SAGE: SEMANTIC AUDIO GENERATIVE ENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio autoencoders compress waveforms into compact latent representations that serve as the interface between raw audio and downstream models. Current systems navigate a three-way trade-off between reconstruction quality, semantic structure of the latent space, and inference speed, typically favoring one or two of these at the expense of the others. This paper introduces SAGE, *Semantic Audio Generative Encoder*: a compact variational autoencoder, trained solely on publicly available music, that shapes its latent by distilling embeddings from a pretrained audio-text model. This 105M-parameter model runs at the inference cost of Stable Audio Open and reaches the listening-test quality of SAME-L, an autoencoder $8\times$ larger and $4\times$ slower, while surpassing both on objective perceptual and distributional metrics of reconstruction. Furthermore, it sets the state of the art on all nineteen probing tasks of latent semantics, in domain and out of domain. These results establish SAGE as a lightweight audio autoencoder that strikes the best balance of the three-way trade-off among those we evaluate, combining high reconstruction fidelity, state-of-the-art semantic structure, and fast inference.

1 INTRODUCTION

State-of-the-art music generation follows the latent diffusion recipe (Rombach et al., 2022). An autoencoder compresses the waveform into a compact latent. A generative model then learns to sample in that latent space. The autoencoder therefore runs at both ends of the pipeline since it encodes the training corpus of the generator and decodes every generated sample. While few-step samplers (Song et al., 2023) cut sampling from tens or hundreds of denoising steps to only a few, the decoder still runs at full cost on every sample and accounts for a large share of inference latency. For interactive, editing, and long-form applications, its cost is thus critical.

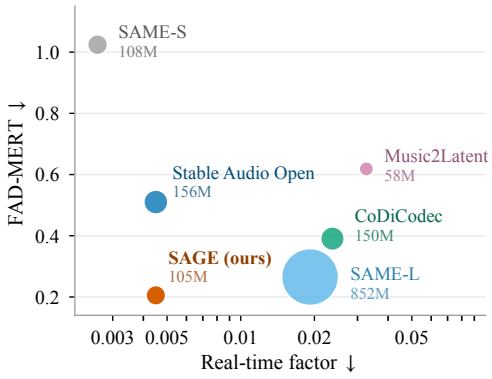


Figure 1: Distributional fidelity against inference cost on the MoisesDB mixtures; marker area is the parameter count.

Yet current music autoencoders force a choice between speed and fidelity. Stable Audio Open (Evans et al., 2025) is fast, a convolutional stack on the waveform at the sample rate, but trails in fidelity. SAME (Parker et al., 2026) reaches the best fidelity of the systems we compare and is the only one that shapes its latent semantically, but its large configuration has 852M parameters and four times the inference cost, and the small one, distilled from it to run faster, pays for the speed in fidelity. Music2Latent (Pasini et al., 2024) and CoDiCodec (Pasini et al., 2025) operate on the spectrogram and decode through a consistency model; like Stable Audio Open, they optimize the latent for reconstruction alone, so any semantic structure in it is incidental. No system offers the speed of the first, the quality of the second, and a semantically organized latent at once.

A spectrogram is a two-dimensional time-frequency signal. Vision already offers efficient hierarchical encoders for such signals. Swin (Liu et al., 2021; 2022) restricts attention to local windows, so its cost grows linearly with input size. Its stages downsample progressively and build multiscale features. In audio, this backbone is established for classification (Chen et al., 2022), not for generative autoencoding. Latent semantics has a similar gap: in images, aligning generative representations

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

with pretrained encoders speeds up training and improves samples (Yu et al., 2025; Yao et al., 2025). In audio, the same idea has relied on large models or closed codecs, costing some fidelity.

We propose SAGE, a compact VAE that encodes and compresses the complex STFT of stereo music with a hierarchical SwinV2 backbone (Liu et al., 2022) and distills a frozen CLAP encoder (Wu et al., 2023) into its latent, yielding a semantically organized representation that directly predicts the complex spectrum without the need for vocoders or phase reconstruction. We evaluate SAGE along two axes: reconstruction, through perceptual and distributional metrics, and latent semantics, through nineteen music probing tasks in the format of MAEB (El Assadi et al., 2026), six from the benchmark itself and thirteen we build on FMA (Defferrard et al., 2017) and MoisesDB (Pereira et al., 2023). Our 105M-parameter model sits at the corner of the trade-off in Figure 1: it matches the real-time factor of Stable Audio Open and is outrun only by the distilled SAME-S, whose Fréchet distance is several times higher, and it matches the subjective quality of SAME-L, the strongest baseline, at an eighth of its parameters and a quarter of its inference cost. On objective perceptual and distributional metrics it surpasses both, in domain and out of domain, and it leads every one of the nineteen probing tasks (Figure 2).

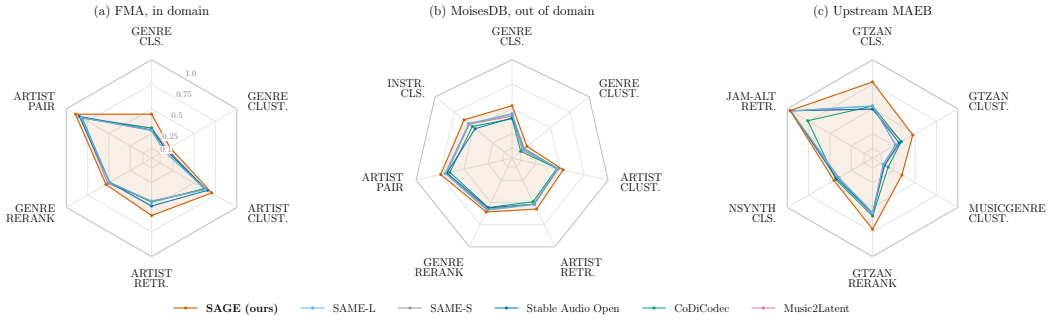


Figure 2: Probing on the nineteen tasks, one panel per block; higher is better on every task.

Our contributions are:

- **A spectrogram-native hierarchical autoencoder.** We adapt a SwinV2 encoder-decoder from images to the complex STFT, with rectangular patches and attention windows. At 105M parameters, the result runs as fast as Stable Audio Open and matches SAME-L in a listening test, surpassing both on perceptual and distributional reconstruction metrics.
- **Semantic distillation on a delayed schedule.** We distill CLAP into the latent, delayed during training so that reconstruction is established before alignment begins; we sweep the introduction of the latent term, the loss weight and the teacher, and identify the knee at which semantic structure is gained at the least cost in fidelity.
- **A thorough evaluation of audio autoencoders.** We measure every compared system through a single harness on five corpora. Reconstruction is scored with perceptual, distributional and sample-exact metrics, complemented by a MUSHRA listening test (ITU-R, 2015) on held-out commercial music, in which SAGE and SAME-L are statistically indistinguishable; latent semantics is scored with nineteen probing tasks in the MAEB format, thirteen of them instantiated on the same corpora.

Weights, code and evaluation harness will be released upon publication.

2 Method

2.1 Audio-Adapted Hierarchical Autoencoder

SAGE is a hierarchical VAE (Kingma & Welling, 2014) that compresses an audio signal $\mathbf{X}$ into a latent x by progressively shrinking the input via SwinV2 (Liu et al., 2022) blocks. Its input is the stacked $\mathbf{X} \in \mathbb{R}^{4\times F\times T}$ representation of the left and right channels, and real and imaginary parts,

of a dual-complex spectrogram, obtained via a short-time Fourier transform (STFT) with a 2048-sample Hann window and a hop of 512 samples, of a 44.1 kHz stereo waveform. To ease the blocks' compression, we drop the Nyquist bin and rescale the spectrum's magnitudes with a power law, as in CoDiCodec. We predict the complex spectrum directly, so no vocoder or phase reconstruction is needed.

