis value is

$$V_w(Z_b) = \mathcal{D}_w^*(\varnothing) - \mathcal{D}_w^*(Z_b) \geq 0, \tag{14}$$

where the first term excludes latent input. Combining this reduction with β times Equation 10 gives the ideal encoder-dependent preference

$$V_w(Z_b) + \beta\rho I(Z_b; C) - \beta I(X; Z_b) - \beta\mathcal{E}_\rho. \tag{15}$$

Thus the rate construction determines the cost of encoded information, while the decoder rewards information that reduces synthesis uncertainty. This interpretation concerns Bayes risk; Appendix B.4 gives the finite-decoder and acoustic-loss correction.

3.4 Training and Inference

We jointly optimize the encoder, rate model, and waveform decoder with parameters (ϕ, ψ, θ) by minimizing

$$\mathcal{J}_\rho = \mathcal{L}_{\mathrm{FM}} + \kappa_{\mathrm{STFT}}\mathcal{L}_{\mathrm{MRSTFT}} + \beta\overline{R}_\rho, \qquad \overline{R}_\rho = B_a + (1-\rho)J_u + \rho J_c, \tag{16}$$

where $\mathcal{L}_{\mathrm{FM}}$ is the sampled clean-target loss associated with Equation 12. The multiresolution spectral-convergence and log-magnitude term $\mathcal{L}_{\mathrm{MRSTFT}}$ explicitly penalizes spectral discrepancies across time–frequency resolutions. The path costs J_u and J_c fit the shared prior ψ to p_u and $p_c(\cdot \mid C)$, the rate and reconstruction terms shape the encoder ϕ, and the waveform terms train the decoder θ. We implement β through the element-mean coefficient β_{mean}, with $\beta = \beta_{\mathrm{mean}}/d$. Training proceeds in two stages. Stage 1 uses only the null condition ($\rho = 0$), so the rate reduces to $\overline{R}_u = B_a + J_u$. Stage 2 starts from Stage 1, feeds the same prior either C or the null condition, and sets $\rho > 0$.

Reading $\mathcal{L}_{\mathrm{FM}}$ as the risk of Section 3.3 and omitting MRSTFT, the Bayes-optimal form of $\mathcal{J}_\rho$ equals the negative of Equation 15 up to an encoder-independent constant. We retain MRSTFT as a deliberate practical inductive bias for multiscale spectral fidelity, not as part of the variational derivation; consequently, the complete objective is not an exact waveform ELBO. We do not isolate the contribution of MRSTFT in the present controls. Appendix B.4 gives the finite-model and acoustic-loss corrections.

Inference. For source reconstruction, the encoder produces $U = E_\phi(X)$. Channel-consistent reconstruction obtains Z_b through Equation 3, whereas clean-latent reconstruction conditions the decoder directly on U. In either case, the decoder integrates $\widehat{v}_\theta$ from Gaussian noise to audio. For synthesis without a source waveform, the rate model samples Z_b from p_u or $p_c(\cdot \mid C)$ before waveform decoding. For downstream transfer, we instead fit a separate generator to U, without reusing the rate prior.

Table 1: **Audio reconstruction comparisons on sound and music benchmarks.** Sound and music metrics are dataset-macro averages. rFAD averages AudioCaps and Song Describer; MOS is the mean $\pm$ sample standard deviation over 60 item-level scores, each averaged across five listeners in the balanced speech, music, and sound panel. AuDiAE-H denotes the high-compression $4096\times$ configuration. Compression factors and native decoders differ, so the table is a system comparison rather than a rate-matched ablation; bold marks the best observed value.

Autoencoder	Temporal compression	Sound avg.			Music avg.			rFAD	MOS
		SI-SDR ↑	STFT ↓	MEL ↓	SI-SDR ↑	STFT ↓	MEL ↓	↓	↑
Single-pass continuous autoencoders									
ϵar-VAE	$1024\times$	8.15	0.060	0.054	13.28	0.085	0.071	0.054	3.921 ± 0.321
Stable Audio Open	$2048\times$	1.24	0.067	0.055	7.30	0.097	0.074	0.101	3.837 ± 0.330
SAME-L	$4096\times$	8.49	0.062	0.050	12.84	0.089	0.061	0.088	3.921 ± 0.336
Generative continuous autoencoders									
Moûsai (v2)	$1024\times$	-30.19	0.094	0.111	-30.23	0.127	0.142	0.341	2.675 ± 0.140
CoDiCodec	$\sim 4096\times$	-4.64	0.073	0.068	0.73	0.106	0.099	0.074	3.579 ± 0.413
Ours									
AuDiAE	$2048\times$	**10.66**	**0.055**	**0.037**	**14.61**	**0.083**	**0.051**	**0.041**	$\mathbf{4.243 \pm 0.273}$
AuDiAE-H	$4096\times$	7.74	0.063	0.048	12.31	0.092	0.064	0.061	4.126 ± 0.280

4 Experiments

4.1 Experimental Setup

Implementation details. We use AuDiAE for the default $2048\times$, 192-channel configuration and AuDiAE-H for the high-compression $4096\times$, 256-channel configuration. Appendix C.1–C.3 specifies the data, architectures, optimization, and inference contracts.

Baselines and metrics. We compare with single-pass continuous autoencoders, ϵar-VAE (Wang et al., 2025), Stable Audio Open (Evans et al., 2025), and SAME-L (Parker et al., 2026), and with generative continuous autoencoders, Moûsai (v2) (Schneider et al., 2024) and CoDiCodec (Pasini et al., 2025). Reconstruction reports SI-SDR (Le Roux et al., 2019) and `log1p` multiresolution STFT/mel errors. Fréchet Audio Distance (Kilgour et al., 2019) in the LAION-CLAP audio space (Wu et al., 2023) is denoted rFAD for reconstructions and FAD for generated audio. We report text–audio CLAP for generation. Twelve listeners rate a fixed, balanced 60-source speech/music/sound panel on a five-point mean opinion score (MOS) scale; each item–system pair receives five ratings under a balanced incomplete-block assignment.

Evaluation protocols. Primary reconstruction covers sound with AudioCaps (Kim et al., 2019) and VGGSound (Chen et al., 2020), and music with Song Describer (Manco et al., 2023) and MusicCaps (Agostinelli et al., 2023). Paired metrics are first averaged within each dataset and then equally over the two datasets in each domain; reconstruction rFAD is computed separately on AudioCaps and Song Describer and equally averaged in Table 1. For downstream generation, every frozen autoencoder is paired with a separate bidirectional flow-matching Transformer. All compared representations share the generator family, optimization budget, and latent-sampling procedure. Appendix C fixes the complete training, inference, population, and aggregation contracts; equal generator updates match data exposure rather than compute.

4.2 Main Results

4.2.1 Audio reconstruction

Table 1 gives two observations. (1) At $2048\times$, AuDiAE attains the most favorable observed value in every reported column, including against the $1024\times$ systems. Relative to Stable Audio Open at equal temporal compression, sound/music SI-SDR rises from 1.24/7.30 to 10.66/14.61 dB; across all baselines, AuDiAE lowers rFAD to 0.041 and gives the highest MOS point estimate, 4.243 ± 0.273. (2) At $4096\times$, AuDiAE-H improves rFAD over SAME-L (0.061 vs. 0.088) and has a slightly higher MOS point estimate (4.126 vs. 3.921), whereas SAME-L retains higher SI-SDR and lower STFT error. AuDiAE-H nevertheless exceeds the generative CoDiCodec baseline in every reported column.

Table 2: **Text-to-audio generation on the AudioCaps test set.** Every frozen autoencoder is paired with the same generator family trained on 45,044 AudioCaps items for 200k updates. Sampling, prompts, and evaluation are fixed across rows; equal updates do not imply equal compute.

