

HOW MUCH AUDIO IS LEFT IN AN EMBEDDING? AN INVERSION AUDIT OF AUDIO ENCODERS

Marios Glytsos Brian McFee

Music and Audio Research Laboratory, New York University, NY, USA

ABSTRACT

Pretrained audio encoders are reused for downstream tasks that are often unknown when the encoder is trained, so their usefulness depends partly on which signal properties survive the pretext objective. We study this retained information through paired source reconstruction. Using a shared Stable Audio Open latent diffusion decoder, we reconstruct five-second, 44.1-kHz stereo music from frozen representations produced by supervised classifiers (VGGish, ConvNeXt), an audio-text contrastive model (CLAP), and a waveform-reconstruction model (EnCodec). These objectives impose different pressures to preserve source detail, while their exposed interfaces vary substantially in temporal and spectral resolution. Evaluating on the Million Song Dataset (MSD), we find clear differences in reconstructability across encoder families, while within encoder comparisons show improved recovery when finer temporal or spectral structure is exposed. Even compressed task oriented embeddings support reconstructions that preserve measurable source specificity and high level musical content.

Index Terms— audio embedding inversion, representation audit, latent diffusion, reconstruction, audio representations

1. INTRODUCTION

Pretrained audio encoders increasingly serve as reusable interfaces between audio and downstream systems. Unlike traditional engineered features, learned embeddings range from highly compressed task oriented vectors to dense representations trained for reconstruction, making it unclear how much source information survives these different abstractions. This matters when representations are reused for tasks unknown during training, and when embeddings are shared instead of audio in privacy or copyright sensitive settings. We therefore ask: how much source information, beyond what the training objective required, survives in the representation?

We study this through paired source reconstruction. Given a frozen encoder Φ and representation $\mathbf{e} = \Phi(\mathbf{x})$, we train a decoder to reconstruct the paired source $\mathbf{x}$. Because Φ need not be injective, exact deterministic inversion is not always well defined. A conditional generative decoder is therefore

natural. It can use the embedding for source structure that is preserved while relying on its learned audio prior for ambiguous details. Successful reconstruction provides evidence that source specific information is accessible from the embedding. Failure however, may also reflect limitations of the decoder.

We use a standardized reconstruction audit rather than an encoder specific worst case attack. Adaptive inversion can exploit the structure of each target encoder [1, 2], potentially recovering more information but confounding comparisons across representations. Instead, we start from a strong pre-trained generator designed to synthesize audio from highly abstract text conditioning, providing an expressive prior for probing what source information an embedding can support without claiming maximal recoverability.

We compare four representative encoders spanning three training objectives. VGGish [3] provides a long-standing classification baseline, ConvNeXt [4] represents a recent AudioSet trained classifier used as a general purpose audio representation, CLAP [5] is a widely used audio/text representation, and EnCodec [6] provides a reconstruction trained contrast. Their objectives impose different requirements on source detail. Classification rewards discriminative information, contrastive learning emphasizes language aligned semantics, and reconstruction requires local signal detail.

We focus on music because reconstructing a recording requires preserving fine grained pitch, timbre, rhythm, and temporal structure, making it a demanding test of how much source information survives an embedding. Music is also a practically relevant setting because licensing constraints often motivate sharing derived representations rather than the original recordings, and instrumental music already lies within the pretrained decoder’s domain, reducing domain shift.

The paper makes the following contributions: (1) we adapt a pretrained latent diffusion audio generator to paired source reconstruction from frozen embeddings, and (2) we compare classification, contrastive, and reconstruction trained encoders. The comparison quantifies reconstructability and provides caution when embeddings derived from sensitive recordings are distributed in place of audio.

Code, listening examples and more detailed per-genre scores are available at <https://mariosgly.github.io/research/audio-embedding-inversion/>.

2. RELATED WORK

Representation inversion has long been used to study what information learned features retain. In vision, Mahendran and Vedaldi optimize inputs to match target representations [1], while Dosovitskiy and Brox learn a decoder that reconstructs image structure from high-level features [7]. Related work instead frames inversion as a privacy attack [8].