Due to the nature of audio spectrograms, we introduce three adaptations to the SwinV2 architecture (Figure 3): strongly rectangular patches, matched by rectangular attention windows, and a shallow three-stage hierarchy. The activation function in the MLP is replaced by a SwiGLU (Shazeer, 2020), and attention is made exclusive (Zhai, 2026), orthogonalizing each token's attention output against its own value vector, so that it gathers only information complementary to the token itself. Otherwise, blocks retain their original design (scaled-cosine attention, log-spaced continuous position bias, post-residual normalization).

The encoder output x defines a diagonal Gaussian posterior, and the VAE bottleneck samples the latent z from it. A Kullback-Leibler term $\mathcal{L}_{\mathrm{KL}}$, averaged over the batch and the latent grid, pulls this posterior towards a standard normal prior. As in other autoencoders that feed into a second-stage generative model, the weight on this term is deliberately small (Section 2.3): in the rate-distortion trade-off it governs (Higgins et al., 2017), a light penalty confines the latent to a smooth, approximately standardized region without discarding the fine spectral structure the decoder must recover. Throughout our experiments, we hold the latent at the compression ratio of Stable Audio Open, so the two are compared at the same $\times 64$ budget. Finally, a mirrored decoder reverses the encoder's hierarchy and produces the reconstruction $\hat{\mathbf{X}} = D(z)$, followed by a lightweight residual refinement layer that reduces discontinuities between reconstructed frequency bands.

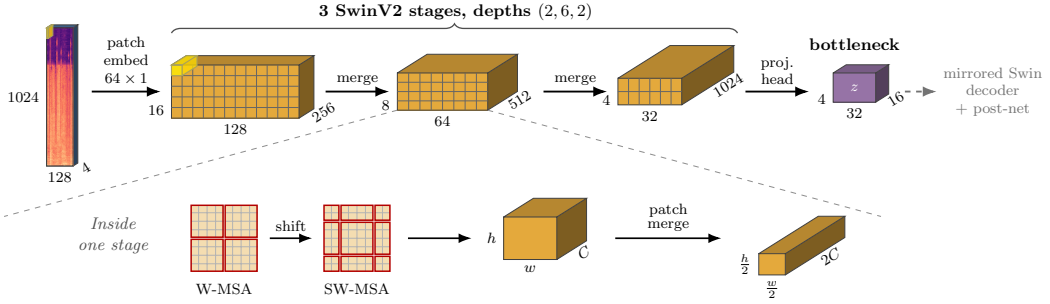


Figure 3: The SAGE encoder: (top) the tensor pipeline through the three SwinV2 stages to the latent bottleneck; (bottom) the internals of one stage.

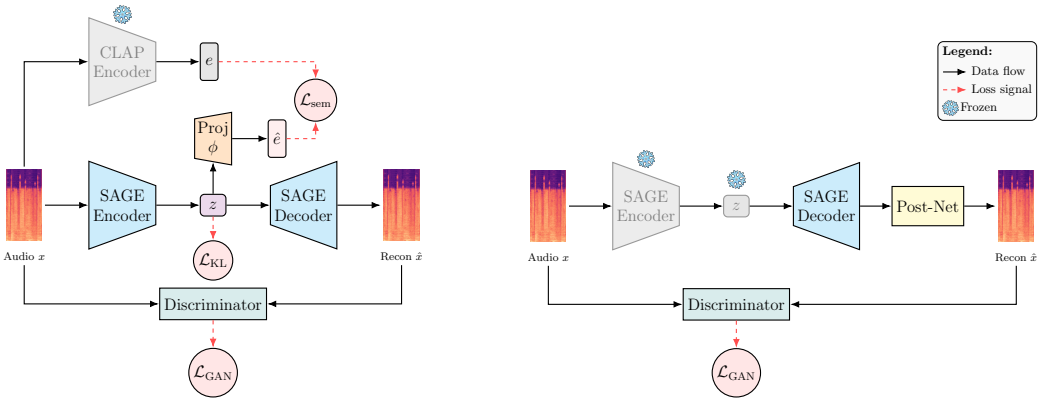
2.2 Semantic Latent Distillation

Reconstruction quality, the distortion axis of the rate-distortion-perception trade-off, says nothing about how the latent is organized, yet that organization is what probing, retrieval and classification consume directly and what a downstream generative model must learn from (Section 4). SAGE therefore distills (Hinton et al., 2015) its latent towards the embedding space of CLAP, a contrastive language-audio model whose audio arm, HTS-AT (Chen et al., 2022), maps a sample to a single 512-dimensional embedding aligned with natural-language descriptions of its content. During training, CLAP remains frozen, and the alignment follows the clip-level scheme of SALAD-VAE (Braun et al., 2026); the resulting branch is the upper pathway of Figure 4a.

Folding $z \in \mathbb{R}^{16\times4\times32}$ along frequency gives $16 \cdot 4 = 64$ channels over 32 frames, averaged over time into a clip descriptor $\bar{z} \in \mathbb{R}^{64}$. A learned linear head ϕ projects the descriptor into the teacher's space, and the loss maximizes its cosine similarity to the teacher embedding e of the same clip

$$\mathcal{L}_{\mathrm{sem}} \;=\; 1 - \cos\big(\phi(\bar{z}),\, e\big), \tag{1}$$

bounded in $[0, 2]$, which keeps its scale commensurate with the other terms. The teacher consumes the same segment downmixed to mono and resampled to its native 48 kHz. Teacher and head are used only during training and discarded at inference; furthermore, the head is optimized at a lower



(a) Pretraining (b) Decoder fine-tuning

Figure 4: The two training phases of SAGE.

learning rate than the rest of the model, since a head that adapts too quickly would shoulder most of the alignment, discouraging the encoder from pursuing it and making it unaligned at inference time.

Detached warm-up. For the first s_0 training steps the descriptor entering equation 1 is $\text{sg}[\bar{z}]$, the stop-gradient of $\bar{z}$: the latent develops its structure under the reconstruction and KL objectives alone, while ϕ, active from the first step, learns to track it. From step s_0 on, $\bar{z}$ enters undetached and $\mathcal{L}_{\text{sem}}$ reshapes a representation that already encodes the signal. Later placement of the gate s_0 favors reconstruction and earlier placement favors semantic structure; we set it at the empirical sweet spot of Section 3.5.

2.3 Two-Phase Training Curriculum

A decoder trained only on average spectral distances regresses to the mean of the reconstructions compatible with a latent, which is smoother than any of them and lacks the high-frequency detail that carries much of the perceived realism. Following neural audio codecs, we add an adversarial objective (Goodfellow et al., 2014) against the composite discriminator of WavTokenizer (Ji et al., 2025), applied to the waveform obtained by inverting the predicted spectrogram. Since this discriminator judges each channel independently, we also feed it the mid $M = L + R$ and side $S = L - R$ signals. We use the relativistic RpGAN formulation (Jolicoeur-Martineau, 2019; 2020), which rates real audio against its own reconstruction; the generator receives the adversarial term $\mathcal{L}_{\text{adv}}$ and a feature-matching term $\mathcal{L}_{\text{fm}}$ (Kumar et al., 2019). Fidelity is enforced by three terms (Appendix F): $\mathcal{L}_{\text{STFT}}$, a squared error on the power-compressed complex spectrogram, penalizing magnitude and phase jointly; $\mathcal{L}_{\text{mel}}$ (Yamamoto et al., 2020), a multi-resolution mel distance on the reconstructed waveform; and $\mathcal{L}_{\text{SD}}$, the sum-and-difference multi-resolution STFT loss of Stable Audio Open. The last term is needed for the stereo image: a squared error weighs the low-energy side only by its share of the energy, and without $\mathcal{L}_{\text{SD}}$ the side is reconstructed about 12 dB too quiet (Appendix B).

