UNIVERSAL MIXUP LOSS: TRAINING AUDIO AUTOENCODERS TO DECODE THE POSITIVE-WEIGHT MIXING CONE

Anonymous authors
Paper under double-blind review

ABSTRACT

Waveform mixing is linear, but mixing the latent codes of an audio autoencoder rarely produces a representation that decodes to the corresponding audio mixture. This limits the use of latent arithmetic for audio editing, source separation, and generative mixing. We show that supervising a few gain and addition identities leaves large regions of the mixing space unconstrained. We introduce the *Universal Mixup Loss*, a decoder-side objective that instead supervises positive-weight combinations of two latent codes throughout their mixing cone. Training points are sampled in polar coordinates, decoupling an annealed radial scale from a uniformly distributed mixing angle. Paired with a scale-sensitive soft-norm bottleneck, this supervision substantially reduces mixing commutation errors while preserving reconstruction quality. The additive noise injected at the bottleneck improves tolerance to the residual codes produced by subtraction and supports decoding at extreme gains. As a result, latent subtraction can recover a source from a mixture when the codes of the interfering sources are available, and decoding remains consistent across interpolation laws absent from training. The resulting latent structure also benefits conditional rectified flow: its velocity field is easier to regress, arithmetic remains effective on generated codes, and high-quality outputs are reached with fewer sampling steps.

1 INTRODUCTION

Every audio production pipeline is, at its core, an exercise in *mixing*: a sound engineer sums stems, applies gains, crossfades between takes, and subtracts a bleed track from a recording. These operations are linear in the waveform domain, and this linearity is precisely what makes them composable, interpretable, and invertible. It is therefore natural to ask whether the latent representations of a neural audio autoencoder could inherit the same property. If scaling a latent representation scaled its signal, and summing two representations summed their sources, then source separation, audio editing, and generative mixing could be expressed through elementary vector algebra on compressed representations, without a task-specific architecture for each operation. This would provide an algebra-preserving interface of the kind that vector arithmetic already offers in embedding and generative latent spaces (Mikolov et al., 2013; Radford et al., 2016; Turner et al., 2024). In practice, audio autoencoders do not generally preserve this relationship. Discrete codecs quantize the bottleneck (van den Oord et al., 2017), making continuous arithmetic on their representations difficult. Continuous autoencoders avoid quantization, but their objectives typically ask the decoder to reconstruct individual signals; they do not directly constrain how it decodes sums and rescalings of latent representations. Final layer normalizations (Ba et al., 2016) can further obscure the magnitude information needed to recover the gain of a mixture. Consequently, a model can reconstruct individual recordings accurately while decoding their latent combinations inconsistently.

Existing implicit-linearity training addresses parts of this problem by supervising a gain-scaled representation and an equal-weight sum of two representations. Mixing, however, spans a continuum of source balances and gains. The set of all positive-weight combinations of two representations forms a two-dimensional convex cone, through which different interpolation laws trace different paths: a linear crossfade, a log-fader curve, a constant-power law, or a path specified by a downstream application. Supervising only selected directions does not directly constrain the decoder throughout

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.

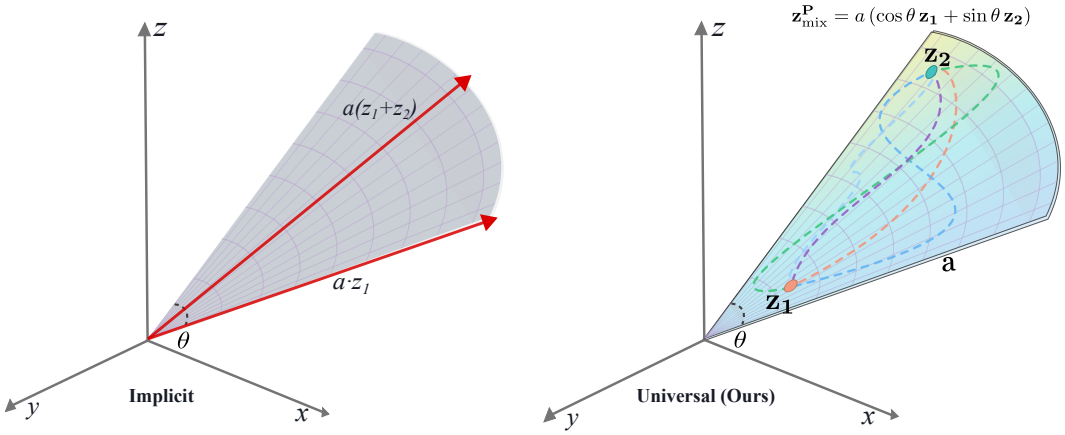


Figure 1: **Geometric comparison of mixing supervision.** Left: Implicit linearity supervises only two discrete rays ($\theta \in \{0, \pi/4\}$). Right: Our Universal Mixup Loss provides uniform coverage over the positive-weight cone.

this cone. The question is therefore how to train a decoder to preserve the correspondence between latent and waveform mixtures across the combinations it may encounter. We address this question by making the positive-weight mixing cone the training distribution. The *Universal Mixup Loss* supervises the decoder on randomly sampled combinations of two latent representations, using polar coordinates to decouple an annealed radial scale from a uniformly distributed mixing angle. This provides continuous angular coverage rather than privileging particular mixing ratios, while radial sampling exposes the decoder to different gains. We apply the explicit mixing constraint at the decoder output: imposing a corresponding constraint on the encoder introduces a trade-off between reconstruction fidelity and latent regularity (Appendix A.6).

Our contributions are as follows:

- **Universal cone supervision.** We introduce the Universal Mixup Loss (Section 3), a decoder-side objective that strictly generalizes prior linearity regularization from two fixed directions of the mixing cone to the whole cone.
- **Theoretical analysis.** We show that pairing the loss with a scale-sensitive soft-norm bottleneck makes the resulting objective interpretable as the intersection of a constant-variance VAE and an ℓ_2-regularized autoencoder. This reading is not merely taxonomic: through an ablation study, we identify the actual role of the bottleneck's additive noise, which is not angular (law-invariance is achieved by the uniform angular sampler alone), but rather the construction of an off-manifold tolerance basin that algebraic source separation relies on, together with radial coverage of the cone at extreme gains (Section 3.4; Appendices A.4 and A.5).
- **Latent algebra as a usable interface.** With this single loss, the commutation errors of a 50.7M-parameter autoencoder drop by one to two orders of magnitude while reconstruction quality improves over the baseline it is built on, fidelity remains flat across interpolation paths the model never followed, and zero-shot, supervised source separation by latent subtraction reaches $+13.7$ dB SI-SDR on MUSDB18-HQ (Rafii et al., 2017) and transfers to unseen corpora and extended sequence lengths (Appendix D.2), where every external baseline falls below -10 dB.
- **Interaction with rectified flow.** Training a rectified-flow prior over the same latent space yields a conditional velocity field that is markedly easier to regress, preserves latent arithmetic on generated codes, and converges in 4 Euler steps instead of 8–16, while maintaining comparable audio quality to a standard non-linear baseline (Section 4.4).