In audio, Comanducci et al. [9] reconstruct speech from intermediate CNN features, while Polyak et al. [10] study resynthesis from discrete self-supervised representations and preservation of content, prosody, and speaker identity. More recent attacks recover audio from intermediate activations [11] or shared gradients [12]. Our goal differs in comparing paired source reconstructability across independently trained representation families using a common pretrained decoder architecture, separately finetuned for each interface.

Audio representations also increasingly serve as conditioning signals for generative models. AudioLDM [13], AudioLDM 2 [14], and Diff-A-Riff [15] show that learned representations can guide realistic, condition-compatible synthesis. These objectives, however, do not require recovery of the particular recording that produced the condition.

A complementary line of work studies representation sensitivity to controlled transformations. Kim et al. [16], Srivastava et al. [17], and Deng et al. [18] examine how known audio changes affect embedding geometry. These methods test which variations are preserved or suppressed, reconstruction instead asks what source-specific information remains jointly recoverable. We apply this view across independently trained representations using a standardized generative probe.

3. EMBEDDING RECONSTRUCTION AUDIT

3.1. Problem formulation

Let $\mathbf{x}$ denote an audio waveform, Φ a frozen encoder, and $\mathbf{e} = \Phi(\mathbf{x})$ its representation. For each encoder, we train a separate conditional generator $G_{\Phi}(\mathbf{e}, \epsilon \mid \theta)$, where θ denotes the generator parameters and $\epsilon \sim \mathcal{N}(0, I)$ the initial diffusion noise. Training uses paired audio and reconstruction is evaluated on held out tracks. All conditions share the same pretrained initialization, generator architecture, target VAE, and training protocol, only the conditioning interface changes.

3.2. Exposed interfaces

Although all targets are fixed five second crops, the encoder interfaces differ substantially in temporal and spectral resolution. We reconstruct fixed five second crops in every condition. Table 1 therefore reports their temporal granularity and the resulting conditioning shape for a five-second crop. We additionally construct alternative CLAP and ConvNeXt interfaces to isolate the effect of temporal and spectral compression while keeping the underlying encoder fixed.

Condition	Temporal granularity	Input shape
VGGish	0.96 s	$128 \times (6-7)$
CLAP 5s	5 s	512×1
CLAP 1s	1 s	512×5
ConvNeXt pooled	≈ 0.32 s	$768 \times (16-17)$
ConvNeXt flattened	≈ 0.32 s	$5376 \times (16-17)$
EnCodec z_q	6.7 ms	128×750

Table 1. Frozen interfaces exposed for one 5 second target. Shapes are feature dimension $\times$ number of conditioning tokens.

VGGish processes mono 16-kHz audio into non-overlapping 0.96-s, 64 bin log-mel patches, producing one 128 dimensional embedding per patch. We use the released PCA-whitened 8-bit output, rescaled from $[0, 255]$ to $[-1, 1]$.

LAION-CLAP [5] processes mono 48-kHz audio with an HTSAT base tower and produces a 512-dimensional embedding. CLAP 5s encodes the full crop once, while CLAP 1s independently encodes five one second segments.

AudioSet ConvNeXt Tiny [4] processes mono 32-kHz audio and produces a final $T \times 768 \times 7$ feature map, yielding 16–17 temporal positions per crop. The *pooled* interface averages the seven frequency positions, while the *flattened* interface preserves them as 5376-dimensional tokens. The original work similarly retains 31 temporal embeddings over ten seconds for audio captioning.

EnCodec [6] processes 48-kHz stereo audio and produces the quantized representation $z_q \in \mathbb{R}^{T \times 128}$ at 150 frames/s, yielding 750 tokens per crop. Our probe conditions only on z_q , omitting the native scale value. EnCodec serves as a reconstruction trained positive control.