Phase 1: pretraining. The full system of Figure 4a trains from scratch on

$$\mathcal{L} = \lambda_{\text{STFT}}\mathcal{L}_{\text{STFT}} + \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}}, \quad (2)$$

with $\mathcal{L}_{\text{sem}}$ under the detached warm-up of Section 2.2. Reconstruction dominates and the adversarial pair is kept well below it (loss weights in Appendix A, Table 6). The discriminator is active from the first step, which suits the WavTokenizer discriminator (Appendix E). An exponential moving average (EMA) of the weights is used for evaluation and to initialize the next phase.

Phase 2: decoder fine-tuning. We reload the EMA weights and freeze the encoder (Figure 4b), so the latent consumed by downstream models stays exactly as pretraining left it. The KL and semantic terms are switched off (Appendix A), a zero-initialized post-net is attached, and the discriminator is re-initialized. The decoder alone then trains further, gaining perceptual quality.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Model configuration. SAGE's architecture (Section 2.1) is instantiated with an embedding size of $C = 256$ in the first block, three stages of depths $(2, 6, 2)$, $(8, 16, 32)$ attention heads, 64×1 patches, 4×32 attention windows, $C_z = 16$ latent channels and stochastic depth at rate 0.1 in encoder and decoder, the decoder carrying in addition the residual post-net, composed of two 64-channel convolutions. The total size, barring the CLAP teacher, the discriminator and the projection head ϕ, is 104.6M parameters. The teacher is the LAION-CLAP checkpoint trained on music and general audio.

Data. SAGE is trained on ≈10.5K hours of 44.1 kHz stereo audio from three public corpora: the FMA-full training split, the MTG-Jamendo training partition (Bogdanov et al., 2019) (excluding tracks used by the Song Describer Dataset), and M4Singer (Zhang et al., 2022), whose solo-singing recordings expose the model to isolated voice alongside full mixes. To avoid over-sampling M4Singer's many short recordings, each epoch uses a different quarter of it alongside all of FMA and Jamendo (≈122.4K tracks, one random 1.5 s crop each).

Reconstruction is evaluated on five held-out sets: the FMA test split (11,263 clips of 30 s, 20× the training window) as the in-domain benchmark; MoisesDB mixtures and isolated stems (1,998 and 1,546 windows of 10 s), professionally produced multitracks that are out of domain; and MusicCaps (Agostinelli et al., 2023) (964 clips of 10 s) and the Song Describer Dataset (Manco et al., 2023) (8,364 chunks of 10 s), the benchmarks of Music2Latent/CoDiCodec and Stable Audio Open/SAME, respectively, so that every baseline is also scored on its authors' own choice of corpus. Semantic probing uses the FMA test split, 1,611 chunks of 30 s from MoisesDB mixtures and stems (the stems enable source-level tasks), and the external corpora of Section 3.2.

Training. Phase 1 trains from scratch for 500 epochs (≈ 239k generator updates) on 16 A100 GPUs at a global batch of 128 segments, with AdamW (Loshchilov & Hutter, 2019) at a peak learning rate of 10^{-3} under an inverse square-root schedule, in full precision. The discriminator and the projection head have their own optimizers; the head runs at the slow constant rate required by Section 2.2. Phase 2 keeps hardware, batch and corpus fixed, lowers the generator learning rate to 10^{-4} with a restarted schedule, and runs for 992 epochs. Appendix A gives the full optimization settings. The two phases cost 1,536 and 3,043 GPU-hours, respectively.

Baselines. In Table 1 we compare against five open-weight continuous-latent audio autoencoders; Stable Audio Open is the convolutional waveform-domain VAE whose compression budget SAGE adopts. SAME is released in two configurations and we evaluate both: SAME-L, the flagship, is by far the largest system in the comparison, and SAME-S is a 108M-parameter variant distilled from it to run on CPU, which puts it at SAGE's size. Both share the 256-dimensional latent and the semantic shaping of SAME (Section 4), making them the only other systems that organize their latent on purpose. CoDiCodec, run in its continuous mode, and Music2Latent are consistency autoencoders on the complex STFT, the former at twice the compression of Stable Audio Open and SAGE; Music2Latent treats the two channels independently rather than modeling them jointly, and is run on both, so its ratio holds per channel. Every baseline is run from its public weights through the same harness, one call per clip with no external chunking or overlap-add. Timings use each authors' implementation at batch size one in full precision, averaged over 100 MoisesDB mixtures; Appendix C reports them uncompiled. SALAD-VAE, whose objective Section 2.2 adapts, is not a baseline: it has released neither weights nor code, and targets general audio.

3.2 EVALUATION PROTOCOL

Reconstruction metrics. We report three families of reconstruction metrics and give priority to the first two. The perceptual family is the cosine similarity between the CLAP embeddings of a clip and of its reconstruction. We compute it under the music checkpoint, the teacher of Section 2.2, and under the general-audio checkpoint, which plays no part in training. A MUSHRA listening test completes the family with the human judgment these similarities approximate. The distributional

Table 1: Inference is the mean encode-decode time of a 10 s clip, compiled with `torch.compile`, on one A100; RTF is that time divided by the clip duration; data hours are those reported by the authors; compression is waveform samples per real latent scalar.

Model	Params	Data (h)	Domain	Latent	Comp.	Infer. (ms) ↓	RTF ↓
Music2Latent	58M	≈ 3,800	STFT	64	×64	326.5	0.0326
SAME-S	108M	≈ 19,500	waveform	256	×32	**26.1**	**0.0026**
CoDiCodec	≈ 150M	≈ 4,030	STFT	64	×128	236.8	0.0237
Stable Audio Open	156M	≈ 7,300	waveform	64	×64	45.2	0.0045
SAME-L	852M	≈ 19,500	waveform	256	×32	191.8	0.0192
SAGE (ours)	105M	10,460	STFT	64	×64	44.8	0.0045

family is the Fréchet audio distance (Kilgour et al., 2019) between the reference and the reconstruction distributions, computed in three embedding spaces, framewise MERT (Li et al., 2024) features, whole-clip embeddings from that same general-audio checkpoint and PANN (Kong et al., 2020) features, since the ranking it induces is known to depend on the space in which it is taken (Gui et al., 2024). Together they answer the question we care about, whether a reconstruction sounds plausible and is hard to tell from the original. The third family is sample-exact and collects the signal-to-distortion ratio (Vincent et al., 2006) and the multi-resolution STFT distance that the objective of Section 2.3 minimizes directly. We keep these in every table and read them as diagnostics, since they charge for inaudible phase rotations and favor a model that regresses the waveform sample by sample over one that predicts a spectrum.

Semantic probing. We freeze the encoder and probe its latents with the nineteen audio-only music tasks. On the corpora of Section 3.1, so that both axes are evaluated on the same data, they are genre classification, genre and artist clustering, artist retrieval, genre reranking and artist pair classification on the FMA test split and the MoisesDB mixtures, plus instrument classification on the isolated MoisesDB stems, a source-level task that mixtures cannot pose. The remaining six are the upstream tasks of MAEB on their original datasets, which none of the compared models have seen in training. For every model we fold the latent along frequency into channels and average it over time. Classification tasks fit a track-grouped, cross-validated logistic regression while clustering, retrieval, reranking and pair classification train nothing and read the pooled latent directly.