Code, checkpoints, and evaluation artifacts are available at `https://anonymous.4open.science/r/Universal-Mixup-Loss-for-Audio-AE-46D0/README.md`.

2 Related Work

Manifold Regularization and Mixup. Data augmentation strategies based on convex linear combinations, such as Mixup (Zhang et al., 2018), Manifold Mixup (Verma et al., 2019), and between-class training for audio (Tokozume et al., 2018), have been widely adopted across machine learning to smooth decision boundaries and enforce linear transitions between representations. The most direct prior attempt to apply similar concepts to continuous audio autoencoders is the implicit linearity regularizer (Torres et al., 2026), which we examine in detail below.

Neural Audio Codecs and Continuous Autoencoders. Recent breakthroughs in high-fidelity audio synthesis rely heavily on neural audio representations (van den Oord et al., 2016; Engel et al., 2020). Discrete neural codecs achieve low bitrate compression using residual vector quantization (van den Oord et al., 2017; Zeghidour et al., 2022; Défossez et al., 2022; Kumar et al., 2023). While effective, their discrete codebooks fragment the latent geometry, precluding continuous vector arithmetic and causing notable degradation in vector subtraction tasks. Discrete token streams nonetheless anchor the dominant audio language-model stack (Dhariwal et al., 2020; Borsos et al., 2023; Copet et al., 2023), which inherits this fragility. Conversely, continuous latent autoencoders maintain continuous topologies suitable for downstream diffusion priors (Caillon & Esling, 2021; Parker et al., 2026; Liu et al., 2023a; 2024; Evans et al., 2024). Nevertheless, their standard bottlenecks typically break additivity because final layer normalizations erase the magnitude information required for positive weight mixtures.

Relation to Implicit Linearity Regularization. While both frameworks pursue algebraic consistency in continuous autoencoders, the implicit formulation supervises the decoder on only two discrete rays of the latent space: a scalar-gain ray ($\theta = 0$) and an equal-weight sum ($\theta = \pi/4$). As we show formally in Appendix A.7, this formulation is a strictly degenerate special case of the Universal Mixup Loss: it corresponds to replacing our continuous, uniform angular measure over the positive mixing cone with a two-point discrete Dirac distribution. By supervising solely these isolated directions, existing models leave all other positive-weight trajectories entirely out of distribution. In contrast, our formulation integrates supervision across the full continuum of the cone, providing uniform angular training support for every positive-weight direction.

Latent Audio Arithmetic and Source Separation. Conventional source separation systems rely on specialized supervised masking networks (Hershey et al., 2016; Yu et al., 2017; Luo & Mesgarani, 2019; Luo et al., 2020; Luo & Yu, 2023), waveform-domain separators (Défossez et al., 2020; Défossez, 2021; Rouard et al., 2023), or unsupervised mixture invariant objectives (Wisdom et al., 2020). In the supervised setting we study, this removes the need for dedicated separation heads or time-frequency masking. By structuring the latent space to preserve the geometry of the positive-weight mixing cone, we show that zero-shot source separation, given the latent representations of the interfering sources, can emerge through elementary vector subtraction (Section 4.2).

3 Universal Mixup Loss

Mixing supervision should cover the range of source balances and gains that a decoder may encounter. To this end, we propose the *Universal Mixup Loss* ($\mathcal{L}_{\mathrm{UMix}}$), which supervises the decoder of an audio autoencoder over the full positive-weight mixing cone of two latent codes in polar coordinates, so that the decoder learns to approximate the inversion of every mixture on the cone. The polar parameterization is a *design choice* that decouples energy and direction, enables isotropic coverage of the cone, and, when paired with a soft-norm bottleneck, yields an objective that admits an interpretation as the intersection of a constant-variance VAE (CV-VAE) (Kingma & Welling, 2014; Rezende et al., 2014) and an ℓ_2-regularized autoencoder (RAE) (Ghosh et al., 2020), coupled with a geometric training distribution over the latent cone.

3.1 Setup and notation

Let $\mathcal{E}_\psi : \mathbb{R}^T \to \mathbb{R}^P$ be an encoder and $\mathcal{D}_\phi : \mathbb{R}^P \to \mathbb{R}^T$ a decoder, with parameters ψ, ϕ and flat latent codes $\mathbf{z} \in \mathbb{R}^P$, where P denotes the product of latent channels and time tokens. We denote by $\mathcal{L}_{\mathrm{rec}} : \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}_{\geq 0}$ the multi-resolution spectral reconstruction loss of Parker et al. (2026), which is Parseval-consistent and behaves as a weighted $\|\cdot\|^2$ energy measure; this property

is leveraged in Section 3.4. In-batch pairs are formed via a circular shift of the batch axis. Writing $\mathbf{X}$ and $\mathbf{Z}$ for the batches of signals and latent codes, respectively, a sample $\mathbf{x}_i$ with latent code $\mathbf{z}_i = \mathcal{E}_\psi(\mathbf{x}_i)$ is paired with $\mathbf{x}_{i+} = \mathrm{roll}\mathbf{X}, 1)_i$ and $\mathbf{z}_{i+} = \mathrm{roll}(\mathbf{Z}, 1)_i$. When clear from context, we omit the index i.

Scale-sensitive bottleneck. To preserve the magnitude information required for positive-weight mixtures, we omit the final layer normalization in both the encoder and decoder. This ensures the bottleneck is *scale-sensitive*, so that scaling and summing latent codes survive the projection, as we show in Appendix C.2. This is a prerequisite for the loss introduced in this section, as it allows the decoder to infer output scale from latent magnitude alone.