Autoencoder	FAD ↓	CLAP ↑
Stable Audio Open	0.160	0.318
SAME-L	0.205	0.284
CoDiCodec	0.116	0.290
AuDiAE	**0.075**	**0.328**
AuDiAE-H	0.104	0.316

This pattern is consistent with the generative decoder favoring distributional and perceived realism over exact sample-aligned fidelity at extreme compression, but the cross-system comparison does not isolate the decoder as the cause. These are system-level comparisons across different native decoders; Section 4.3 provides the matched controls.

4.2.2 AUDIO GENERATION

To compare downstream learnability, we freeze each autoencoder and fit the same generator family under matched AudioCaps data, initialization, updates, sampling, and guidance. The $2048\times$, 192-channel AuDiAE attains the lowest FAD and the highest CLAP in Table 2. Among representations at approximately $4096\times$, the 256-channel AuDiAE-H improves on both SAME-L and CoDiCodec in FAD (0.104 vs. 0.205 and 0.116) and CLAP (0.316 vs. 0.284 and 0.290). Because latent length and width differ across rows, equal generator updates do not imply equal compute; these comparisons evaluate complete representations rather than isolating the learned prior.

4.3 CONTROLLED ANALYSIS

Table 3: **Controlled audio reconstruction and text-to-audio generation on the AudioCaps test set.** Panel (a) compares the fixed-Gaussian rate with the learned unconditional diffusion rate. Panel (b) compares unconditional, shuffled-caption, and paired-caption diffusion-rate objectives. All distributional metrics use LAION-CLAP.

(a) Rate-model control

Rate objective	Reconstruction				Generation	
	MEL ↓	STFT ↓	SI-SDR ↑	rFAD ↓	FAD ↓	CLAP ↑
Fixed-Gaussian rate (control)	0.0308	0.0501	13.7645	0.0893	0.142	0.274
Learned unconditional diffusion rate	**0.0284**	**0.0473**	**14.8307**	**0.0741**	**0.093**	**0.305**

(b) Text-conditional rate

Rate objective	FAD ↓	CLAP ↑
Unconditional diffusion rate	0.093	0.305
Mixed diffusion rate, shuffled captions	0.095	0.303
Mixed diffusion rate, paired captions (AuDiAE)	**0.075**	**0.328**

A learned unconditional rate improves fidelity and downstream learnability. Table 3(a) compares rate references under the same prescribed endpoint channel and training contract. Replacing the exact element-mean KL to $\mathcal{N}(0, I)$ with the learned unconditional diffusion bound reduces MEL/STFT by 7.8%/5.6%, raises SI-SDR by 1.07 dB, and lowers rFAD by 17.0%. The resulting representation also improves downstream generation, reducing FAD from 0.142 to 0.093 and increasing CLAP from 0.274 to 0.305. The agreement across paired, distributional, and downstream metrics supports the learned rate model under this fixed-channel contract; the control is not a conventional VAE with learned posterior variance.

The conditional benefit depends on correct text–audio pairing. Panel (b) examines whether the downstream effect follows paired captions rather than merely adding a conditional branch. Shuffled

captions do not improve over the unconditional rate (0.095/0.303 vs. 0.093/0.305), whereas paired captions recover the primary AuDiAE result (FAD 0.075, CLAP 0.328; Table 2). Relative to the unconditional control, paired conditioning reduces FAD by 19.4% and increases CLAP by 0.023. This matched contrast supports an end-to-end benefit of the paired-caption rate objective; the shuffled control rules out extra conditional capacity alone as the explanation.

Evidence boundary. These controls identify system-level effects of the rate objectives, not a single internal cause. In particular, they do not separate encoder reshaping from prior or decoder co-adaptation. Appendix D provides narrower frozen-representation evidence: the diffusion prior gives a tighter bound than the best fitted Gaussian (3.104 vs. 3.264 nats per element), and a residual condition adapter assigns lower rate to paired than mismatched text by 3.79×10^{-3} nats per element. A separate decoder-objective screen does not show an overall improvement.

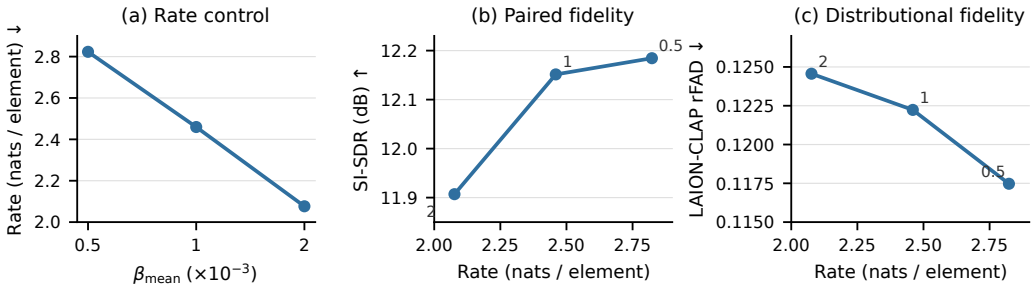


Figure 3: **Effect of rate pressure at fixed endpoint precision.** The three points use the same $2048\times$ controlled-analysis configuration, endpoint, pre-specified 200k-update budget, and fixed evaluation cohort. Panel (a) shows the measured rate as β_{mean} changes. Panels (b–c) reparameterize the same autoencoders by achieved rate; rFAD uses the matched 975-recording AudioCaps subset. Point labels give $10^3\beta_{\text{mean}}$.

Rate pressure controls the achieved bound. Figure 3 varies β_{mean} on a 128-channel, $2048\times$ model under a common 200k-update budget. Increasing β_{mean} from 5×10^{-4} to 2×10^{-3} lowers the measured bound monotonically from 2.823 to 2.077 nats per element, while SI-SDR falls from 12.18 to 11.91 dB and rFAD rises from 0.117 to 0.125. Because this configuration differs from the main systems and the bound includes both mismatch and variational slack, we interpret the within-sweep direction rather than a global optimum or a change in $I(X; Z_b)$ alone.

5 Related Work

Continuous audio autoencoders pair spectral–adversarial reconstruction with variational, structured, or semantically aligned bottlenecks (Evans et al., 2025; Wang et al., 2025; Liu et al., 2026; Parker et al., 2026); generative alternatives use diffusion or consistency decoders (Schneider et al., 2024; Pasini et al., 2024; 2025). Learned-prior decompositions separate transmitted information from reference mismatch (Hoffman & Johnson, 2016; Tomczak & Welling, 2018), while information bottlenecks distinguish total from condition-relevant information (Tishby et al., 1999; Alemi et al., 2017). Building on joint diffusion prior–decoder training (Heek et al., 2026), AuDiAE studies a fixed-noise audio bottleneck with marginal and text-conditional diffusion rate bounds and clean-target waveform decoding. Its variational rate is not an operational bitrate, and the current comparisons do not isolate a causal decoder gain. Appendix A separates these design axes and their causal evidence.

6 Conclusion

We introduced AuDiAE, coupling a learned diffusion rate model with a generative waveform decoder, and AuDiAE-H for higher compression. Its bound separates transmitted information from prior mismatch; the conditional rate favors text-predictable latents; and the decoder objective values denoising utility. Controls show a local rate response and tighter frozen-representation prior fit, but not causal component or conditional/decoder gains.

AI USE STATEMENT

Generative AI tools were used for language polishing and LaTeX typesetting. The authors performed the research design, implementation, experiments, analysis, and verification, and retain responsibility for the final content.