3.3 Reconstruction

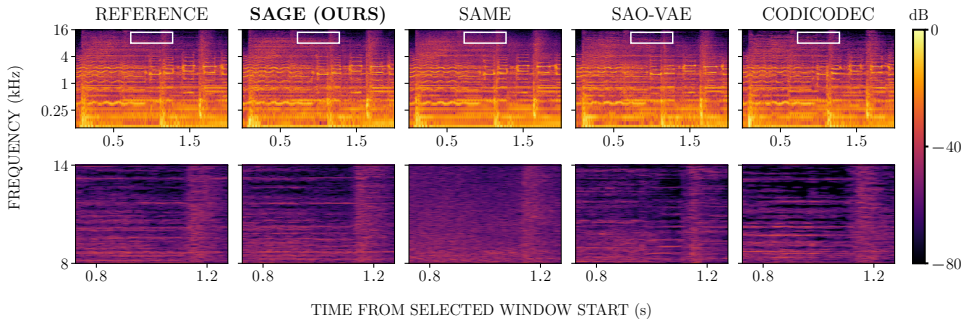


Figure 5: STFT power spectrograms of a selected music excerpt. Top: reference and reconstructions. Bottom: the same highlighted region enlarged for each model.

Table 2 reports reconstruction on the five evaluation sets of Section 3.1. SAGE holds the lowest Fréchet distance and the highest CLAP similarity on twenty-three of the twenty-five measurements the five sets provide, on twenty-four of them against SAME-L, a model eight times larger, and on every one against SAME-S; the exceptions are FAD-MERT on MusicCaps, where CoDiCodec is ahead, and FAD-PANN on the MoisesDB mixtures, where SAME-L is. Figure 1 places these results against inference cost. Stable Audio Open runs at the same real-time factor as SAGE and trails it on

Table 2: Reconstruction on the five evaluation sets, best per column within each set in bold.

		FAD ↓			CLAP ↑			
	Model	MERT	CLAP	PANN	music	audio	SDR ↑	dSTFT ↓
FMA test	**SAGE (ours)**	**0.113**	**0.016**	**1.422**	**0.956**	**0.960**	5.28	**0.929**
	SAME-L	0.210	0.038	3.470	0.897	0.927	**10.91**	1.746
	Stable Audio Open	0.319	0.044	2.891	0.892	0.911	5.64	1.342
	CoDiCodec	0.283	0.026	3.243	0.907	0.934	1.78	1.379
	SAME-S	0.649	0.082	7.169	0.839	0.870	8.80	1.675
	Music2Latent	0.396	0.045	4.599	0.862	0.906	1.21	1.531
MoisesDB mixtures	**SAGE (ours)**	**0.205**	**0.026**	1.438	**0.966**	**0.971**	4.79	0.857
	SAME-L	0.265	0.033	**1.277**	0.954	0.966	**9.49**	**0.816**
	Stable Audio Open	0.510	0.118	2.477	0.933	0.944	5.02	0.956
	CoDiCodec	0.391	0.130	11.531	0.930	0.923	1.84	0.848
	SAME-S	1.024	0.094	4.266	0.914	0.930	7.97	0.839
	Music2Latent	0.619	0.318	3.362	0.930	0.945	1.22	0.943
MoisesDB stems	**SAGE (ours)**	**0.135**	**0.027**	**1.692**	**0.962**	**0.965**	6.69	**1.058**
	SAME-L	0.559	0.087	4.141	0.888	0.911	**12.89**	1.398
	Stable Audio Open	0.552	0.121	3.198	0.934	0.942	7.05	1.190
	CoDiCodec	0.611	0.140	4.418	0.882	0.903	4.00	1.295
	SAME-S	1.348	0.145	7.713	0.852	0.857	9.70	1.357
	Music2Latent	1.650	0.331	6.564	0.834	0.854	3.52	1.646
MusicCaps	**SAGE (ours)**	0.306	**0.029**	**3.169**	**0.930**	**0.957**	5.09	**1.229**
	SAME-L	0.508	0.092	3.529	0.870	0.914	**10.56**	1.741
	Stable Audio Open	0.529	0.125	5.171	0.867	0.909	5.47	1.258
	CoDiCodec	**0.284**	0.056	5.635	0.918	0.940	1.83	1.239
	SAME-S	0.957	0.134	6.484	0.828	0.873	8.56	1.656
	Music2Latent	0.582	0.287	5.713	0.887	0.928	1.12	1.406
Song Describer	**SAGE (ours)**	**0.098**	**0.011**	**0.888**	**0.969**	**0.976**	6.31	**0.878**
	SAME-L	0.186	0.037	1.497	0.940	0.955	**12.17**	1.451
	Stable Audio Open	0.349	0.126	2.417	0.924	0.942	6.86	1.213
	CoDiCodec	0.290	0.085	8.397	0.922	0.920	2.42	1.178
	SAME-S	0.648	0.087	4.732	0.896	0.915	9.97	1.399
	Music2Latent	0.408	0.276	2.847	0.915	0.935	2.24	1.290

every perceptual and distributional measurement; CoDiCodec and Music2Latent trail as well while running five to seven times slower; SAME-L is the only system that listeners rate on par with SAGE (Table 3), at eight times the parameters and more than four times the inference time. SAGE thus pairs the speed of Stable Audio Open with the perceived quality of SAME-L. Figure 5 shows the same trend on a single excerpt.

The only faster system, SAME-S, pays for its speed in the distributional metrics: its FAD-MERT is three to ten times SAGE's, the highest of any system on four of the five sets. Its authors' listening test agrees, rating it below every other system they tested, Stable Audio Open included (66.1 vs. 73.3), despite a higher SI-SDR (9.6 vs. 6.2 dB). The Fréchet distances in the MERT and PANN spaces place it behind Stable Audio Open as the listeners do, which is why Section 3.2 gives them priority and keeps SDR as a diagnostic.

The MUSHRA test uses ten 10 s excerpts of commercial music held out of training, comparing SAGE against SAME-L, Stable Audio Open and CoDiCodec, together with a hidden reference and a 3.5 kHz low-pass anchor. Music2Latent and SAME-S are omitted to shorten the test: the first trails all four on FAD-MERT on every set of Table 2, and the second is rated below Stable Audio Open in the test of its own authors. Participants are filtered on their ability to correctly rate the hidden reference. Table 3 reports the filtered results, for a total of 21 raters. Two groups emerge. SAGE and SAME-L score 81.6 and 81.8 with overlapping 95% confidence intervals, so the listening test cannot separate them, although SAME-L has eight times the parameters and four times the inference cost. Stable Audio Open and CoDiCodec form a second group and trail by fifteen to seventeen points; in particular, at the same real-time factor as Stable Audio Open, SAGE scores 17 points higher.

Table 3: Mean MUSHRA score ± 95% CI, with parameters / RTF under each model; 38 participants, 21 after filtering. The hidden reference scored 97.9 ± 0.7 and the 3.5 kHz low-pass anchor 15.4 ± 2.0.