3.3. Conditional latent diffusion probe

We initialize each decoder from Stable Audio Open variant 2.0 trained on Creative Commons data [19]. Its fixed variational autoencoder maps five second, 44.1kHz stereo audio to 64 channel latent representations with a downsampling ratio of 2048. The pretrained 24 layer diffusion transformer has width 1536 and 24 attention heads and cross attends to 768 dimensional tokens. The latent target, autoencoder, and transformer architecture are identical across encoders.

We replace the pretrained T5 conditioning interface with each frozen audio representation. Each crop yields a sequence $\mathbf{e} \in \mathbb{R}^{T \times D}$ whose length and dimension depend on the encoder. A learned linear layer maps D to the transformer’s 768 dimensional conditioning space. The pretrained cross attention was learned for T5 features, while the replacement audio features differ in scale, geometry, and temporal density. We therefore initialize the new projection close to zero, with $W_{\text{proj}} \sim \mathcal{N}(0, 10^{-3})$ and $b_{\text{proj}} = 0$, to reduce the initial perturbation introduced by the new conditioning interface. Our experiments indicate more stable optimization and better con-

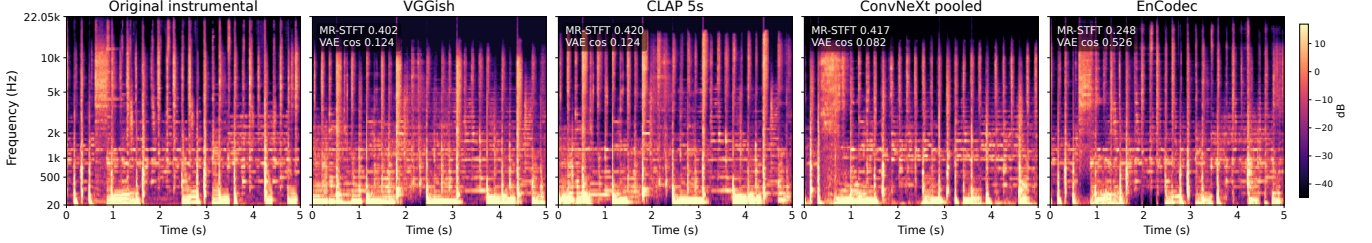


Fig. 1. Qualitative reconstruction of an MTG-Jamendo track using the standard interfaces.

vergence with this initialization.

Dense conditioning also makes temporal correspondence unnecessarily difficult if every latent query can attend equally to every conditioning token. We assign normalized times $t_i, s_j \in [0, 1]$ to latent and conditioning tokens and modify the cross-attention score as

$$a_{ij} = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}} + b_h(\text{bucket}(|t_i - s_j|)). \quad (1)$$

The learned bias is head-specific and initialized to zero. A hard mask also restricts attention to $|t_i - s_j| \leq 0.2$. Together, these terms express the weak prior that corresponding audio regions should remain locally aligned. A query reconstructing 5.0–5.2 s of audio is therefore biased toward conditioning tokens from the same region rather than distant context. This prior also eases finetuning under our limited compute budget, for interfaces with many conditioning tokens, such as EnCodec. The bucket widths and mask radius were set heuristically based on temporal locality, while learning rate schedules were selected using validation-set performance.

3.4. Training objective

Stable Audio Open uses the v -prediction objective. Let $\mathbf{z}_0$ denote the fixed VAE representation of the target audio and $\epsilon \sim \mathcal{N}(0, I)$. At diffusion time τ ,

$$\mathbf{z}_\tau = \alpha_\tau \mathbf{z}_0 + \sigma_\tau \epsilon, \quad \mathbf{v}_\tau = \alpha_\tau \epsilon - \sigma_\tau \mathbf{z}_0, \quad (2)$$

where α_τ and σ_τ are the signal and noise coefficients with $\alpha_\tau^2 + \sigma_\tau^2 = 1$. The diffusion transformer f_θ predicts $\mathbf{v}_\tau$ conditioned on the encoder representation $\mathbf{e}$, and is trained to minimize

$$\mathcal{L}_v = \mathbb{E}_{\mathbf{x}, \tau, \epsilon} \|\mathbf{f}_\theta(\mathbf{z}_\tau, \tau, \mathbf{e}) - \mathbf{v}_\tau\|_2^2. \quad (3)$$

Because this objective does not directly optimize source-waveform similarity, validation loss is used as an optimization diagnostic rather than a measure of reconstruction quality.