REPRODUCIBILITY STATEMENT

Appendix B states the assumptions and boundaries of the formal claims. Appendix C specifies data populations, optimization, inference, preprocessing, metric aggregation, and the balanced 60-item incomplete-block MOS protocol. Appendix D separates matched mechanism evidence from sensitivity and negative diagnostics, while Appendix E records the remaining causal, statistical, and efficiency boundaries and Appendix E.1 documents the human-participant safeguards.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. URL https://arxiv.org/abs/2301.11325.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.00410.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020. URL https://www.robots.ox.ac.uk/~vgg/publications/2020/Chen20/.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract.html.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BysvGP5ee.

Michael Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pp. 316–323, 2017. URL https://archives.ismir.net/ismir2017/paper/000075.pdf.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888461.

Frédéric Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 411–412, 2013. doi: 10.1145/2502081.2502245.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

Dongning Guo, Shlomo Shamai (Shitz), and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005. doi: 10.1109/TIT.2005.844072. URL https://arxiv.org/abs/cs/0412108.

Jonathan Heek, Emiel Hoogeboom, Thomas Mensink, and Tim Salimans. Unified latents (UL): How to train your latents. *arXiv preprint arXiv:2602.17270*, 2026. URL https://arxiv.org/abs/2602.17270.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Matthew D. Hoffman and Matthew J. Johnson. ELBO surgery: Yet another way to carve up the variational evidence lower bound. In *NIPS Workshop on Advances in Approximate Bayesian Inference*, 2016.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of Interspeech*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of NAACL-HLT*, pp. 119–132, 2019. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2858.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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, 2023. doi: 10.52202/075280-1214.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR – half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 626–630, 2019. doi: 10.1109/ICASSP.2019.8683855. URL https://arxiv.org/abs/1811.02508.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Huadai Liu, Wen Wang, Kaicheng Luo, Qian Chen, Xiangang Li, and Wei Xue. STAR-VAE: Structured topology-aware regularization for audio reconstruction and generation. *arXiv preprint arXiv:2606.23064*, 2026. URL https://arxiv.org/abs/2606.23064.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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. URL https://arxiv.org/abs/2311.10057.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. doi: 10.1109/TASLP.2024.3419446.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. EXPRESSO: A benchmark and analysis of discrete expressive speech resynthesis. In *Interspeech*, pp. 4823–4827, 2023. doi: 10.21437/Interspeech.2023-1905. URL https://www.isca-archive.org/interspeech_2023/nguyen23_interspeech.html.

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. URL https://arxiv.org/abs/2605.18613.

Marco Pasini, Stefan Lattner, and George Fazekas. Music2latent: Consistency autoencoders for latent audio compression. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, pp. 111–119, 2024. doi: 10.5281/zenodo.14877289.

Marco Pasini, Stefan Lattner, and George Fazekas. CoDiCodec: Unifying continuous and discrete compressed representations of audio. In *Proceedings of the 26th International Society for Music Information Retrieval Conference*, pp. 433–441, 2025. doi: 10.5281/zenodo.17811403.

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

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8050–8068, 2024. doi: 10.18653/v1/2024.acl-long.437. URL https://aclanthology.org/2024.acl-long.437/.

Prem Seetharaman and Rithesh Kumar. Taming audio VAEs via Target-KL regularization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026. doi: 10.1109/ICASSP55912.2026.11460662. URL https://arxiv.org/abs/2605.17085.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55876–55905, 2025. URL https://proceedings.mlr.press/v267/skorokhodov25a.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 1999. URL https://arxiv.org/abs/physics/0004057. 37th Allerton Conference; arXiv:physics/0004057.

Jakub Tomczak and Max Welling. VAE with a VampPrior. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1214–1223, 2018. URL https://proceedings.mlr.press/v84/tomczak18a.html.

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. URL https://arxiv.org/abs/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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), 2019.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Guangting Zheng, Yiyuan Zhang, Tao Yang, Yunpeng Chen, Rui Zhu, Jiajun Deng, and Yanyong Zhang. GenFirst: Generation before reconstruction for stable end-to-end latent generative modeling. *arXiv preprint arXiv:2608.29335*, 2026. URL https://arxiv.org/abs/2608.29335.

A Extended Related Work

This appendix expands the compact positioning in Section 5. We distinguish representation interfaces, reconstruction mechanisms, rate models, and conditional learnability because evidence for one axis does not by itself establish improvement on another.

A.1 Continuous and Discrete Audio Bottlenecks

Continuous audio autoencoders learn compact representations for waveform reconstruction and subsequent latent generative modeling. Stable Audio Open (Evans et al., 2025) combines a convolutional VAE with spectral and adversarial reconstruction objectives. ɛar-VAE adds perceptual K-weighting, phase-derivative losses, and stereo spectral supervision (Wang et al., 2025). STAR-VAE redistributes regularization across latent channels and uses a hybrid convolutional–Mamba architecture (Liu et al., 2026), while Target-KL controls a Gaussian audio VAE toward a selected KL operating point (Seetharaman & Kumar, 2026). SAME instead combines a soft-normalization bottleneck with phase-sensitive, adversarial, semantic, and auxiliary generative objectives (Parker et al., 2026). These systems differ in posterior family, latent shaping, and reconstruction supervision; temporal compression alone therefore does not specify the representation-learning problem.

Discrete neural codecs such as Improved RVQGAN quantize audio into RVQ token streams with an operational bitrate (Kumar et al., 2023). AuDiAE instead studies a continuous fixed-noise bottleneck. Its variational rate controls information during learning and should not be interpreted as the bitrate of an entropy-coded file. This distinction is also why rate-matched continuous controls, rather than nominal temporal compression alone, are needed to attribute representation changes to the regularizer.

A.2 Generative Reconstruction

Spectral–adversarial decoders produce a single waveform for a fixed latent. Generative alternatives model conditional waveform variation given the encoded representation. Moûsai conditions a diffusion autoencoder decoder on compressed audio features (Schneider et al., 2024); Music2Latent develops end-to-end consistency autoencoding with encoder–decoder cross-connections and single-step reconstruction (Pasini et al., 2024); and CoDiCodec couples consistency reconstruction with summary embeddings and finite-scalar- quantization dropout to support continuous and discrete interfaces (Pasini et al., 2025). These works establish generative reconstruction as a viable design family, but differences in bottleneck, decoder, and training objective prevent an uncontrolled system comparison from isolating the decoder's causal contribution.

Variational Lossy Autoencoder formalizes how decoder capacity changes which source information is useful to encode (Chen et al., 2017). Weighted diffusion objectives further connect denoising losses to likelihood bounds under stated conditions (Kingma & Gao, 2023). AuDiAE uses conditional waveform denoising to define the ideal value of latent information as a reduction in Bayes denoising risk. This is an analysis of information allocation, not an empirical claim that a generative decoder necessarily outperforms a matched deterministic decoder. Its practical objective also contains MRSTFT supervision, which is a spectral inductive bias rather than a term in the exact variational derivation.

A.3 Learned Priors and Variational Rate Control

The VAE objective couples reconstruction to a divergence between the encoder posterior and a prior (Kingma & Welling, 2014). Aggregate-posterior decompositions separate transmitted information, latent dependence, and reference mismatch (Hoffman & Johnson, 2016; Chen et al., 2018). Learned priors relax the fixed-reference family; VampPrior, for example, evaluates a mixture of variational posteriors at learned pseudo-inputs (Tomczak & Welling, 2018). For a fixed encoder distribution, a better prior can reduce aggregate–prior mismatch without changing the information already present in the representation. End-to-end training is different: the prior and encoder may co-adapt, so both mismatch and information allocation can change.