SAGE (ours)	SAME-L	Stable Audio Open	CoDiCodec
105M / 0.0045	852M / 0.0192	156M / 0.0045	150M / 0.0237
$\mathbf{81.6 \pm 2.7}$	$\mathbf{81.8 \pm 2.6}$	64.6 ± 3.7	66.4 ± 3.5

3.4 Semantic Probing

Figure 2 reports the nineteen tasks in their three blocks, with the absolute scores and a CLAP-oracle reference row in Table 9 of Appendix D. Every score is the metric MAEB defines for its task type (Table 9); all are bounded by one, so the figure draws every task on the same radius. SAGE leads every one of the nineteen tasks, in domain, out of domain and on the upstream corpora that none of the compared systems has trained on. On the block averages of Table 4, which pool metrics of different kinds as the MAEB leaderboards do, the five baselines lie within 8% of one another and SAGE stands 10% to 28% above the best of them, most of all on the upstream tasks. Figure 6 shows that geometry with no model fitted at all, projecting with UMAP (McInnes et al., 2018) the descriptors of the isolated MoisesDB stems, colored by their ground-truth instrument: the stems group by instrument, and related sources lie side by side, the bass next to the guitar, the other keys next to the piano and the percussion at the edge of the drums.

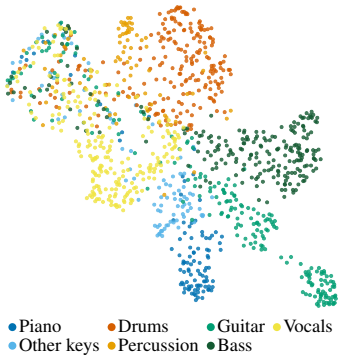


Figure 6: UMAP of SAGE.

Table 4: Average score over the tasks of each block.

Block	**SAGE (ours)**	SAME-L	SAME-S	Stable Audio Open	CoDiCodec	Music2Latent
FMA (6)	**0.563**	0.470	0.468	0.490	0.474	0.484
MoisesDB (7)	**0.544**	0.493	0.487	0.469	0.456	0.479
Upstream MAEB (6)	**0.622**	0.473	0.472	0.481	0.471	0.487

3.5 Ablations

Table 5: Semantic branch on FMA-medium, probes averaged over the six FMA tasks. Gate steps are optimizer steps, one generator update every three of them.

	FAD ↓		CLAP ↑				
Configuration	MERT	CLAP	music	audio	SI-SDR ↑	STFT ↓	Probes ↑
Gate and strength, with the WavTokenizer discriminator							
$s_0 = 25$k, $\lambda_{\text{sem}} = 1$ (adopted)	0.240	0.051	0.881	0.879	0.76	**1.099**	0.637
$s_0 = 50$k, $\lambda_{\text{sem}} = 1$	0.248	**0.046**	**0.889**	**0.885**	0.35	1.117	0.629
$s_0 = 100$k, $\lambda_{\text{sem}} = 1$	**0.218**	0.051	0.878	0.881	**1.46**	1.121	0.614
$s_0 = 25$k, $\lambda_{\text{sem}} = 0.5$	0.250	0.052	0.885	0.882	1.01	1.102	0.631
$s_0 = 25$k, $\lambda_{\text{sem}} = 2$	0.382	0.061	0.872	0.870	−0.06	1.134	**0.641**
Teacher, without the adversarial branch							
none	**1.453**	**0.41**	**0.645**	**0.624**	**2.47**	1.083	0.495
MERT	1.490	0.44	0.625	0.607	2.28	**1.052**	0.521
CLAP (adopted)	1.746	0.48	0.613	0.575	1.60	1.071	**0.642**

Table 5 sweeps the two knobs of Section 2.2 on FMA-medium, with a reduced backbone that predates the SwiGLU and XSA variants of the final one, so what transfers is the ordering inside a study and not the absolute values. The sweep traces the trade the method rests on: doubling the weight

buys the best probing average of the sweep and pays for it with a FAD-MERT 60% worse, opening the gate four times later reverses both, and the configuration we adopt sits at the knee. The lower block first removes the teacher and then changes it, everything else held fixed. Distillation costs fidelity on every column but the STFT distance, and what it returns is probing structure: little of it with the audio-only teacher, a great deal with the language-audio one. The CLAP-based metrics move against the model that distills CLAP on both knobs: the strongest weight gives the worst CLAP cosine and the worst FAD-CLAP of the sweep, and in the lower block the CLAP-distilled arm scores below the undistilled one on both CLAP similarities, the music one being the teacher itself. A metric inflated by its own training signal would move the other way. Removing the teacher drops the probing average from 0.642 to 0.495, and with it most of the margin Section 3.4 reports over the compared systems.

Appendix E carries two further studies. Any discriminator moves the distributional and perceptual metrics a long way, and the best moment to introduce one depends on its family; depth and width both convert parameters into quality, depth at a larger return per parameter.

4 Related Work

Audio autoencoders. Recent audio autoencoders either keep a continuous, typically variational, latent or discretize it through residual vector quantization (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023). Neural codecs target the best fidelity at a given bitrate and quantize to reach bitrates a continuous latent cannot; the autoencoders behind latent diffusion stay continuous, lightly regularized towards a Gaussian prior, and trade some fidelity for a latent the second stage can model more easily. The consistency autoencoders Music2Latent and CoDiCodec decode in a single step with a consistency model (Song et al., 2023), trading sample-exactness for plausibility.

Audio autoencoders also differ in their input representation. Waveform-domain models such as Stable Audio Open and SAME learn phase-aware filtering from samples; models on magnitude or mel spectrograms, as in AudioLDM (Liu et al., 2023), discard phase and need a separately trained vocoder; models on the complex STFT, Music2Latent, CoDiCodec and SAGE among them, recover the waveform with a single inverse transform.

Semantic alignment. Most of the systems above optimize the latent for reconstruction alone, whereas recent work shapes it on purpose, arguing that the encoder is the first stage of a generative system and the geometry it leaves behind is what the second stage has to model. In audio, the idea has been pursued from two directions. SAME shapes its latent with linear regressors onto hand-crafted chroma and stereo-image targets and a contrastive critic trained jointly over latent, audio, and text. SALAD-VAE (Braun et al., 2026) distills CLAP embeddings into a clip-level latent descriptor.

Music generative models. Text-to-music systems generate in the latent of a pretrained autoencoder, and their quality is bounded by it. Discrete models such as MusicLM (Agostinelli et al., 2023) and MusicGen (Copet et al., 2023) run an autoregressive transformer over codec tokens; continuous ones such as AudioLDM and the Stable Audio family (Evans et al., 2026) run latent diffusion over a VAE bottleneck. In both cases the generator must learn from the latent alone which directions correspond to instrument, genre or timbre, a task the image-domain evidence of the Introduction suggests an organized latent eases. Matching the budget and channel count of Stable Audio Open, SAGE is a drop-in first stage for diffusion architectures designed at that rate.

5 Conclusions

We presented SAGE, a compact variational autoencoder that shapes its latent by distilling a pretrained audio-text model on a delayed schedule, so that reconstruction settles before the semantic gradient is admitted. At 105M parameters, SAGE runs at the inference cost of Stable Audio Open, behind only the distilled SAME-S, and matches the listening-test quality of SAME-L, a model eight times larger and four times slower, surpassing both on the perceptual and distributional metrics of reconstruction; it also leads all nineteen probing tasks, in domain and out of domain.

AI USE STATEMENT

Large language models were used extensively to write and refactor the training code. All code was reviewed and run by the authors, and the method, experiments, and analysis are the authors' own; the authors take full responsibility for the content.