3.2 The positive-weight cone and its polar parameterization

For a pair of latent codes $(\mathbf{z}_1, \mathbf{z}_2) \in \mathbb{R}^P \times \mathbb{R}^P$, we define the *positive-weight mixing cone* as the closed convex cone

$$\mathcal{C}(\mathbf{z}_1, \mathbf{z}_2) = \{w_1\mathbf{z}_1 + w_2\mathbf{z}_2 \mid w_1, w_2 \geq 0\} \subset \mathbb{R}^P. \tag{1}$$

Two parameterizations of $\mathcal{C}(\mathbf{z}_1, \mathbf{z}_2)$ are relevant to our analysis. The *Cartesian* parameterization expresses a mixture as

$$\mathbf{z}^C_{\mathrm{mix}} = w_1\mathbf{z}_1 + w_2\mathbf{z}_2, \qquad w_1, w_2 \geq 0, \tag{C}$$

while the *polar* parameterization decouples the radial scale $a \geq 0$ from the mixing direction $\theta \in [0, \pi/2]$:

$$\mathbf{z}^P_{\mathrm{mix}} = a\left(\cos\theta\,\mathbf{z}_1 + \sin\theta\,\mathbf{z}_2\right), \qquad a \geq 0, \quad \theta \in \left[0, \frac{\pi}{2}\right]. \tag{P}$$

For $a > 0$, the map $(w_1, w_2) \mapsto (a, \theta)$ defined by $a = \sqrt{w_1^2 + w_2^2}$ and $\theta = \mathrm{atan2}(w_2, w_1)$ is a bijection between the non-negative quadrant excluding the origin and the set $\{a > 0, \theta \in [0, \pi/2]\}$; at $a = 0$, every value of θ represents the origin. Consequently, Eq. (C) and Eq. (P) span the same set $\mathcal{C}(\mathbf{z}_1, \mathbf{z}_2)$. While the two parameterizations describe identical point sets, they induce distinct training distributions; the difference is analyzed in Appendix A.1 and tested empirically in Appendix A.3.

3.3 The Universal Mixup objective

The Universal Mixup Loss supervises the decoder on random points of the cone $\mathcal{C}(\mathbf{z}_1, \mathbf{z}_2)$, sampled in polar coordinates (a, θ). Formally, for a pair of audio signals $(\mathbf{x}_1, \mathbf{x}_2)$ with corresponding latent codes $\mathbf{z}_i = \mathcal{E}_\psi(\mathbf{x}_i)$, we define

$$\mathcal{L}_{\mathrm{UMix}}(\phi, \psi) = \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2)}\mathbb{E}_{a,\theta}\,\mathcal{L}_{\mathrm{rec}}\Big(\mathcal{D}_\phi\big(a(\cos\theta\,\mathbf{z}_1 + \sin\theta\,\mathbf{z}_2)\big),\ \mathrm{clip}_1\big(a(\cos\theta\,\mathbf{x}_1 + \sin\theta\,\mathbf{x}_2)\big)\Big), \tag{2}$$

where $\mathrm{clip}_1 : \mathbb{R}^T \to [-1, 1]^T$ clips waveform amplitudes element-wise to the valid range. The mixing parameters (a, θ) are sampled from

$$\theta \sim \begin{cases} 0 & \text{with probability } 1 - m, \\ \mathcal{U}([0, \frac{\pi}{2}]) & \text{with probability } m, \end{cases} \qquad a \sim \begin{cases} \mathcal{U}([s_{\min}, s_{\max}]) & \text{with probability } p_g, \\ 1 & \text{with probability } 1 - p_g, \end{cases}, \tag{3}$$

with $a \leftarrow 0$ if $a < \delta$, and where the default hyperparameters are $s_{\min} = 0$, $s_{\max} = 3$, $p_g = 0.8$, $m = 0.5$, and $\delta = 0.05$. The gain range is cosine-annealed to $(1, 1)$ over training. Thus, a fraction $1 - m$ of each batch uses $\theta = 0$ (i.e., $\mathbf{z}_{\mathrm{mix}} = a\mathbf{z}_1$), ensuring pure gain-scaled reconstruction is explicitly included, while the remaining fraction samples θ uniformly from $[0, \pi/2]$. The full generator objective combines reconstruction, adversarial (Goodfellow et al., 2014), mixup, and KL terms:

$$\mathcal{L}_{\mathrm{gen}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{UMix}}\mathcal{L}_{\mathrm{UMix}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}, \tag{4}$$

with $\mathcal{L}_{\mathrm{KL}}$ defined in Appendix A.4; training details are in Appendix B.1.

Functional constraint on additivity. If $\mathcal{D}_\phi$ were exactly additive and homogeneous on $\mathcal{C}(\mathbf{z}_1, \mathbf{z}_2)$, reconstructed both inputs exactly, and the target mixture remained unclipped, then

$$\mathcal{D}_\phi(a(\cos\theta\,\mathbf{z}_1 + \sin\theta\,\mathbf{z}_2)) = a(\cos\theta\,\mathcal{D}_\phi(\mathbf{z}_1) + \sin\theta\,\mathcal{D}_\phi(\mathbf{z}_2)) = a(\cos\theta\,\mathbf{x}_1 + \sin\theta\,\mathbf{x}_2),$$

and $\mathcal{L}_{\text{UMix}}$ would vanish. Thus, $\mathcal{L}_{\text{UMix}}$ imposes a soft functional constraint on decoder additivity over the entire cone: it penalizes departures from additive, homogeneous decoding wherever the sampler visits the cone. While a symmetric linearity constraint could naturally be mirrored on the encoder side, enforcing it directly induces a relevant trade-off between reconstruction fidelity and latent manifold regularity, tending to collapse the code distribution to a low-rank subspace. We therefore restrict cone supervision strictly to the decoder, and analyze this encoder-side trade-off in Appendix A.6. We stress that *decoder-only* refers to the constraint, not to the parameter updates: the implementation does not detach the codes, so the mixup gradient backpropagates through the composite latent into the encoder. What the objective omits is an explicit encoder-commutation penalty, $\mathcal{E}_\psi(w_1\mathbf{x}_1 + w_2\mathbf{x}_2) \approx w_1\mathcal{E}_\psi(\mathbf{x}_1) + w_2\mathcal{E}_\psi(\mathbf{x}_2)$, rather than encoder participation in training. The gradient decomposition is provided in Appendix A.1.

Isotropic cone coverage. In the mixup component of Eq. (3), θ is sampled uniformly from $[0, \pi/2]$, so the angular support of the training distribution spans all positive-weight mixing directions: every sector of $\mathcal{C}(\mathbf{z}_1, \mathbf{z}_2)$ receives training mass, and no direction is privileged within this component. Consequently, the decoder is encouraged to invert all positive-weight interpolation paths (e.g., linear, log-fader, constant-power, parabolic), which discourages the memorization of any single interpolation law. This uniform angular support is the core design choice behind the universality of the objective, and it is what Section 4.3 tests on path parameterizations the model never followed.