Classifier free guidance (CFG) normally trades off conditional and unconditional predictions by randomly dropping the condition during training [20]. This is unnecessary for our

setting, the goal is to reconstruct from the observed embedding rather than trade conditioning fidelity for sample diversity. We therefore use zero conditioning dropout and guidance weight $w = 1$, retaining only the conditional prediction.

At inference, we use a 150-step deterministic DDIM sampler ($\eta = 0$). Randomness therefore enters only through the initial noise realization, no noise is injected during denoising. This reduces sampling variability that would otherwise confound comparisons between paired reconstructions.

3.5. Dataset

We use the Million Song Dataset (MSD) [21] with the Tagtraum CD2 genre annotations [22]. After requiring complete paired audio and features, our artist-disjoint split contains 219,810 training, 26,190 validation, and 26,367 test tracks across 15 genres. Because Stable Audio Open was trained on royalty-free instrumental audio and sound effects [19], we remove vocals with Demucs [23] to prevent the base generator’s domain limitations from bottlenecking the reconstruction audit, and use the resulting instrumental mixtures for both encoder extraction and reconstruction targets.

Implementation details. All models are trained across 4 GPUs with a per device batch size of 32. Finetuning the conditional latent diffusion decoder takes approximately 36 hours per encoder conditioning interface.

3.6. Evaluation

The natural first question is whether the paired source itself can be reconstructed from an embedding. We measure acoustic reconstruction using multi-resolution STFT distance (MR-STFT) and cosine similarity in the fixed Stable Audio Open VAE space. The latter helps distinguish waveform reconstruction error from limitations of the generator representation itself, high VAE similarity with poor MR-STFT indicates that acoustically different signals are represented similarly within the decoder’s domain. We calibrate VAE similarity against 200k randomly sampled distinct source–source crop pairs. Figure 2 compares this reference distribution with the paired source–reconstruction distributions for each encoder. We also retrieve each reconstruction against all 26,367 held-out source crops, R@1 and R@10 report whether its paired

Encoder	STFT↓	VAE↑	R@1↑	R@10↑	Tag↑
VGGish	.21±.08	.15±.07	.004	.021	.733±.144
CLAP 5s	.20±.08	.16±.07	.006	.034	.751±.134
CLAP 1s	.19±.08	.18±.07	.024	.091	.771±.128
ConvNeXt pool.	.18±.08	.19±.07	.094	.201	.799±.121
ConvNeXt flat.	.18±.07	.24±.07	.352	.539	.812±.116
EnCodec	.13±.06	.47±.16	.894	.913	.847±.107
EnCodec native	.10±.05	.89±.05	1.000	1.000	.909±.079

Table 2. Reconstruction on our five-second MSD test crops. Values are mean $\pm$ standard deviation where applicable. R@1/R@10 denote paired-source retrieval. Tag is normalized MusicNN top-10 tag overlap. EnCodec native uses its original decoder.

source is among the nearest one or ten sources in VAE space.

Acoustic mismatch need not imply loss of musical content, so we also compare source–reconstruction pairs using the MSD-trained MusicNN [24]. We report their normalized top 10 tag overlap $|\mathcal{T}_{\text{src}} \cap \mathcal{T}_{\text{rec}}|/10$, as a complementary measure of preserved high-level musical content.

4. RESULTS AND DISCUSSION

Table 2 shows a consistent ordering in reconstructability, from VGGish and CLAP through ConvNeXt to EnCodec. This ordering is consistent with the degree of source detail encouraged by the different pretraining objectives, though the encoder families also differ in architecture and training data. The within encoder comparisons are more direct, finer CLAP temporal resolution and preserving ConvNeXt frequency structure both improve reconstruction, showing that interface compression itself limits recoverability.