ETHICS STATEMENT

SAGE was trained solely on publicly available data, and the commercial music of the listening test was used only for evaluation. The listening test was conducted with volunteer participants who gave informed consent; no personal or sensitive data were collected, and responses were analyzed anonymously. As a first stage for generative models, SAGE contributes to making audio and music generation better and cheaper. That progress may displace the work of composers, musicians and audio engineers, and it raises questions of attribution, consent and misuse. We believe these models should be developed alongside tools that support audio professionals rather than replace them.

REPRODUCIBILITY STATEMENT

The code, weights and evaluation harness of SAGE will be released upon publication. Since all training data are publicly available and the full configuration of both training phases is given in Appendix A, the released code is sufficient to reproduce our results. The random seeds used for the data splits will be included in the configurations and, even without those, the code will include the procedure to regenerate equivalent splits, so any remaining variation should come from the partitioning alone.

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, 2023. URL https://arxiv.org/abs/2301.11325.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML)*, Long Beach, CA, United States, 2019. URL http://hdl.handle.net/10230/42015.

Sebastian Braun, Hannes Gamper, and Dimitra Emmanouilidou. Salad-vae: Semantic audio compression with language-audio distillation. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, January 2026.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pp. 646–650, 2022. doi: 10.1109/ICASSP43922.2022.9746312.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Michaël 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, ISMIR 2017, Suzhou, China, October 23-27, 2017*, pp. 316–323, 2017.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Trans. Mach. Learn. Res.*, 2023.

Adnan El Assadi, Isaac Chung, Chenghao Xiao, Roman Solomatin, Animesh Jha, Rahul Chand, Silky Singh, Kaitlyn Wang, Ali Sartaz Khan, Marc Moussa Nasser, Sufen Fong, Pengfei He, Alan Xiao, Ayush Sunil Munot, Aditya Shrivastava, Artem Gazizov, Niklas Muennighoff, and Kenneth Enevoldsen. MAEB: Massive audio embedding benchmark, 2026. URL https://arxiv.org/abs/2602.16008.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888461.

Zach Evans, Julian D. Parker, Matthew Rice, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio 3, 2026. URL https://arxiv.org/abs/2605.17991.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2672–2680, 2014.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 1331–1335, 2024. doi: 10.1109/ICASSP48485.2024.10446663.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

ITU-R. Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems. International Telecommunication Union, 2015.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Alexia Jolicoeur-Martineau. On relativistic f-divergences. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, pp. 4931–4939, 2020.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pp. 2350–2354, 2019. doi: 10.21437/INTERSPEECH.2019-2219.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C. Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14881–14892, 2019.

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 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger B. Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. MERT: acoustic music understanding model with large-scale self-supervised training. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202, pp. 21450–21474, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer V2: scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 11999–12009, 2022. doi: 10.1109/CVPR52688.2022.01170.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

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. In *Machine Learning for Audio Workshop, NeurIPS*, 2023. URL https://arxiv.org/abs/2311.10057.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2018. arXiv:1802.03426.

Julian D. Parker, Zach Evans, CJ Carr, Zack Zukowski, Josiah Taylor, Matthew Rice, and Jordi Pons. SAME: A semantically-aligned music autoencoder, 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, ISMIR 2024, San Francisco, California, USA and Online, November 10-14, 2024*, 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, ISMIR 2025, Daejeon, South Korea, September 21-25, 2025*, pp. 433–441, 2025. doi: 10.5281/ZENODO.17811403.

Igor Pereira, Felipe Araújo, Filip Korzeniowski, and Richard Vogl. Moisesdb: A dataset for source separation beyond 4-stems. In *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, pp. 619–626, 2023. doi: 10.5281/ZENODO.10265363.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Noam Shazeer. GLU variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pp. 32211–32252, 2023.

Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Trans. Speech Audio Process.*, 14(4):1462–1469, 2006. doi: 10.1109/TSA.2005.858005.

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 ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 6199–6203, 2020. doi: 10.1109/ICASSP40776.2020.9053795.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 15703–15712, 2025. doi: 10.1109/CVPR52734.2025.01464.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30: 495–507, 2022. doi: 10.1109/TASLP.2021.3129994.

Shuangfei Zhai. Exclusive self attention, 2026. URL https://arxiv.org/abs/2603.09078.

Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, and Zhou Zhao. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

A HYPERPARAMETERS

Table 6 lists the complete configuration of the model reported in the paper, together with the two training phases of Section 3.1.

Table 6: Full configuration of SAGE and of the two training phases.

Input representation	
Sample rate, channels	44.1 kHz, stereo
STFT window / hop	2048 (Hann) / 512
Frequency bins, frames	$F = 1024$ (Nyquist dropped), $T = 128$
Training segment	65,024 samples per channel (≈ 1.5 s)
Power-law compression α, scale β	0.65, 0.35
Architecture	
Patch size, embedding width	64×1, $C = 256$
Stage depths, attention heads	$(2, 6, 2)$, $(8, 16, 32)$
Attention window	4×32
Feed-forward	SwiGLU, hidden ratio 3.0
Attention variant, normalization	exclusive self-attention, post-residual
Latent	$C_z = 16$, grid 4×32
Compression rate	130,048 waveform samples $\rightarrow$ 2,048 latent scalars ($63.5\times$)
Stochastic depth	0.1
Post-net (decoder only)	2 convolutions, 7×3 kernels, 64 channels
Parameters	104,629,380
Stereo terms	
Sum-and-difference MR-STFT	M, S, L, R; FFT 2048 to 64, hop FFT/4, A-weighted
Discriminator inputs	L, R, M, S folded into the batch
Optimization, phase 1	
Hardware, global batch	16 A100 (4×4), 128 segments
Epochs, batches	500, 956 per epoch ($\approx$ 478k batches in total)
Generator updates	$\approx$ 239k, one every two batches, alternating with the discriminator
Generator optimizer	AdamW, lr 10^{-3}, $\beta = (0.9, 0.98)$, wd 10^{-4} (matrices only)
Schedule	inverse square root, $\gamma = 2 \times 10^5$, power $1/2$
Warm-up	exponential, 90% of peak at $\approx$ 2,400 steps
Discriminator optimizer	AdamW, lr 2×10^{-4}, $\beta = (0.8, 0.99)$, wd 10^{-3}
Projection head optimizer	AdamW, lr 10^{-5}, constant
Semantic gate s_0	$\approx$ 8.3k generator updates (25,000 steps over all three optimizers)
Loss weights, order of equation 2	$(1, 0.5, 1, 10^{-4}, 1, 0.1, 0.2)$
Gradient clipping, precision	norm 20, FP32
EMA decay, seed	0.9998, 94
Optimization, phase 2 (differences only)	
Epochs	992
Generator learning rate	10^{-4}, schedule and warm-up restarted
Encoder, discriminator	frozen, re-initialized from scratch
Loss weights	$\lambda_{\mathrm{KL}} = \lambda_{\mathrm{sem}} = 0$, $\lambda_{\mathrm{SD}} = 1$

B STEREO IMAGE

The objective of Section 2.3 carries $\mathcal{L}_{\mathrm{SD}}$ for the stereo image alone, and Table 7 measures what the term does. The four quantities are computed on the MoisesDB mixtures against their references, on the multi-corpus pretraining of Section 3.1 rather than on the reduced setup of Section 3.5. The level of the side, in decibels relative to the side of the reference, is zero when the image is as wide as it should be and negative when the reconstruction is narrower; d_{width} is a distance between the width envelopes of reconstruction and reference; and the last two columns are the scale-invariant signal-to-distortion ratios of the side and of the mid, which separate the level of the side from its content.