3.4 The structural role of noise in the soft-norm bottleneck

Our bottleneck inherits the soft-norm design of Parker et al. (2026), which incorporates additive Gaussian noise at the decoder input. While originally motivated by robustness to downstream diffusion models, this design admits an interpretation, at the level of loss terms, as a constant-variance VAE (Kingma & Welling, 2014; Ghosh et al., 2020) while sidestepping the well-known training pathologies of the full formulation (Alemi et al., 2018). We derive this interpretation, along with its exact loss-term decomposition, in Appendix A.4. Notably, this formulation yields a noise corruption floor that is invariant to the mixing angle θ.

Empirically, our ablations (Appendix A.5) reveal that this constant-variance noise is not required for interpolation law-invariance, but it serves two distinct structural roles. First, it builds an off-manifold *tolerance basin*: without noise, algebraic source separation degrades substantially (+7.8 dB drops to -2.6 dB), as the decoder becomes brittle to the structured residual of latent subtraction. Second, its fixed variance for individual codes supports robustness across radial scales: the model retains mixing fidelity at extreme gains (near-silence and $a \geq 2.2$), where the training sampler is sparse or annealed away. The fixed-variance stochastic injection is therefore the mechanism that our ablations identify as extending latent arithmetic beyond the explicitly supervised unit-gain shell.

4 Experiments

Our evaluation tests whether full-cone supervision reduces commutation errors without harming reconstruction, enables zero-shot algebraic operations (crossfading and source separation via subtraction), and facilitates downstream rectified-flow generation. We evaluate a shared 50.7 M-parameter Transformer backbone (width 512, 64-channel continuous bottleneck) under two complementary regimes: an *internal 16 kHz benchmark* for controlled ablations across Minimal, Soft-Norm, and Universal variants; and a *44.1 kHz high-fidelity regime* fine-tuned from the 16 kHz checkpoint to benchmark against the production codecs SAME-L and DAC, together with the pretrained linear CAE of Torres et al. (2026) (Lin-CAE), the closest prior approach to audio latent linearity. Architecture specifications, training schedules, and metrics are detailed in Appendices B.1 and B.2. To ensure the robustness of our findings, generative training and evaluations were conducted across multiple independent random seeds. Primary comparisons are assessed with paired statistical tests, with the comprehensive framework and p-values deferred to Appendix B.3; where a difference is not statistically robust, we report it as such (Section 4.4).

Controlled Training Baselines. To isolate the benefit of full-cone supervision, we also retrained our exact backbone with the implicit linearity loss of Torres et al. (2026) (supervising only scalar scaling and equal-weight sum rays) and a constant-power arc variant (*Trig mixup*, restricted to unit

Table 1: Additivity metric suite for the three core configurations (2 s clips, 16 kHz, seed 42). Encoder/decoder and NormErr entries: mean ± std over 1000 validation pairs. MS-STFT: mean ± SEM over 2000 reconstruction clips. Conv and Eff-Rank: pool statistics ± bootstrap std (B=500, 256 clips). Lower is better (↓) except for effective rank (↑). Bold marks the best entry per category, underline indicates second-best.

Metric	Minimal	Soft-Norm	Universal
Reconstruction			
MS-STFT ↓	$\mathbf{0.534 \pm 0.013}$	0.617 ± 0.014	$\underline{0.601 \pm 0.014}$
Encoder			
single ↓	48.33 ± 21.40	$\underline{1.688 \pm 0.216}$	$\mathbf{1.375 \pm 0.228}$
lerp ↓	2.211 ± 1.672	$\underline{0.798 \pm 0.170}$	$\mathbf{0.699 \pm 0.123}$
smooth ↓	4.3 ± 10.3	$\underline{0.147 \pm 0.078}$	$\mathbf{0.126 \pm 0.091}$
Decoder ($\times 10^{-4}$)			
single ↓	347.98 ± 334.24	$\underline{323.82 \pm 277.66}$	$\mathbf{28.19 \pm 70.42}$
lerp ↓	$\underline{16.28 \pm 25.60}$	308.53 ± 59.38	$\mathbf{1.11 \pm 2.01}$
smooth ↓	$\underline{(1.7 \pm 12) \times 10^{-2}}$	1.40 ± 0.62	$\mathbf{(8.1 \pm 60) \times 10^{-5}}$
Latent Space			
NormErr ↓	202.59 ± 102.56	$\mathbf{2.68 \pm 2.06}$	$\underline{3.40 \pm 3.14}$
Conv ↓	224.38 ± 1.55	$\underline{112.65 \pm 0.13}$	$\mathbf{105.93 \pm 0.44}$
Eff-Rank ↑	9.03 ± 1.01	$\mathbf{107.20 \pm 2.15}$	$\underline{78.28 \pm 3.59}$

energy); the pretrained counterpart of the implicit-linearity model appears as the external baseline *Lin-CAE* in Sections 4.2 and 4.3. As detailed in Appendix A.7 (with full benchmarks in Appendices C and D), neither discrete rays nor 1D arcs prevent algebraic degradation across the full mixing cone, where our cone supervision remains the most robust.

4.1 Additivity of the Encoder-Decoder Interface

To assess how mixing supervision shapes the latent space, we evaluate three internal 16 kHz configurations: *Minimal* (unregularized reconstruction bound), *Soft-Norm* (generative regularization without mixing supervision), and *Universal* (combining soft-norm with our cone objective). All paired differences discussed below are statistically significant at $p < 10^{-4}$ (Table 15, Appendix B.3).

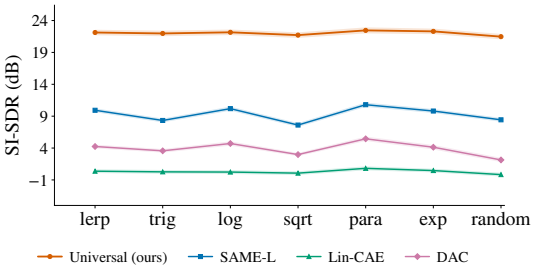


Figure 2: Interpolation fidelity across mixing laws (600 MUSDB18-HQ pairs, 44.1 kHz). Our cone supervision maintains flat fidelity across unseen paths, whereas external baselines fluctuate or degrade.