EnCodec serves as a positive control for the common probe, while its native decoder provides a topline reference. The common probe approaches native EnCodec in MR-STFT and paired-source retrieval, and listening examples are recognizably close to the source, providing qualitative confirmation that the decoder can effectively use detailed conditioning. At the same time, these reconstructions remain farther away and more variable in SAO-VAE space. Across different models MR-STFT and SAO-VAE similarity are only weakly correlated ($|r| = 0.02\text{--}0.17$, $R^2 < 0.03$), which supports their use as complementary rather than interchangeable reconstruction measures capturing different aspects of the reconstruction.

The complementary metrics reveal that musical content and source specificity are not equivalent. For compressed representations such as VGGish and CLAP, reconstructions may miss exact pitch or timing while still retaining similar timbral texture and mixtures of instruments. The relatively strong MusicNN tag agreement is consistent with this behavior. Paired source retrieval, by contrast, improves much more sharply as finer temporal and spectral structure is exposed. Figure 2 provides a distributional view of source specificity, most source–reconstruction similarities lie above the random

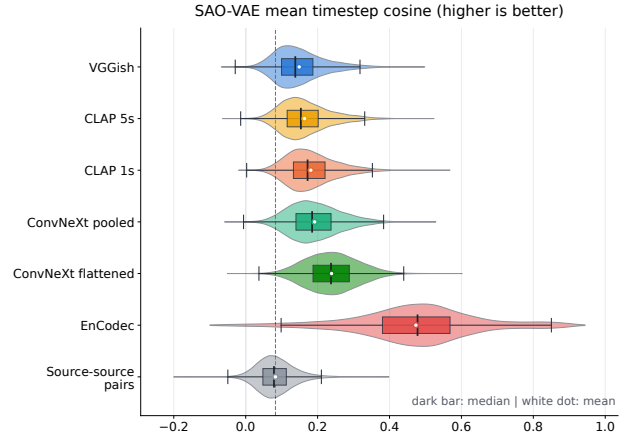


Fig. 2. Distribution of paired source–reconstruction SAO-VAE cosine similarity across encoder conditions, compared with 200k randomly sampled distinct source–source pairs.

source–source distribution, indicating that the reconstructions remain systematically closer to their paired source than to an unrelated test track.

Figure 1 reveals several qualitative differences between the representations. Notably, CLAP 5s retains frequency-specific temporal patterns despite using a single embedding, with low-frequency regions switching on and off similarly to the source. VGGish likewise preserves coarse rhythm and the overall spectral character, but follows source-specific harmonic structure less faithfully, consistent with its relatively strong MusicNN agreement despite weaker acoustic reconstruction. ConvNeXt pooled more closely tracks both the temporal pattern and harmonic structure, although frequency pooling removes finer spectral detail that the generative decoder must infer. EnCodec most closely reproduces the source’s time–frequency structure.

Conditioning ablations which can be found on the project page, show clearer gains in validation denoising loss from near zero projection initialization and temporally local attention than in final reconstruction metrics, particularly for denser interfaces, while learning rate choice also affects reconstruction variability.

5. CONCLUSION

Audio embeddings are useful abstractions, but downstream utility does not reveal how much source detail remains reconstructable. We frame inversion as a controlled audit and compare held-out source recovery from supervised, contrastive, and reconstruction-trained representations using a shared pre-trained diffusion probe. Future work should extend this analysis beyond music, to broader encoder families and more direct privacy leakage measures.