Without the term the side comes out about 12 dB too quiet, and the deficit is the same after three hundred and after fifteen hundred epochs, which rules out an effect of under-training. With it the

level is essentially matched, the width distance falls by a third and the side gains more than nine decibels of content, while the mid is left where it was.

Table 7: Stereo image on the MoisesDB mixtures, with and without $\mathcal{L}_{\mathrm{SD}}$ in pretraining.

Pre-training	Side level (dB)	$d_{\mathrm{width}} \downarrow$	SI-SDR side ↑	SI-SDR mid ↑
without $\mathcal{L}_{\mathrm{SD}}$, epoch 319	−12.40	0.092	−14.81	4.87
without $\mathcal{L}_{\mathrm{SD}}$, epoch 1499	−11.99	0.091	−14.32	4.87
with $\mathcal{L}_{\mathrm{SD}}$, epoch 499	−0.20	0.060	−5.09	4.84

C Inference Timing without Compilation

Inference timings are averaged over 100 MoisesDB mixtures of 10 s, at batch size one and in full precision on a single A100. Each baseline runs its authors' released implementation and public weights through the same harness, with one encode-decode call per clip and no external chunking or overlap-add. Table 1 reports timings from this loop compiled with `torch.compile`; Table 8 reports the same measurement without compilation, so that eager and compiled inference costs can be read side by side. RTF is the encode-decode time divided by the clip duration.

SALAD-VAE, whose objective Section 2.2 adapts, is excluded from the comparison because it has released neither weights nor code. It targets general audio and reports fidelity on speech rather than music.

Table 8: Same measurement as Table 1, without `torch.compile`.

Model	Infer. (ms) ↓	RTF ↓
Music2Latent	401.6	0.0402
SAME-S	**30.8**	**0.0031**
CoDiCodec	445.9	0.0446
Stable Audio Open	97.3	0.0097
SAME-L	209.8	0.0210
SAGE (ours)	57.0	0.0057

D Semantic Probing by Task

Table 9 gives the absolute scores behind Figure 2 and Table 4, each under the metric MAEB defines for its task type and with the protocol of Section 3.2; retrieval is audio-to-audio throughout.

Averaged over the nineteen tasks, SAGE retains 84% of the CLAP-oracle score. The gap concentrates in genre clustering, whose tasks give three of the suite's four lowest figures, from 55% to 69%, and 76% on Music Genre, the same level as MoisesDB genre classification and NSynth; artist clustering, equally training-free, retains 87% to 98%, so the drop tracks genre rather than the absence of supervision. Every other task retains above 82%, and on Jam-ALT artist retrieval SAGE exceeds the oracle outright.

E Additional Ablations

The two studies below isolate one factor at a time on FMA-medium, with a reduced backbone and, where indicated, no adversarial branch; the winner of each is what Section 3.1 adopts. As in the semantic study of Section 3.5, the runs predate the SwiGLU and XSA variants of the final backbone, so what transfers is the ordering inside a study, not the absolute values.

Adversarial branch. Any discriminator moves the distributional and perceptual metrics a long way, the WavTokenizer ensemble cutting FAD-MERT sevenfold at a cost in SI-SDR, and the objective matters as much as the architecture: on the same EnCodec critic the relativistic formulation

Table 9: Semantic probing on the nineteen tasks, best per row among the six autoencoders in bold, block averages in the last row of each block. *SAME-L and SAME-S are probed at their native 256 dimensions, four times the width of every other descriptor. †CLAP-oracle, underlined throughout, probes the frozen 512-dimensional teacher of Section 2.2 under the same protocol; it cannot be inverted back to audio and is a reference, not a seventh autoencoder, so it is excluded from the bolding.

Task	Metric	**SAGE**	SAME-L*	SAME-S*	Stable Audio Open	CoDiCodec	Music2-Latent	CLAP-oracle†
FMA test split, in domain								
Genre classification	accuracy	**0.447**	0.285	0.279	0.287	0.305	0.307	0.529
Genre clustering	V-measure	**0.215**	0.146	0.145	0.171	0.168	0.185	0.385
Artist clustering	V-measure	**0.706**	0.634	0.634	0.655	0.617	0.632	0.719
Artist retrieval	nDCG@10	**0.583**	0.442	0.438	0.486	0.448	0.435	0.587
Genre reranking	MAP@1000	**0.535**	0.491	0.491	0.493	0.493	0.513	0.626
Artist pair classification	max AP	**0.893**	0.824	0.822	0.849	0.812	0.833	0.935
Average		**0.563**	0.470	0.468	0.490	0.474	0.484	0.630
MoisesDB, out of domain								
Genre classification	accuracy	**0.533**	0.453	0.444	0.409	0.400	0.427	0.698
Genre clustering	V-measure	**0.193**	0.157	0.141	0.155	0.112	0.132	0.351
Artist clustering	V-measure	**0.532**	0.466	0.475	0.479	0.467	0.472	0.612
Artist retrieval	nDCG@10	**0.574**	0.524	0.517	0.517	0.491	0.518	0.693
Genre reranking	MAP@1000	**0.607**	0.588	0.580	0.561	0.557	0.573	0.682
Artist pair classification	max AP	**0.744**	0.697	0.695	0.679	0.648	0.680	0.826
Instrument classification (stems)	accuracy	**0.624**	0.563	0.560	0.480	0.516	0.555	0.695
Average		**0.544**	0.493	0.487	0.469	0.456	0.479	0.651
Upstream MAEB tasks, on their original datasets								
GTZAN genre classification	accuracy	**0.774**	0.531	0.530	0.500	0.529	0.494	0.840
GTZAN genre clustering	V-measure	**0.471**	0.272	0.269	0.311	0.337	0.336	0.679
Music Genre clustering	V-measure	**0.344**	0.126	0.125	0.139	0.182	0.144	0.450
GTZAN genre reranking	MAP@1000	**0.722**	0.555	0.554	0.557	0.588	0.573	0.816
NSynth instrument source	accuracy	**0.453**	0.396	0.396	0.419	0.432	0.406	0.595
Jam-ALT artist retrieval‡	nDCG@10	**0.9672**	0.9601	0.9592	0.9611	0.7603	0.9668	0.923
Average		**0.622**	0.473	0.472	0.481	0.471	0.487	0.717

‡ Autoencoders at four decimals, since SAGE and Music2Latent coincide at three.

reaches 0.387 where the hinge loss stops at 1.048 and drives SI-SDR negative. The lower block of Table 10 moves the same critics between the two phases, and the best moment to introduce one depends on which one it is. The WavTokenizer ensemble is best present from the first step, 0.208 against 0.419 when it arrives only at fine-tuning, whereas the EnCodec-style critic is better introduced late, 0.548 against 0.959; resuming its weights into the second phase rather than re-initializing them collapses every column.

Table 10: Discriminator type and placement on FMA-medium; only the first two rows of the lower block distill a teacher.