Commutation behaviour. As reported in Table 1, cone supervision reduces decoder commutation errors by two orders of magnitude (e.g., lerp error drops from 308.53 to 1.11×10^{-4}). Pure reconstruction (*Minimal*) fails to commute with mixing, while unguided regularization (*Soft-Norm*) distorts magnitude algebra further. A parallel gain occurs in the encoder, yielding smoother latent paths and a closer approximation to an interpolation homomorphism. Importantly, this algebraic compliance improves acoustic reconstruction relative to the soft-norm base (MS-STFT 0.601 vs. 0.617). While the unconstrained Minimal baseline achieves lower MS-STFT (0.534), it suffers from dimensionality collapse (Eff-Rank = 9.03) and non-functional algebra, rendering it unsuitable as a generative prior. Relative to the regularized continuous baselines tested, cone supervision offers a favorable trade-off between algebraic compliance and reconstruction quality, and remains robust across compression factors (Appendix C.3).

Table 2: Zero-shot, supervised algebraic source separation on 100 held-out MUSDB18-HQ multi-tracks at 44.1 kHz (seed 42). Metrics are mean ± SEM over non-silent trials. Best results in **bold**, second best underlined. Full stem breakdown in Appendix D.1.

	Overall		Drums		Vocals	
Model	**SI-SDR ↑**	**MS-STFT ↓**	**SI-SDR ↑**	**MS-STFT ↓**	**SI-SDR ↑**	**MS-STFT ↓**
Universal (Ours)	$\mathbf{+13.71 \pm 0.73}$	$\mathbf{0.013 \pm 0.001}$	$\mathbf{+10.04 \pm 1.29}$	$\mathbf{0.018 \pm 0.002}$	$\mathbf{+9.42 \pm 2.08}$	$\mathbf{0.012 \pm 0.001}$
SAME-L	$\underline{-10.65 \pm 0.64}$	2.075 ± 0.061	$\underline{-10.59 \pm 1.39}$	1.794 ± 0.081	-14.50 ± 1.68	2.203 ± 0.109
Lin-CAE	-14.88 ± 1.04	$\underline{0.896 \pm 0.073}$	-17.27 ± 2.28	$\underline{1.139 \pm 0.204}$	$\underline{-12.02 \pm 2.03}$	$\underline{0.472 \pm 0.054}$
DAC	-16.62 ± 0.48	82.72 ± 0.47	-19.73 ± 0.78	81.13 ± 0.99	-18.34 ± 1.54	83.47 ± 0.84

Latent geometry and downstream readiness. Latent regularity and algebraic additivity often conflict. Although *Soft-Norm* in isolation yields the lowest norm error, it significantly degrades linear compositionality. The *Universal* model reconciles this trade-off: it preserves a well-dispersed representation (Eff-Rank $= 78.28$) and controlled norm error while improving manifold convexity (Conv $= 105.93$). This geometric balance avoids out-of-distribution voids without sacrificing algebraic properties, establishing our Universal model as the reference configuration scaled to 44.1 kHz for external benchmarking.

4.2 Supervised algebraic source separation

If an audio autoencoder learns an algebraically additive interface, source separation emerges directly via vector subtraction, without task-specific training, separation heads, or time-frequency masks. We evaluate this setting as a diagnostic of latent algebra rather than as a blind-separation benchmark, on multi-track mixtures ($x_{\text{mix}} = \sum_k s_k$): given the known latent representations of interfering sources, we isolate a target stem s_{target} by passing the algebraic residual through the unconstrained decoder:

$$z_{\text{sep}} = z_{\text{mix}} - \sum_{k \neq \text{target}} z_k, \qquad \hat{s}_{\text{target}} = D\big(z_{\text{sep}}\big). \tag{5}$$

Because the acoustic inputs are exact linear sums, any degradation in $\hat{s}_{\text{target}}$ reflects non-additivity in latent space or off-manifold decoder operations. Since the pipeline requires access to the interfering stems, its scores are not comparable to those of task-trained blind separators (Défossez, 2021); the unsupervised setting is studied in Section 4.4.

We evaluate this zero-shot pipeline on 100 held-out MUSDB18-HQ (Rafii et al., 2017; Stöter et al., 2018) multi-tracks at 44.1 kHz (400 trials across 4 stems) using SI-SDR and multi-scale STFT MSE (Table 2). As a cross-domain verification, we also report speech denoising across diverse colored noises on 10 s clips from LJSpeech in Appendix D.2.

Separation fidelity and baseline collapse. Our Universal model achieves $+13.71$ dB overall SI-SDR and 0.013 MS-STFT error, successfully recovering complex signals such as vocals ($+9.42$ dB) and percussion ($+10.04$ dB). In contrast, all external baselines collapse below -10 dB, leaving interference and acoustic artifacts to dominate the reconstructed signal. This indicates that representations not trained for cone coverage, whether production codecs or the pretrained implicit-linearity model, lack the geometric structure needed for subtraction. When controlling for architecture on our 16 kHz backbone (Appendix D), the constant-power Trig mixup also separates (12.31 dB) but full cone supervision adds $+1.4$ dB with $25\times$ lower spectral error, while fixed-direction training (Torres et al.) degrades on attenuated interior paths (Appendix A.7). Notably, this algebraic separation generalizes well to unseen acoustic domains: as detailed in Appendix D.2, zero-shot speech denoising on LJSpeech reaches $+16.19$ dB SI-SDR, outperforming the best external baseline by nearly 13 dB. Complete stem results and baseline limitation analyses are provided in Appendix D.1.

4.3 Held-out interpolation paths

Sampling the mixing angle uniformly over $\theta \sim \mathcal{U}([0, \pi/2])$ gives the training distribution full angular coverage of the positive cone $\mathcal{C}(z_1, z_2)$. Since the interpolation paths evaluated below lie inside this supervised region, we test whether reconstruction fidelity remains flat across path parameterizations never encountered during training.

Setup and path evaluation. An interpolation law is defined by non-negative weighting functions $\Lambda = (w_1, w_2) : [0,1] \to \mathbb{R}^2_{\geq 0}$ (Appendix E). We evaluate 600 stem pairs from MUSDB18-HQ across nine evenly spaced points $\alpha \in [0,1]$. Applying the law in latent space yields $z_\Lambda(\alpha) = w_1(\alpha) z_A + w_2(\alpha) z_B$, which is decoded and scored against the acoustic ground truth $x_\Lambda(\alpha) = w_1(\alpha) x_A + w_2(\alpha) x_B$ using SI-SDR and MS-STFT error. To quantify trajectory robustness across the law set $\mathfrak{L}$, we define the invariance index $\mathcal{I}_d = \mathrm{std}_{\Lambda \in \mathfrak{L}}[\bar{d}_\Lambda]$, where $\bar{d}_\Lambda$ is the mean metric along path Λ.