Variational Diffusion Models provide continuous-data likelihood bounds and an SNR parameterization (Kingma et al., 2021). AuDiAE uses this construction as a learned rate reference at a finite-noise

audio endpoint and retains the corresponding boundary term. Its frozen-representation prior comparison is therefore the direct test of prior fit; reconstruction or downstream changes under joint training additionally contain encoder, decoder, and optimization effects. The current comparison with a full-channel Gaussian does not isolate a uniquely non-Gaussian benefit because that Gaussian omits temporal dependence.

A.4 Conditional Learnability and Joint Diffusion Training

The information bottleneck distinguishes total input information from condition-relevant content (Tishby et al., 1999), with neural variational realizations developed by Alemi et al. (2017). We do not claim this preference as a new principle. AuDiAE realizes it by mixing marginal and text-conditional diffusion rate bounds: the marginal component controls total information, while the conditional component lowers the relative cost of information predictable from text. Unlike direct semantic feature matching, text enters the rate prior rather than the audio encoder or waveform decoder.

Diffusability work studies how readily a diffusion model fits a learned latent distribution (Skorokhodov et al., 2025). For conditional audio generation, marginal regularity is only part of the target: identical marginal latent laws can have different dependence on text. SAME addresses related concerns through semantic and auxiliary generative objectives (Parker et al., 2026); AuDiAE instead changes the encoder's rate preference. A paired-versus-shuffled end-to-end contrast can establish that a downstream benefit depends on correct captions. Because the encoder, prior, and decoder co-adapt, however, it does not isolate a representation mechanism. The present evidence therefore supports a contract-specific system benefit rather than a universal conditional-representation gain.

Unified Latents is the closest joint diffusion construction: it couples encoder noise to a diffusion prior, uses a diffusion decoder, and learns representations for subsequent latent generation (Heek et al., 2026). The shared foundation is joint diffusion prior–decoder training. AuDiAE studies an audio-specific fixed-noise channel, a marginal/conditional rate mixture, clean-target waveform flow matching, and explicit finite-boundary, mismatch, and variational-slack accounting. Its two training paths share the deterministic encoder output and endpoint law, not a particular noisy sample.

GenFirst analyzes posterior entropy and asymmetric generation–reconstruction optimization, motivating a generation-before-reconstruction strategy (Zheng et al., 2026). Its optimization perspective complements the present focus on rate-induced representation preferences. Evaluating a separately initialized generator on frozen latents is essential in this setting: it distinguishes transfer under a common generator contract from the fit of the jointly trained prior, while still leaving native waveform decoders as part of the end-to-end audio system.

B Derivations and Scope

Table 4 provides a compact reference for the notation used in the derivations below.

B.1 Marginal and Conditional Rate Decompositions

The caption feature C introduced in Section 3 induces the data law for (X, C). All expectations below use the joint law $q(x, c, z) = p_{\mathrm{data}}(x, c)q^b_\phi(z \mid x)$. Assume the relevant KL divergences and mutual informations are finite. Inserting the aggregate posterior into the previously defined marginal rate gives

$$R_u = \mathbb{E}_q \log \frac{q^b_\phi(z \mid x)}{q^b_\phi(z)} + \mathbb{E}_q \log \frac{q^b_\phi(z)}{p_u(z)} = I(X; Z_b) + M_u. \tag{17}$$

Expanding the previously defined conditional rate with $q^b_\phi(z \mid c)$ instead gives

$$R_c = \mathbb{E}_q \log \frac{q^b_\phi(z \mid x)}{q^b_\phi(z \mid c)} + \mathbb{E}_q \log \frac{q^b_\phi(z \mid c)}{p_c(z \mid c)} = I(X; Z_b \mid C) + M_c. \tag{18}$$

The first term is conditional mutual information because $q^b_\phi(z \mid x, c) = q^b_\phi(z \mid x)$. The chain rule then yields

$$I(X; Z_b) = I(X, C; Z_b) = I(C; Z_b) + I(X; Z_b \mid C). \tag{19}$$

Combining the two rate decompositions with $G_s = \overline{R}_s - R_s \geq 0$ proves Equation 10. In particular,

$$\overline{R}_\rho \geq (1-\rho) I(X; Z_b) + \rho I(X; Z_b \mid C) \geq (1-\rho) I(X; Z_b). \tag{20}$$

This explains why the mixed value cannot directly be reported as an unconditional bitrate. Differences between $\overline{R}_u$ and $\overline{R}_c$ also contain differences in prior mismatch and bound slack; they do not directly measure $I(Z_b; C)$.

Diffusion rate estimator. For two log-SNR levels $a \leq \lambda < \lambda' \leq b$, the variance-preserving forward Markov process, running from b toward a, has transition

$$q(z_\lambda \mid z_{\lambda'}) = \mathcal{N}\left(\frac{\alpha_\lambda}{\alpha_{\lambda'}} z_{\lambda'}, \left[1 - \frac{\alpha_\lambda^2}{\alpha_{\lambda'}^2} \right] I \right). \tag{21}$$

At b, its single-level conditional marginal given X is the Gaussian in Equation 3. The estimator samples η_λ at a random level λ, independently of the endpoint draw η_b supplied to the decoder; neither denotes one shared noise realization along an entire Markov path. For branch $s \in \{u, c\}$, let $\chi_u = \varnothing$ and $\chi_c = C$. Let W_λ be standard d-dimensional Brownian motion. The reverse prior is the endpoint law at b of the Itô diffusion

$$dZ_\lambda = \left[\tfrac{1}{2} \sigma_\lambda^2 Z_\lambda - \sigma_\lambda \epsilon_\psi(Z_\lambda, \lambda, \chi_s) \right] d\lambda + \sigma_\lambda \, dW_\lambda, \qquad Z_a \sim \mathcal{N}(0, \mathbf{I}), \tag{22}$$

where λ increases from a to b. This specifies the generative process whose score parameterization is $-\epsilon_\psi / \sigma_\lambda$. Under the usual absolute-continuity and finite-energy assumptions, path-space KL and marginalization give the stated diffusion bound. The shared high-noise boundary is B_a from Equation 7. For either text condition, the diffusion variational construction gives $R_s \leq B_a + \frac{1}{2} \int_a^b \mathbb{E} \|\eta_\lambda - \epsilon_\psi(Z_\lambda, \lambda, \chi_s)\|^2 d\lambda$. Equation 8 is the marginal importance-sampling estimator; replacing the null condition by C yields its conditional counterpart. The derivation concerns the specified reverse process at b. A finite-step approximate sampler need not have that same distribution. All dimensions, boundary terms, sampling densities, and duration normalizations are required when reporting variational bits per second.

B.2 Latent Scale and Stochastic Precision

For a centered encoder coordinate, Equation 3 gives $\mathrm{SNR}_j = e^b \operatorname{Var}(U_j)$. Replacing U by kU for a scalar $k > 0$ while keeping the additive noise fixed changes the stochastic channel. In contrast, applying an invertible affine map to the complete Z_b transforms both signal and noise and preserves $I(X; Z_b)$. The latter transformation can change numerical conditioning without removing information learned by the encoder.

A scalar Gaussian example separates these effects analytically. Let one encoder coordinate satisfy $U_j \sim \mathcal{N}(0, \nu)$ and use the corresponding scalar Gaussian channel. The expected KL to a standard Gaussian prior is

$$R_G(\nu) = \frac{1}{2} \left[\alpha_b^2 \nu + \sigma_b^2 - 1 - \log \sigma_b^2 \right]. \tag{23}$$

For the optimal diffusion denoiser, the noise-prediction MSE at log-SNR λ equals $e^\lambda \nu / (1 + e^\lambda \nu)$. Integrating it and including B_a gives

$$\overline{R}_D^*(\nu) = \frac{1}{2} \log(1 + e^b \nu) + \frac{1}{2} \left[V_a - 1 - \log V_a \right], \qquad V_a = \alpha_a^2 \nu + \sigma_a^2. \tag{24}$$

The first term is $I(U_j; Z_{b,j})$; the second is the mismatch at the finite-noise initialization. It vanishes as $a \to -\infty$ for fixed ν, but cannot in general be discarded at a finite endpoint. This example shows why learned and fixed priors can impose different costs on scale without identifying large variance with a difficult conditional distribution.