		FAD ↓		CLAP ↑		
Discriminator		MERT	CLAP	music	audio	SI-SDR ↑
Type, present from the first step						
WavTokenizer-RpGAN		**0.203**	**0.056**	**0.881**	**0.879**	1.55
EnCodec-RpGAN		0.387	0.122	0.821	0.804	0.55
EnCodec-hinge		1.048	0.265	0.693	0.660	−2.21
none		1.453	0.414	0.645	0.624	**2.47**
Placement across the two phases						
Pre-training	Fine-tuning					
WavTokenizer	WavTokenizer	**0.208**	**0.045**	**0.892**	**0.889**	0.80
none	WavTokenizer	0.419	0.064	0.860	0.864	0.49
none	EnCodec-hinge	0.548	0.130	0.819	0.813	**2.49**
EnCodec-hinge	EnCodec-hinge, reset	0.959	0.200	0.730	0.701	−1.98
EnCodec-hinge	EnCodec-hinge, resumed	2.553	0.450	0.473	0.465	−7.87

Capacity. Table 11 grows the backbone along one axis at a time. Deepening the middle stage from six blocks to eighteen, the step Swin itself takes between its tiny and its small configuration, adds 83M parameters and improves both Fréchet distances and both CLAP similarities. Widening the embedding to $C = 512$ costs four times the parameters and improves FAD-CLAP by a quarter and the two CLAP similarities, conceding a little on FAD-MERT. Neither change asks for a different architecture, and the return per parameter is larger along the depth axis; the configuration we report sits at the compact end of both curves.

Table 11: Capacity on FMA-medium, without the adversarial branch; values compare inside a study, not across the two.

		FAD ↓		CLAP ↑		
Configuration	Params	MERT	CLAP	music	audio	SI-SDR ↑
Width, $\lambda_{\text{mel}} = 0.5$						
$C = 512, (2, 6, 2)$	388M	1.050	**0.259**	**0.700**	**0.720**	1.58
$C = 256, (2, 6, 2)$	97M	**0.963**	0.347	0.680	0.674	**1.68**
Depth, $\lambda_{\text{mel}} = 0.3$						
$C = 256, (2, 18, 2)$	180M	**1.298**	**0.38**	**0.667**	**0.650**	**2.94**
$C = 256, (2, 6, 2)$	97M	1.453	0.41	0.645	0.624	2.47

F Training Objective

Let $a \in \mathbb{R}^{2\times N}$ be a stereo segment with channels L, R, and let $M = L + R$, $S = L - R$; hats mark the reconstructed waveform $\hat{a}$ and the signals derived from it. Throughout, $\|\cdot\|_1$ and $\|\cdot\|_2^2$ denote means over all entries (batch, channels, frequency, time; $|\cdot|^2$ for complex entries), $\|\cdot\|_F$ the Frobenius norm over the frequency-time plane of a single signal, and $\mathbb{E}$ the mean over the batch.

Spectral reconstruction. With $c(Z) = \beta\,|Z|^{\alpha}e^{i\angle Z}$, $\alpha = 0.65$, $\beta = 0.35$, let $\mathbf{X} = c\big(\mathrm{STFT}(a)\big)$ be the power-compressed complex spectrogram of Section 2.1 (window 2048, hop 512, Nyquist bin dropped), whose real and imaginary parts form the encoder input. The decoder predicts $\hat{\mathbf{X}} = D(z)$ in the same compressed domain, the waveform is recovered as $\hat{a} = \mathrm{iSTFT}\big(c^{-1}(\hat{\mathbf{X}})\big)$ with $c^{-1}(Z) = (|Z|/\beta)^{1/\alpha}e^{i\angle Z}$, and

$$\mathcal{L}_{\text{STFT}} = \big\|\hat{\mathbf{X}} - \mathbf{X}\big\|_2^2, \tag{3}$$

which penalizes magnitude and phase jointly.

Mel reconstruction. For windows $w \in \{2048, 1024, 512\}$, Mel_w projects the magnitude STFT (Hann window w, hop $w/4$) onto 128 mel bands from 0 Hz to Nyquist, and with $\epsilon = 10^{-5}$

$$\mathcal{L}_{\text{mel}} = \sum_{w} \Big(\big\|\mathrm{Mel}_w(a) - \mathrm{Mel}_w(\hat{a})\big\|_1 + 2\,\big\|\log_{10}\max\big(\mathrm{Mel}_w(a), \epsilon\big) - \log_{10}\max\big(\mathrm{Mel}_w(\hat{a}), \epsilon\big)\big\|_1 \Big). \tag{4}$$

Sum-and-difference loss. Let Y_n be the STFT (Hann window n, hop $n/4$) of the signal y after A-weighting, with magnitudes floored at 10^{-4}. Over the four signals and six FFT sizes, n doubling from 64 to 2048,

$$\mathcal{L}_{\text{SD}} = \frac{1}{2} \sum_{y\in\{M,S,L,R\}} \frac{1}{6} \sum_{n} \left[\mathbb{E}\, \frac{\big\|\,|\hat{Y}_n| - |Y_n|\,\big\|_F}{\big\|\,|\hat{Y}_n|\,\big\|_F} + \big\|\ln|Y_n| - \ln|\hat{Y}_n|\big\|_1 \right]. \tag{5}$$

Why a squared error neglects the stereo image. Let $e_L = \hat{\mathbf{X}}_L - \mathbf{X}_L$ and $e_R = \hat{\mathbf{X}}_R - \mathbf{X}_R$ be the errors on the two channels. For any two vectors,

$$\|e_L\|_2^2 + \|e_R\|_2^2 = \tfrac{1}{2}\big(\|e_L + e_R\|_2^2 + \|e_L - e_R\|_2^2\big), \tag{6}$$

so $\mathcal{L}_{\text{STFT}}$ weighs the sum and the difference of the channel errors equally, and the difference, which carries the side, enters only through its squared norm, a small share of the total when the side is weak.

On music, where S is typically much weaker than M, even discarding the side barely moves the loss. The spectral-convergence term of $\mathcal{L}_{\mathrm{SD}}$ instead divides by the norm of the reconstructed magnitude of each signal, so a side shrunk towards zero inflates the term, while a near-mono reference keeps it at most about one (Appendix B).

Adversarial terms. We use the discriminator of WavTokenizer, but keep its multi-period branch only once, since its DAC component repeats it. The result has eleven sub-discriminators $\mathcal{D}_1, \ldots, \mathcal{D}_{11}$. Five are multi-period discriminators on the waveform, with periods $2, 3, 5, 7, 11$. Three are multi-resolution discriminators on the magnitude STFT, with rectangular windows of 1024, 2048 and 512 samples and a hop of a quarter of the window. The last three are the multi-band discriminators of DAC on the complex STFT, with FFT sizes 2048, 1024, 512 and five frequency bands each; their input has its DC offset removed and is peak-normalized. The four signals L, R, M, S are folded into the batch, so every sub-discriminator judges each of them as a separate mono signal. Each $\mathcal{D}_b$ returns a map of logits, and the relativistic pairing compares every real segment with its own reconstruction, position by position. With $f(t) = \log(1 + e^t)$,

$$\mathcal{L}_{\mathrm{disc}} = \frac{1}{11} \sum_{b=1}^{11} \mathbb{E}\big[f\big(\mathcal{D}_b(\hat{a}) - \mathcal{D}_b(a)\big)\big], \tag{7}$$

$$\mathcal{L}_{\mathrm{adv}} = \frac{1}{11} \sum_{b=1}^{11} \mathbb{E}\big[f\big(\mathcal{D}_b(a) - \mathcal{D}_b(\hat{a})\big)\big], \tag{8}$$

where the mean runs over the batch and the logit positions. Feature matching compares the K_b hidden feature maps $\mathcal{D}_b^{(k)}$ that sub-discriminator b computes before its logits, averaging the distances first within each sub-discriminator and then across the eleven,

$$\mathcal{L}_{\mathrm{fm}} = \frac{1}{11} \sum_{b=1}^{11} \frac{1}{K_b} \sum_{k=1}^{K_b} \big\| \mathcal{D}_b^{(k)}(a) - \mathcal{D}_b^{(k)}(\hat{a}) \big\|_1. \tag{9}$$