Fidelity and path invariance. As illustrated in Figure 2 (with full numerical breakdowns deferred to Table 24 in Appendix E), the Universal model exhibits flat fidelity across diverse interpolation trajectories. Its spectral error remains stable across paths ($\mathcal{I}_{\mathrm{STFT}} = 0.0006$), representing a $53\times$ higher consistency than SAME-L and $157\times$ over DAC. Moreover, its mean SI-SDR of 22.0 dB across paths sits within 1.7 dB of its single-stem reconstruction anchor (23.8 dB). In contrast, external baselines fluctuate by several dB depending on trajectory geometry: DAC suffers a $3\times$ spectral error spike between linear and trigonometric laws, while Lin-CAE collapses near 0 dB across all settings. Because all evaluated trajectories reside strictly inside $\mathcal{C}(z_1, z_2)$, these behaviors suggest that decoders lacking continuous cone supervision tend to degrade along directions they were never supervised on.

4.4 Cone structure under rectified flow

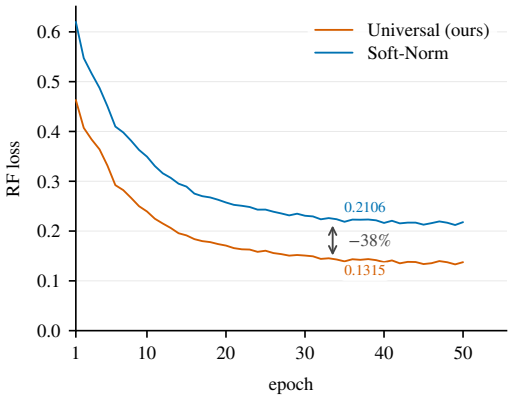


Figure 3: Rectified-flow validation loss. The cone latent yields a markedly easier conditional velocity field, converging to a 38% lower loss than the baseline.

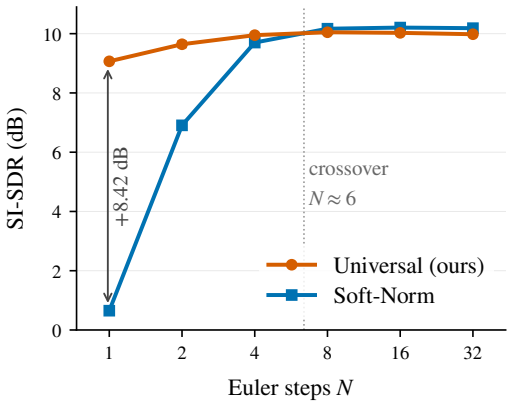


Figure 4: Sampling efficiency across Euler steps. `univ` converges by $N = 4$ and substantially outperforms the baseline in low-budget regimes.

Section 4.2 established zero-shot algebraic source separation with stem access. Here, we evaluate whether cone geometry facilitates unassisted generative modelling in latent space, the generative route that diffusion-based speech enhancement pursues in the waveform domain (Richter et al., 2023); as in Section 4.2, the evaluation probes latent geometry and the survival of latent arithmetic, not competitive blind-separation quality. We train conditional rectified-flow transformers (Liu et al., 2023b; Lipman et al., 2023; Peebles & Xie, 2023), the latent-space counterpart of diffusion-based generative priors (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022), to isolate voice latents from mixtures, comparing our cone-supervised space (`univ`) against the soft-norm baseline (`softnorm`) under identical architectures, data, and budgets (Appendix F.1).

Field geometry and few-step sampling. The conditional velocity field is markedly easier to learn on the cone latent. Validation flow loss for `univ` is 38% lower than `softnorm` (0.1315 vs. 0.2106; Figure 3), and the trajectory is measurably straighter ($S = 0.9869$ vs. 0.9806; Appendix F.2). This geometric regularity translates into substantial few-step efficiency (Salimans & Ho, 2022; Song et al., 2023): across 3 DiT seeds (Appendix F.3), `univ` outperforms `softnorm` by $+8.42$ dB at a single Euler step (9.07 dB vs. 0.65 dB) and nears convergence by $N = 4$ (9.95 dB; Figure 4). In contrast, `softnorm` collapses at low budgets and requires 8 to 16 steps to reach parity. At large

Table 3: Unsupervised voice latent recovery (2 s clips, 3 DiT seeds). The *regressor* estimates the conditional mean in a single pass. The *latavg* row decodes the algebraic mean of generated codes to test arithmetic survival. Complete ladders are in Appendix F.2.

	SI-SDR (dB) ↑		MS-STFT ↓	
Estimator	`univ`	`softnorm`	`univ`	`softnorm`
Reference bounds (no generative model)				
ceiling, $D_\phi(z_{\text{voice}})$	17.53	16.50	0.336	0.358
Latent subtraction, $D_\phi(z_{\text{mix}} - \Sigma_k z_k)$	16.70	−4.48	0.503	5.698
no-op mixture floor, $D_\phi(z_{\text{mix}})$	2.72	2.58	3.120	3.127
Flow matching & algebraic properties				
FM, single draw ($N = 32$)	8.04 ± 0.02	8.26 ± 0.14	1.487	1.419
FM, *latavg* (mean of generated codes)	$\mathbf{9.96 \pm 0.03}$	0.54 ± 0.04	**1.248**	4.458
Regressor, $f_\theta(z_{\text{mix}})$ (conditional mean)	$\mathbf{6.71 \pm 0.06}$	3.42 ± 0.08	**1.769**	2.727

budgets ($N \geq 8$), `softnorm` shows a small audio margin (+0.20 dB at $N = 32$) whose per-seed significance is inconsistent (e.g., $p = 0.21$ in Seed 42), which we read as practical equivalence rather than a meaningful advantage (Appendix B.3). Furthermore, both models eventually exhibit a performance decline at higher steps; `softnorm` simply experiences this drop later and milder (from 9.15 dB at $N = 8$ down to 8.82 dB at $N = 32$), as discussed in Appendix F.5.