If prior fitting additionally covers $\lambda \in (b, h]$, that supervision can change the learned denoiser but does not enlarge the encoder's rate integral for Z_b. Likewise, a prior sampler used to supply the decoder ends at b to match its training channel. Using h as the sampling endpoint when $h > b$ changes the latent precision and no longer matches the decoder's training channel.

B.3 An Information Preference without a Feature-Alignment Target

For a finite discrete source $\widetilde{X}$, a matched marginal prior, and an exact conditional log-loss decoder, the ideal objective is

$$H(\widetilde{X} \mid Z) + \beta I(\widetilde{X}; Z) = H(\widetilde{X}) + (\beta - 1) I(\widetilde{X}; Z). \tag{25}$$

It does not distinguish two latents with the same encoded information. Under the conditional rate mixture, the corresponding expression becomes

$$H(\widetilde{X}) + (\beta - 1) I(\widetilde{X}; Z) - \beta\rho I(Z; C). \tag{26}$$

Thus the condition introduces a preference that is absent from the marginal rate alone. This identity concerns ideal distributional log-loss; it is not an exact expression for the practical acoustic and denoising losses.

For an explicit example, let $\widetilde{X} = (Q, N)$ contain two independent fair bits, let the text condition be $C = Q$, and let B_ε be an independent Bernoulli noise bit with probability $\varepsilon \in (0, 1/2)$. Compare $Z_Q = Q \oplus B_\varepsilon$ and $Z_N = N \oplus B_\varepsilon$. Writing $H_\varepsilon = h_2(\varepsilon)$ for binary entropy, define the matched-prior mixed rate $R_\rho^{\mathrm{match}} := (1-\rho) I(\widetilde{X}; Z) + \rho I(\widetilde{X}; Z \mid C) = I(\widetilde{X}; Z) - \rho I(Z; C)$. The resulting costs in bits are

Latent	$I(\widetilde{X}; Z)$	$I(Z; C)$	$H(\widetilde{X} \mid Z)$	R_ρ^{match}
Z_Q	$1 - H_\varepsilon$	$1 - H_\varepsilon$	$1 + H_\varepsilon$	$(1-\rho)(1 - H_\varepsilon)$
Z_N	$1 - H_\varepsilon$	0	$1 + H_\varepsilon$	$1 - H_\varepsilon$

Both latents have the same marginal distribution and information quantity. Only Z_Q carries information about the condition, so it receives a lower mixed rate. These are analytic toy quantities, not measured audio results. The example establishes a preference rather than a computational advantage: an independent fair bit is itself easy to generate. An audio representation also needs to preserve factors omitted by its caption when those factors are needed for faithful synthesis.

For a finite normalized log-loss decoder $p_{\mathrm{dec}}(\widetilde{x} \mid z)$ under the induced source–latent law q_{toy}, define $D_{\log} := \mathbb{E}[-\log p_{\mathrm{dec}}(\widetilde{X} \mid Z)]$ and let $A_{\mathrm{dec}} = \mathbb{E}_Z \, \mathrm{KL}(q_{\mathrm{toy}}(\widetilde{X} \mid Z) \| p_{\mathrm{dec}}(\widetilde{X} \mid Z))$. The corrected objective is

$$D_{\log} + \beta \overline{R}_\rho = H(\widetilde{X}) + (\beta - 1) I(\widetilde{X}; Z) - \beta\rho I(Z; C) + A_{\mathrm{dec}} + \beta \mathcal{E}_\rho. \tag{27}$$

A variational decoder bound introduces additional slack. These corrections explain why a lower training objective alone cannot identify more encoded information or improved downstream generation.

B.4 Denoising Optima and the Value of a Latent Representation

Recall from Section 3.3 that $S = A(X)$, where A is fixed independently of the learned model and may be the identity. Assume finite second moments and a positive, integrable weighted risk on the chosen waveform noise interval. At any fixed t, let $\mu_Z = \mathbb{E}[S \mid S_t, t, Z]$. For a measurable predictor F,

$$\frac{1}{d_x} \mathbb{E} \|S - F\|_2^2 = \frac{1}{d_x} \mathbb{E} \|S - \mu_Z\|_2^2 + \frac{1}{d_x} \mathbb{E} \|F - \mu_Z\|_2^2. \tag{28}$$

The cross term vanishes by conditional expectation. Multiplying by a positive time-only weight and integrating establishes the Bayes predictor and $\mathcal{D}_w^*(Z)$ used in Section 3.3. Since $\xi = (S_t - tS)/(1-t)$ for $t < 1$,

$$\mathbb{E}[S - \xi \mid S_t, t, Z] = \frac{\mathbb{E}[S \mid S_t, t, Z] - S_t}{1 - t}. \tag{29}$$

This is the conditional marginal flow velocity. Exact generation additionally requires the appropriate endpoints, regularity of the field, and accurate integration; denoiser optimality alone is not a finite-step sampling theorem.

Let $\mu_\varnothing = \mathbb{E}[S \mid S_t, t]$. At $t = t(\ell)$, define the coordinate-mean Bayes risks $m_Z(\ell) = d_x^{-1} \mathbb{E} \|S - \mu_Z\|_2^2$ and $m_\varnothing(\ell) = d_x^{-1} \mathbb{E} \|S - \mu_\varnothing\|_2^2$. A second conditional variance decomposition gives

$$m_\varnothing(\ell) - m_Z(\ell) = \frac{1}{d_x} \mathbb{E} \|\mu_Z - \mu_\varnothing\|_2^2 \geq 0. \tag{30}$$

Integrating gives the nonnegative value in Equation 14:

$$V_w(Z) = \frac{1}{2}\int_{\ell_-}^{\ell_+} w(\ell)e^{\ell}[m_{\varnothing}(\ell) - m_Z(\ell)]\, d\ell. \tag{31}$$

The reference risk is independent of the encoder because S and its noise process are fixed. Substitution of Equation 10 then yields Equation 15. For a finite waveform model, the nonnegative excess risk $A_w(F, Z) = \mathcal{D}_w(F, Z) - \mathcal{D}_w^*(Z)$ appears as an additional term; it can change with both the architecture and the learned representation.

Adding MRSTFT supervision, the finite-model objective is

$$\begin{aligned}\mathcal{J} = \mathcal{D}_w^*(\varnothing) - V_w(Z_b) + \beta I(X; Z_b) - \beta\rho I(Z_b; C)\\ + \beta\mathcal{E}_\rho + A_w(F_\theta, Z_b) + \kappa_{\mathrm{STFT}}\mathcal{L}_{\mathrm{MRSTFT}}.\end{aligned} \tag{32}$$

The first term is independent of the encoder. The remaining approximation and acoustic terms explain why the ideal information preference does not determine the outcome of finite-model optimization by itself.

For the I–MMSE interpretation, rescale the observation as $S_t/(1-t) = \sqrt{\gamma}\, S + \xi$, where $\gamma = e^{\ell}$. Applying the Gaussian-channel identity (Guo et al., 2005) conditionally on Z gives

$$\frac{d}{d\gamma} I(S; \sqrt{\gamma}\, S + \xi \mid Z) = \frac{d_x}{2} m_Z(\log \gamma). \tag{33}$$

Thus the optimal denoising risk integrates information increments across noise levels with weight w/d_x. Finite endpoints contribute endpoint terms; this is not an identification of the risk with an unrestricted entropy. Noise levels also do not specify semantic categories or acoustic frequency bands.