Acknowledgments

This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

6. REFERENCES

- [1] Aravindh Mahendran and Andrea Vedaldi, “Understanding deep image representations by inverting them,” in *Proc. CVPR*, 2015, pp. 5188–5196.
- [2] Karla Pizzi, Franziska Boenisch, Ugur Sahin, and Konstantin Böttinger, “Introducing model inversion attacks on automatic speaker recognition,” in *Proc. SPSC*, 2022, pp. 11–16.
- [3] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, et al., “Cnn architectures for large-scale audio classification,” in *Proc. ICASSP*, 2017, pp. 131–135.
- [4] Thomas Pellegrini, Ismail Khalfaoui-Hassani, Etienne Labbé, and Timothée Masquelier, “Adapting a ConvNeXt model to audio classification on AudioSet,” in *Proc. Interspeech*, 2023, pp. 4169–4173.
- [5] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, et al., “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *Proc. ICASSP*, 2023, pp. 1–5.
- [6] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, “High fidelity neural audio compression,” *Trans. Mach. Learn. Res.*, 2023.
- [7] Alexey Dosovitskiy and Thomas Brox, “Inverting visual representations with convolutional networks,” in *Proc. CVPR*, 2016, pp. 4829–4837.
- [8] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proc. ACM CCS*, 2015, pp. 1322–1333.
- [9] Luca Comanducci, Paolo Bestagini, Marco Tagliasacchi, Augusto Sarti, and Stefano Tubaro, “Reconstructing speech from cnn embeddings,” *IEEE Signal Process. Lett.*, vol. 28, pp. 952–956, 2021.
- [10] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux, “Speech resynthesis from discrete disentangled self-supervised representations,” in *Proc. Interspeech*, 2021, pp. 3615–3619.
- [11] Xianglong Zhang, Haoming Luo, Mingda Han, Qihao Dong, Guoming Zhang, Yanni Yang, Qianli Li, and Pengfei Hu, “Invercrs: Generative audio inversion attack in collaborative recognition systems,” in *Proc. ICDSCS*, 2025, pp. 516–526.
- [12] Xijie Zeng and Frank Rudzicz, “How to recover long audio sequences through gradient inversion attack with dynamic segment-based reconstruction,” in *Proc. Interspeech*, 2025, pp. 5118–5122.
- [13] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, et al., “Audioldm: Text-to-audio generation with latent diffusion models,” in *Proc. ICML*, 2023, pp. 21450–21474.
- [14] 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 Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2871–2883, 2024.
- [15] Javier Nistal, Marco Pasini, Cyran Aouameur, Maarten Grachten, and Stefan Lattner, “Diff-a-riff: Musical accompaniment co-creation via latent diffusion models,” in *Proc. ISMIR*, 2024, pp. 272–280.
- [16] Jaehun Kim, Julián Urbano, Cynthia C. S. Liem, and Alan Hanjalic, “Are nearby neighbors relatives? testing deep music embeddings,” *Frontiers in Applied Mathematics and Statistics*, vol. 5, pp. 53, 2019.
- [17] Sangeeta Srivastava, Ho-Hsiang Wu, João Rulff, Magdalena Fuentes, Mark Cartwright, Cláudio T. Silva, Anish Arora, and Juan Pablo Bello, “A study on robustness to perturbations for representations of environmental sound,” in *Proc. EUSIPCO*, 2022, pp. 125–129.
- [18] Victor Deng, Changhong Wang, Gaël Richard, and Brian McFee, “Investigating the sensitivity of pre-trained audio embeddings to common effects,” in *Proc. ICASSP*, 2025, pp. 1–5.
- [19] Zach Evans, C. J. Carr, Josiah Taylor, Scott H. Hawley, Jordi Pons, et al., “Stable audio open,” *arXiv preprint arXiv:2407.14358*, 2024.
- [20] Jonathan Ho and Tim Salimans, “Classifier-free diffusion guidance,” 2022.
- [21] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere, “The million song dataset,” in *Proc. ISMIR*, 2011, pp. 591–596.
- [22] Hendrik Schreiber, “Improving genre annotations for the million song dataset,” in *Proc. ISMIR*, 2015, pp. 241–247.
- [23] Alexandre Défossez, “Hybrid spectrogram and waveform source separation,” in *Proc. ISMIR*, 2021, pp. 137–144.
- [24] Jordi Pons and Xavier Serra, “MusiCNN: Pre-trained convolutional neural networks for music audio tagging,” in *Proc. ISMIR LBD*, 2019.