Conditional mean estimation and generative arithmetic. The clearest evidence of latent linearity appears when the flow model is replaced by a deterministic conditional mean regressor (Table 3). On `univ`, this single-pass estimator acts as an effective separator (+4.0 dB over mixture floor), whereas on `softnorm` it improves only marginally over the mixture floor at 2 s and falls below the floor at 10 s (Appendix F.2), indicating that its conditional mean is not a reliably decodable state. This result is structural in our experiments: it persists when scaling up capacity, as expanding the `softnorm` regressor by 7.8× (to 129 M parameters) still leaves it 5.25 dB below our lightweight 16.5 M generative model (Appendix F.4). Furthermore, algebraic properties survive the generative process: averaging K generated codes prior to decoding (*latavg*) yields 9.96 dB on `univ`, but collapses to 0.54 dB on `softnorm`. Thus, cone supervision preserves algebraic operations even on generated representations.

5 Conclusion

In this work, we presented an in-depth geometric investigation into latent linearity for continuous audio autoencoders, examining how bottleneck constraints and mixing objectives shape the invertibility of linear operations. Through this analysis, we introduced the Universal Mixup Loss. Our results indicate that a faithful algebraic interface is not obtained by verifying a few algebraic identities: it calls for coverage of the mixtures the decoder must invert, continuous and isotropic over the positive-weight cone rather than restricted to discrete rays or fixed-energy trajectories, together with tolerance to the residual codes that subtraction produces. By combining uniform angular sampling with annealed radial scaling, and anchoring the decoder on a scale-sensitive soft-norm bottleneck, the resulting space yields an angle-independent noise floor that supports robust off-manifold decoding. Empirically, this comprehensive coverage delivers two orders of magnitude lower commutation errors, unlocks supervised zero-shot algebraic stem separation via latent subtraction on unseen multi-tracks, and maintains flat fidelity across held-out continuous interpolation laws. In downstream generative modeling, the structured latent space renders the conditional velocity field straighter, enabling substantial few-step sampling efficiency under rectified flow while preserving algebraic properties across generated representations. While our analysis highlights distinct operating trade-offs at asymptotic sampling budgets and decoder-only supervision (discussed further in Appendices A.6 and F.5), our findings delineate geometric principles that can make latent algebra a reliable primitive for generative audio workflows.

USE OF GENERATIVE AI

In this work, we used generative AI tools (specifically large language models) for tasks with recommended disclosure, strictly limited to retrieving and summarizing related literature, brainstorming narrative structures, and editing the manuscript to improve readability and grammatical precision. All AI-assisted outputs were thoroughly vetted, fact-checked, and revised by the authors against the primary literature and experimental artifacts. The authors take full responsibility for the entire content of this work, including all text, claims, mathematical formulations, and results.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, all core components of this work are thoroughly documented across the manuscript and the supplementary material. Formal mathematical derivations, the probabilistic formulation of the bottleneck, and gradient decompositions are detailed in Appendix A. Full specifications of the network architectures, optimization schedules, training datasets, and preprocessing steps are provided in Appendix B.1, while exact definitions of all evaluation metrics and statistical testing procedures are given in Appendices B.2 and B.3. Downstream generative modeling setups and baseline configurations are comprehensively described in Appendix F.1. Furthermore, complete source code, training and evaluation scripts, and model checkpoints are available in the accompanying repository referenced at the end of Section 1.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. In *International Conference on Machine Learning (ICML)*, volume 80, pp. 159–168, 2018.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023.

Antoine Caillon and Philippe Esling. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. In *International Conference on Digital Audio Effects (DAFx)*, pp. 278–285, 2021.

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 (NeurIPS)*, volume 36, pp. 27768–27791, 2023.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain. In *International Conference on Learning Representations (ICLR)*, 2020.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable digital signal processing. In *International Conference on Learning Representations (ICLR)*, 2020.

Yariv Ephraim and David Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, 1984.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

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

Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. In *International Conference on Learning Representations (ICLR)*, 2020.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pp. 2672–2680, 2014.

John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35, 2016.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Tero Karras, Miika Aittamäki, Samuli Hellsten, Sami Laine, Jaakko Lehtinen, and Timo Aila. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. *arXiv preprint arXiv:2306.06546*, 2023.

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 (ICASSP)*, pp. 626–630, 2019.

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 (ICLR)*, 2023.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning (ICML)*, volume 202, pp. 21450–21474, 2023a.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27 (8):1256–1266, 2019.

Yi Luo and Jianwei Yu. Music source separation with band-split RNN. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901, 2023.

Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 46–50, 2020.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 746–751, 2013.

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. Stability AI.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning (ICML)*, volume 202, pp. 28100–28127, 2023.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016.

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

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, volume 32, pp. 1278–1286, 2014.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2351–2364, 2023.

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)*, pp. 10674–10685, 2022.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*, 2018.

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, volume 202, pp. 32211–32252, 2023.

Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In *Latent Variable Analysis and Signal Separation (LVA/ICA)*, volume 10891 of *Lecture Notes in Computer Science*, pp. 293–305. Springer, 2018.

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

Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Learning from between-class examples for deep sound recognition. In *International Conference on Learning Representations (ICLR)*, 2018.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Bernardo Torres, Manuel Moussallam, and Gabriel Meseguer-Brocal. Learning linearity in audio consistency autoencoders via implicit regularization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. arXiv:2510.23530.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 6306–6315, 2017.

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

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). University of Edinburgh, The Centre for Speech Technology Research (CSTR), 2017.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning (ICML)*, 2019.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron J. Weiss, Kevin Wilson, and John R. Hershey. Unsupervised sound separation using mixture invariant training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 3846–3857, 2020.

Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training for robust monaural speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

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

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1583–1594, 2020.

CONTENTS OF APPENDICES

To facilitate navigation, the supplementary material is organized as follows:

- **Appendix A: Theoretical Derivations and Method Ablations.** Details on backpropagation, computational cost, sampler behavior, the role of CV-VAE noise, and comparisons to implicit linearity.
- **Appendix B: Global Experimental pipeline.** Network architectures, training schedules, metric formulations, and the statistical significance framework.
- **Appendix C: Extended Results for Experiment 1 (Additivity).** Full ablation tables and compression-factor scaling.
- **Appendix D: Extended Results for Experiment 2 (Source Separation).** MUSDB18-HQ baselines, reconstruction anchors, and zero-shot speech denoising on LJSpeech.
- **Appendix E: Extended Results for Experiment 3 (Interpolation Laws).** Formal definitions of the evaluated trajectories and full numerical results.
- **Appendix F: Extended Results for Experiment 4 (Rectified Flow).** DiT architecture details, exhaustive step sweeps, cross-length controls, and decoder sensitivity analyses.