B.5 Prediction Space, Sampling Density, and Effective Weight

For the linear waveform path, define $\widehat{\xi}_\theta = (S_t - t\widehat{S})/(1-t)$ and $\widehat{v}_\theta = (\widehat{S} - S_t)/(1-t)$. The samplewise identities are

$$\|\widehat{\xi}_\theta - \xi\|^2 = e^{\ell}\|\widehat{S} - S\|^2, \qquad \|\widehat{v}_\theta - (S - \xi)\|^2 = \frac{\|\widehat{S} - S\|^2}{(1-t)^2}. \tag{34}$$

These identities hold for the same predictor and input. Replacing a network with a different bare output parameterization can change finite-capacity optimization even when its outputs are algebraically convertible.

For a normalized density π_x on $[\ell_-, \ell_+]$, sampling $\ell \sim \pi_x$ assigns the coordinate-mean clean-target error the multiplier

$$a_x(\ell) = \frac{w(\ell)e^{\ell}}{2\pi_x(\ell)}. \tag{35}$$

Let π_t denote the induced sampling density of t. In the t coordinate, where $\ell'(t) = 2/[t(1-t)]$, the multiplier is

$$a_x(t) = \frac{w(\ell(t))e^{\ell(t)}\ell'(t)}{2\pi_t(t)}. \tag{36}$$

A scalar normalization of this weight also rescales the synthesis term relative to rate pressure, so the normalization should be reported. Equation 13 fixes the waveform reduction to a coordinate mean, removing a dimension-dependent ambiguity from this accounting.

For the reference $\mathrm{logit}(t) \sim \mathcal{N}(0, 1)$, the interior log-SNR density is $\pi_x(\ell) = \mathcal{N}(\ell; 0, 4)$. Direct coordinate-mean clean-signal MSE therefore has effective integral weight $2\pi_x(\ell)e^{-\ell}$. A sigmoid multiplied directly into this MSE is a different risk from a sigmoid specified in the noise-error integral. Endpoint clipping, importance-weight clipping, and normalization by a random minibatch mean also change the exact objective or estimator. Controlled comparisons specify these choices and use a fixed population normalization when matching an integral risk.

The general waveform risk in Equation 13 is a denoising objective. A likelihood interpretation requires a specified diffusion generative process and its boundary terms. Under the applicable

assumptions, monotone weighting admits an ELBO interpretation for noise-augmented data (Kingma & Gao, 2023); it does not certify that a particular sigmoid is optimal for audio. The MRSTFT term in Equation 16 lies outside this squared-error analysis.

Table 4: **Notation reference.** Main-text notation and appendix-only indices. Latent and waveform noise processes use different coordinates and independent noise.

Symbol	Definition
X, Y, C	Audio, paired caption, and frozen caption feature $T_{\text{text}}(Y)$
ϕ, ψ, θ	Parameters of the encoder, rate prior, and waveform decoder
U, Z_b	Deterministic encoder output and stochastic decoder-endpoint condition
$q_\phi(u \mid c), q_{\text{ch}}(z \mid u)$	Encoder-output law and prescribed endpoint channel
$q_\phi^b(z \mid x), q_\phi^b(z \mid c), q_\phi^b(z)$	Endpoint posterior, conditional aggregate, and marginal aggregate
T_z, d_z, d	Latent time steps, channels, and total coordinates $d = T_z d_z$
λ, a, b, h	Latent log-SNR; high-noise start, decoder endpoint, prior-fitting upper endpoint
$\alpha_\lambda, \sigma_\lambda$	Variance-preserving signal/noise coefficients; squared sum is one
$\eta_b, \eta_\lambda, \pi_z$	Independent endpoint/path noise draws and log-SNR sampling density
p_u, p_c	Marginal and conditional reverse diffusion priors at b
$s \in \{u, c\}, \chi_s$	Rate branch and its prior-conditioning input
$R_s, \overline{R}_s, M_s, G_s$	Expected KL, variational bound, aggregate-prior mismatch, and bound slack
J_s, B_a	Diffusion path cost and finite high-noise boundary
$\rho, \mathcal{E}_\rho$	Conditional-rate mixture weight and mixed mismatch/slack
A, S, d_x	Fixed waveform preprocessing, target $A(X)$, and waveform dimension
t, ℓ, ℓ_-, ℓ_+	Waveform flow time, log-SNR $2\,\text{logit}(t)$, and integration limits
ξ, π_x, w, a_x	Waveform noise, log-SNR density, integral weight, and sampled-loss multiplier
$\mathcal{D}_w, V_w, A_w$	Denoising risk, optimal latent value, and finite-predictor excess risk
$\mathcal{L}_{\text{FM}}, \mathcal{L}_{\text{MRSTFT}}, \kappa_{\text{STFT}}$	Clean-target loss, spectral loss, and coefficient
$\beta, \beta_{\text{mean}}$	Summed-rate and element-mean coefficients; $\beta = \beta_{\text{mean}}/d$

C Implementation and Evaluation Protocols

This section records the contracts behind the main reconstruction, generation, and human-study results: training and evaluation populations, model interfaces, optimization, inference, preprocessing, and aggregation. These details make the reported comparisons auditable; they do not introduce a separate model-selection claim.

C.1 Data, Architecture, and Optimization

Training populations. Stage 1 prior and rate-pressure studies use nominal three-second crops from FMA (Defferrard et al., 2017) and WavCaps (Mei et al., 2024). We use the Freesound (Font et al., 2013), SoundBible, and AudioSet Strongly-Labelled (Gemmeke et al., 2017) partitions of WavCaps and exclude its BBC Sound Effects partition. Stage 2 uses captioned examples from the same non-BBC WavCaps selection and AudioCaps (Kim et al., 2019), sampled in the ratio $59 : 5$. Conditional and shuffled-caption comparisons reuse the same audio population and crop policy. A caption permutation changes the pairing only; it does not change the audio batches.

Primary architecture. The encoder uses 128 base channels and a six-layer query stack. AuDiAE uses $(r, d_z) = (2048, 192)$, while AuDiAE-H uses $(4096, 256)$; these shapes correspond to approximately 21.5 and 10.8 latent frames per second at 44.1 kHz. The rate prior is a four-block, width-512 DiT (Peebles & Xie, 2023) with eight heads and RoPE (Su et al., 2024). The waveform decoder is a 12-block, width-1024 DiT with 16 heads and predicts the clean waveform target. The primary configuration uses $(a, h, b) = (-10, 5, 5)$ and $\beta_{\text{mean}} = 10^{-3}$. Conditional Stage 2 sets $\rho = 0.5$ and uses 2,048-dimensional features from a frozen Qwen3-1.7B text encoder (Yang et al., 2025). Text is provided to the prior and never to the waveform decoder.

External systems use released continuous-latent checkpoints and native decoders. Stable Audio Open, ϵar-VAE, and SAME-L use their published temporal interfaces. CoDiCodec uses its continuous output and parallel consistency decoder; its summary embeddings give a nominal $\sim 4096\times$ sequence interface but are not individually time-aligned. Moûsai uses its released diffusion autoencoder and native iterative decoder. Temporal downsampling and latent width are reported separately because neither is an entropy-coded bitrate.

Full-clip training pads or crops inputs to 450,560 samples, uses global batch 64, and updates the autoencoder with AdamW (Loshchilov & Hutter, 2019) at 2×10^{-5}. Unless a controlled screen

Table 5: **Evaluation populations.** Primary objective metrics are averaged within dataset and then equally over the two datasets in each sound/music domain. The MOS panel draws ten items from each of all six listed datasets.

Role	Domain / dataset	Population	Scoring unit
Primary	Sound / AudioCaps (Kim et al., 2019)	975 recordings	Recording
Primary	Sound / VGGSound (Chen et al., 2020)	14,483 recordings	Recording
Primary	Music / Song Describer (Manco et al., 2023)	706 recordings	Recording
Primary	Music / MusicCaps (Agostinelli et al., 2023)	5,352 recordings	Recording
Listening	Speech / VCTK (Yamagishi et al., 2019)	2,160 recordings	Recording
Listening	Speech / EXPRESSO (Nguyen et al., 2023)	648 recordings	Recording

states otherwise, the selected two-stage models start Stage 2 from the 300k-update Stage 1 EMA checkpoint and receive 600k Stage 2 updates. The encoder, rate prior, and waveform decoder are optimized jointly as described in Section 3.4; reported reconstruction uses EMA weights. The rate follows the element-mean convention in Section 3.4. Condition dropout selects conditional or null prior inputs and is distinct from the mixture weight ρ. The high-noise decoder variant multiplies the clean-target squared error by $1 + 0.5\tanh(-\ell/2)$ while keeping the decoder architecture, output parameterization, and MRSTFT fixed.

Matched fixed-Gaussian reference control. Table 3(a) retains the deterministic encoder output U and prescribed endpoint channel $q_{\mathrm{ch}}(z_b \mid U) = \mathcal{N}(\alpha_b U, \sigma_b^2 I)$. Its control replaces the learned diffusion bound by the exact element-mean KL to the fixed reference $\mathcal{N}(0, I)$, $\frac{1}{2d}[\alpha_b^2 \mathbb{E}\|U\|^2 + d(\sigma_b^2 - 1 - \log \sigma_b^2)]$. The two systems are trained separately with matched architecture, data, endpoint, rate coefficient, optimization exposure, decoder, and downstream generator contract. This isolates the rate-reference choice within the fixed channel; it is not a conventional VAE control with a learned posterior variance.

Evaluation populations. Table 5 gives the objective-metric and listening cohorts. AudioCaps contains one waveform for each of the 975 available official test recordings. VGGSound contains 14,483 valid test recordings after removing unavailable files and source videos shared with AudioCaps or MusicCaps. Song Describer uses the fixed 706-recording cohort, and MusicCaps contains 5,352 available recordings. Rate-pressure controls use frozen subsets under a separate analysis manifest; their rFAD uses the matched 975-recording AudioCaps cohort. The listening panel additionally uses the fixed VCTK speaker-balanced cohort of 2,160 mic2 recordings and the 648-recording EXPRESSO test cohort. It selects ten items from each of these two speech datasets and from each of the four primary sound/music datasets, yielding the balanced 60-source panel. Before training, all six cohorts are fully de-duplicated against the union of the complete Stage 1 and Stage 2 manifests using dataset-native source identifiers and exact audio fingerprints after canonical decoding. WavCaps is expanded into its constituent records before matching, and every matched training item is removed. Thus no evaluation recording occurs in either stage's training audio.

C.2 Reconstruction Protocol and Metrics

Inference interface. The benchmark uses the deterministic encoder output U and does not add endpoint noise before decoding. The waveform flow is integrated on $[0, 0.36, 0.66, 0.94, 1]$ with a third-order Adams–Bashforth multistep solver, requiring four decoder evaluations. References are downmixed to mono. Released models retain their native sample rates and channel interfaces; stereo-only models receive duplicated mono inputs, and their reconstructed channels are averaged before scoring. Waveforms preserve natural amplitude for paired fidelity metrics. Full valid recordings are scored; inputs longer than the decoder context are processed in overlapping 30-s windows with a

one-second linear crossfade. AudioCaps rFAD exports use mono 44.1-kHz PCM16 waveforms of exactly ten seconds. Model padding is removed before scoring.

Paired fidelity. For each FFT size n, let $\mathcal{S}_n$ denote the magnitude STFT and P_n its fixed mel projection. We compute

$$d_{\mathrm{STFT}}(x, \widehat{x}) = \frac{1}{|\mathcal{K}_{\mathrm{STFT}}|} \sum_{n \in \mathcal{K}_{\mathrm{STFT}}} \mathrm{mean} \left| \log(1 + \mathcal{S}_n(x)) - \log(1 + \mathcal{S}_n(\widehat{x})) \right|, \tag{37}$$

$$d_{\mathrm{MEL}}(x, \widehat{x}) = \frac{1}{|\mathcal{K}_{\mathrm{MEL}}|} \sum_{n \in \mathcal{K}_{\mathrm{MEL}}} \mathrm{mean} \left| \log(1 + P_n\mathcal{S}_n(x)) - \log(1 + P_n\mathcal{S}_n(\widehat{x})) \right|. \tag{38}$$

Here $\mathcal{K}_{\mathrm{STFT}} = \{128, 256, 512, 1024, 2048, 4096\}$ and $\mathcal{K}_{\mathrm{MEL}} = \{1024, 2048, 4096\}$. We use periodic Hann windows, hop $n/4$, centered reflect padding, unnormalized magnitude spectra, and 64 HTK mel bands from zero to Nyquist. SI-SDR (Le Roux et al., 2019) is computed on paired natural-amplitude waveforms and reported in decibels.

Distributional fidelity. Fréchet Audio Distance (Kilgour et al., 2019) uses LAION-CLAP audio embeddings (Wu et al., 2023). We write rFAD for reconstructions and FAD for generated samples. Primary reconstruction rFAD is computed separately on the 975 AudioCaps and 706 Song Describer recordings with identical prediction and reference ID sets, then equally averaged in Table 1. Controlled analyses that use AudioCaps-only rFAD state that contract explicitly. rFAD is not used in place of paired fidelity.

Human listening study. Twelve listeners evaluated a fixed panel of 60 source items: 20 each from speech, music, and sound, comprising ten items from each of the six datasets in Table 5. The item list was selected without reference to model outputs and shared by every system. We used a balanced incomplete-block assignment at the source level: each source was assigned to five listeners, and each assigned listener rated the outputs of all seven systems in Table 1. Thus every item–system pair received five scores, while each listener rated 25 sources (175 trials) in two sessions containing 12 and 13 sources.

Only listeners who passed a headphone check proceeded to a short practice round and the two scoring sessions. Overall audio quality was rated on an integer scale from 1 (bad) to 5 (excellent), without access to system identity or model metadata. Trial order was independently randomized within each session, with the same rating question throughout.

For each system and item, the five listener scores were averaged to obtain an item-level MOS. Table 1 reports the mean and sample standard deviation of these 60 item-level values. Equal numbers from the three domains and the two datasets within each domain give every domain and dataset equal weight, while the shared panel makes system contrasts paired at the item level. No automatic MOS predictor or source-consistency rating is included. We treat the result as descriptive perceptual evidence and report no hypothesis test or claim of statistical significance between systems.

Scope of the protocol. The six cohorts have informed development and are not treated as blind test sets. This limits claims of blind-test generalization but does not create train–evaluation contamination: the cohorts are disjoint from all training audio under the manifest and fingerprint procedure above.

C.3 Downstream Generation Protocol

Each autoencoder is frozen before fitting its downstream generator. The generator is a bidirectional, non-autoregressive flow-matching Transformer initialized from the same Qwen backbone for every representation. It models the complete deterministic latent sequence U without a true-latent prefix. A channel-wise population mean and one scalar population variance define the fixed latent normalization. Sampling applies its inverse once and passes the result directly to the waveform decoder, without endpoint noise.

Generator training uses 45,044 AudioCaps training items and nominal ten-second waveforms. Every representation uses global batch size 64, seed 4396, a learning rate of 10^{-4} with 2,000 warm-up updates, and a 200k-update budget. Native latent lengths and widths differ, so equal updates match

data exposure rather than FLOPs. Each representation is represented by one generator training run; the comparison does not quantify generator-training variation.

AudioCaps generation uses the 4,875 official captions associated with 975 test recordings; each caption receives one ten-second sample. The completed common suite reports CLAP over all 4,875 outputs. FAD uses the fixed common-reference intersection of 879 recordings, or 4,395 caption–output pairs. The remaining 96 recordings (480 caption–output pairs) are absent from that reference intersection and are excluded identically for every row before scoring, without reference to model outputs. Sampling uses EMA weights, 24 latent-field evaluations, classifier-free guidance (Ho & Salimans, 2022) scale 5, and each autoencoder's native decoder; AuDiAE and AuDiAE-H use four waveform-decoder evaluations. Input and output projections adapt to each latent dimension. Main generation results report LAION-CLAP FAD and text–audio CLAP. A valid comparison requires the same caption and reference manifests within a dataset and a complete output population.

D Controlled Evidence for the Learned Bottleneck

The controls are organized by the claim they can support. Frozen-representation prior fitting directly tests density fit without changing encoded information. Conditional-training and decoder-objective screens instead diagnose whether an end-to-end benefit can be localized to the representation or decoder objective.

D.1 Frozen-Representation Prior Fit

Table 6 freezes the $2048\times$, 192-channel AuDiAE encoder and stochastic channel, fits each reference density on the same latent training split, and evaluates the same held-out cohort. This isolates prior fit because the representation and its mutual information do not change. The conventional KL baseline instead learns a diagonal-Gaussian posterior and therefore changes both the posterior family and the information constraint; a broader system comparison requires common maturity, rate–distortion matching, and separately trained generators. Table 3(a) asks the complementary end-to-end question under the matched fixed-channel contract. Together, the frozen comparison identifies tighter reference fit and the joint comparison measures its system-level consequence; neither establishes that KL regularization or Gaussian VAEs are generally inferior.

Table 6: **Frozen-representation prior fit.** All rows use the same encoder, stochastic channel, latent split, and held-out cohort. Gaussian rates are exact; the diffusion result is a variational upper bound.

Reference density	Dependence represented	Reported quantity	Nats / latent element ↓
Standard Gaussian	Fixed isotropic	Exact rate R_p	6.953
Fitted diagonal Gaussian	Coordinatewise mean / scale	Exact rate R_p	3.479
Fitted full-channel Gaussian	Within-step channel covariance	Exact rate R_p	3.264
Learned diffusion prior	Channel and temporal dependence	Bound $\overline{R}_u$	**3.104**

Table 7: **Frozen-prior text-use diagnostics.** Mismatched minus paired rate contrast; a positive value supports use of the paired caption. The encoder and stochastic channel are fixed in both rows.

Contract	Variant	Updated	Rate contrast $\times 10^3$
Frozen	Shared conditional prior	Prior only	$0.276\ [-0.167, 0.683]$
Frozen	Residual condition adapter	Adapter only	$\mathbf{3.79}\ [2.43, 5.27]$

Table 7 tests text use without changing the encoder or stochastic channel. For the shared conditional prior, the mismatched-minus- paired rate contrast is 0.276×10^{-3} nats per element and its interval spans zero. Updating only a residual condition adapter yields a positive contrast of 3.79×10^{-3} with interval $[2.43, 5.27] \times 10^{-3}$. Thus paired text is detectably used under the adapter contract, but the shared- prior result alone is inconclusive; neither comparison tests whether joint training reshapes the representation.

D.2 Conditional-Training and Decoder-Objective Diagnostics

The remaining conditional analysis asks whether the end-to-end downstream benefit in Table 3(b) is already visible in the jointly trained representation at a common checkpoint. The separate frozen-representation decoder screen then tests a local objective change.

Table 8: **Joint conditional-training reconstruction diagnostic.** Conditional minus marginal at a common checkpoint. Negative favors conditional training for the rate and error metrics; positive favors SI-SDR. The interval applies to the rate contrast.

Updated	**Rate contrast** $\times 10^3$	**ΔMEL** ↓	**ΔSI-SDR** ↑	**ΔrFAD** ↓
Encoder / prior / decoder	$0.602\ [-0.111, 1.337]$	-1.8×10^{-5}	$+0.010$	$+2.49\times10^{-4}$

Table 8 reports the separate same-checkpoint joint diagnostic. Its interval spans zero and the paired metrics are unchanged at the displayed precision. Table 3(b) adds the matched downstream comparison: paired-caption training improves FAD/CLAP over both the marginal and shuffled-caption controls under that system contract. Because the encoder, prior, and decoder are updated together and each representation has one generator run, this does not isolate a representation mechanism or seed-level effect.

Table 9: **Frozen-representation decoder-objective screen.** The encoder and rate prior are fixed, so only the decoder objective changes.

Decoder objective	$\mathrm{MEL}_{\mathrm{log1p}}$ ↓	$\mathrm{STFT}_{\mathrm{log1p}}$ ↓	**SI-SDR** ↑	**rFAD** ↓
Baseline clean-target loss	0.029	0.047	**14.79**	**0.077**
High-noise-weighted loss	**0.028**	0.047	14.78	0.081

Table 9 instead freezes the representation and rate prior: high-noise weighting changes paired metrics only slightly and raises rFAD. This negative diagnostic does not support selecting the reweighted decoder objective as an overall improvement.

The full rate-by-decoder grid does not share a complete common endpoint, and matched downstream results for the decoder-objective cells are unavailable. We therefore report neither a decoder-objective main effect nor a rate–decoder interaction.

E Limitations and Evidence Boundaries

Causal scope. The frozen-representation prior comparison establishes tighter density fit under its matched encoder and channel contract. The residual-adapter diagnostic establishes paired-text use under its frozen contract, whereas the shared-prior interval spans zero. Table 3(b) additionally supports a contract-specific downstream benefit of paired-caption joint training, but the joint updates do not isolate a representation mechanism. The decoder-objective screen is negative: it does not support a causal advantage for the reweighted objective or for generative decoding over a matched deterministic decoder.

System comparisons. Tables 1 and 2 compare complete systems with different latent shapes and native decoders. They support the reported operating points, but do not isolate the learned prior, decoder family, or latent width. Equal generator updates match data exposure, not FLOPs. The matched fixed-channel standard-Gaussian control in Table 3(a) is not a conventional learned-variance KL-VAE; no mature rate–distortion-matched conventional KL system is available as an end-to-end control.

Evaluation scope. The reconstruction cohorts are disjoint from training audio but informed model development and are not blind test sets. Each downstream representation uses one generator run, so generator-training variation is not quantified. The 60-item listening study is descriptive and reports no hypothesis test or listener-subgroup analysis. The current evidence therefore does not support claims of seed-level robustness or blind-test generalization.

Efficiency scope. The study does not report parameters, FLOPs, latency, and peak memory for every system under one hardware and precision contract. Temporal compression and latent width should therefore be read as interface descriptors, not as complete measures of storage bitrate or end-to-end efficiency.

E.1 Ethics Statement

The only human-participant component is the MOS listening study in Appendix C.2. Participation was voluntary and limited to perceptual quality ratings of the fixed, balanced 60-source panel. Each source was assigned to five of the 12 listeners, and every assigned listener rated all seven system outputs for that source in two sessions. Listeners passed a headphone check, system identities were hidden, and presentation order was randomized. We collected only five-point quality scores: no names, contact details, demographic attributes, listener audio, or other personally identifying information were collected. Ratings are reported only as aggregate statistics, and no participant-level results are reported or released. The study makes no claims about listener subgroups or individual speakers.