A THEORETICAL DERIVATIONS AND METHOD ABLATIONS

A.1 BACKPROPAGATION: PARAMETERIZATION AND SAMPLING DISTRIBUTION

We analyze how the choice of cone parameterization affects backpropagation. While the polar and Cartesian formulations describe the identical conical point set $\mathcal{C}(z_1, z_2)$, they expose distinct gradient coordinate systems and induce different sampling measures. Notably, the functional loss gradients at any fixed point in the cone are parameterization-invariant.

Latent gradient decomposition. The forward evaluation graph for Universal Mixup is

$$\ell = \ell_{\mathrm{rec}}\big(\hat{y},\ x_{\mathrm{tgt}}\big), \qquad \hat{y} = D_\phi\big(z_{\mathrm{mix}}\big), \qquad z_{\mathrm{mix}} = w_1 z_1 + w_2 z_2 = a\big(\cos\theta\, z_1 + \sin\theta\, z_2\big), \quad (6)$$

where $z_i = E_\psi(x_i)$, and x_{tgt} denotes the clipped target mixture in waveform space. Backpropagation through the decoder maps the loss gradient $\nabla_{\hat{y}}\ell$ back to the decoder input space:

$$g \ \triangleq\ \nabla_{z_{\mathrm{mix}}}\ell \ =\ \Big[\frac{\partial D_\phi(z_{\mathrm{mix}})}{\partial z_{\mathrm{mix}}}\Big]^\top \nabla_{\hat{y}}\ell. \quad (7)$$

All downstream gradients represent direct projections of g. Specifically, evaluating gradients with respect to latent codes z does not bypass the decoder backpropagation; rather, g represents the decoder's backpropagated loss vector expressed directly in the latent domain.

Gradients at fixed points. Applying the chain rule through Eq. (6), the decoder parameter update is given by standard backpropagation:

$$\nabla_\phi \ell = \Big[\frac{\partial D_\phi(z_{\text{mix}})}{\partial \phi}\Big]^\top \nabla_{\hat{y}} \ell, \tag{8}$$

where z_{mix} acts purely as a fixed evaluation input ($\partial z_i/\partial \phi = 0$). Propagating g backward through the mixing layer yields the branch gradients:

$$\text{Cartesian:} \quad \nabla_{z_1}\ell\big|_C = w_1\, g, \qquad \nabla_{z_2}\ell\big|_C = w_2\, g, \tag{9}$$

$$\text{Polar:} \quad \nabla_{z_1}\ell\big|_P = a\cos\theta\, g, \qquad \nabla_{z_2}\ell\big|_P = a\sin\theta\, g. \tag{10}$$

These propagate into the encoder as $\nabla_\psi \ell = \sum_i (\partial z_i/\partial \psi)^\top \nabla_{z_i}\ell$. Because latent codes are not detached in our implementation, this path actively updates the encoder. The objective is "decoder-only" strictly in the sense that it imposes no explicit algebraic homomorphism penalty on the encoder (Appendix A.6).

At any matched point on $\mathcal{C}(z_1, z_2)$ where $w_1 = a\cos\theta$ and $w_2 = a\sin\theta$, the parameter gradients across both parameterizations are identical: Eq. (8) shares the same forward value z_{mix}, and Eq. (9)–Eq. (10) evaluate to identical branch vectors. Parameterization re-indexes coordinates without altering fixed-point optimization dynamics.

Gradients with respect to mixing coordinates. Because mixing coefficients are sampled non-learnable variables, their partial derivatives serve as diagnostic signals and curriculum indicators rather than SGD updates:

$$\frac{\partial \ell}{\partial a} = \big\langle g,\ \cos\theta\, z_1 + \sin\theta\, z_2 \big\rangle = \frac{1}{a}\langle g, z_{\text{mix}}\rangle \quad (\text{for } a > 0), \tag{11}$$

$$\frac{\partial \ell}{\partial \theta} = a\,\big\langle g,\ -\sin\theta\, z_1 + \cos\theta\, z_2 \big\rangle. \tag{12}$$

Here, $\partial\ell/\partial a$ isolates the radial (energy) response, while $\partial\ell/\partial\theta$ measures angular sensitivity (source balance). Importantly, the sampling directions $(\cos\theta, \sin\theta)$ and $(-\sin\theta, \cos\theta)$ are orthonormal in parameter space. In Cartesian coordinates, radial and angular variations are inherently coupled: radial scaling projects across both $\partial\ell/\partial w_1$ and $\partial\ell/\partial w_2$, while pure angular adjustments require maintaining the non-linear constraint $w_1\delta w_1 + w_2 \delta w_2 = 0$.

Sampling distributions and induced density. The primary distinction between parameterizations lies in their natural sampling measures (Table 4). The polar sampler provides decoupled marginals: the angle follows $\theta \sim \mathcal{U}([0, \pi/2])$, ensuring strict angular isotropy across the cone, while the density on the (w_1, w_2) plane scales as $\propto 1/a$.

Conversely, independent uniform sampling $w_i \sim \mathcal{U}([0, S])$ (Cartesian-I) induces an angular density obtained by integrating the area element $a\, da\, d\theta$ over the square boundary:

$$p(\theta) = \int_0^{S/\max(\cos\theta, \sin\theta)} a\, da \ \propto\ \frac{1}{\max(\cos\theta, \sin\theta)^2}. \tag{13}$$

This density introduces a pronounced diagonal bias: it peaks at $\theta = \pi/4$ ($1.57\times$ uniform density) and thins near the coordinate axes ($0.79\times$). Rectifying this distortion in Cartesian space (Cartesian-II) requires importance sampling with the inverse Jacobian $1/\sqrt{w_1^2 + w_2^2}$, emulating at added complexity what the polar parameterization yields natively.

Operational advantages of the polar form. The polar formulation provides three direct design benefits:

- **(M1) Decoupled energy curriculum:** The scale a can be dynamically annealed toward 1 without perturbing the uniform angular coverage $\theta \sim \mathcal{U}([0, \pi/2])$.
- **(M2) Isotropic cone coverage:** Uniform sampling over θ gives every angular direction of the cone equal sampling probability, preventing the decoder from overfitting to fixed mixing trajectories.
- **(M3) Orthogonal diagnostic monitoring:** Quantities Eq. (11) and Eq. (12) decouple energy and balance errors into independent scalar metrics, enabling targeted diagnostic tracking.

Table 4: Sampling distributions over the positive-weight cone $\mathcal{C}(z_1, z_2)$. "Isotropic in θ" denotes a uniform angular marginal over $[0, \pi/2]$ (every ray is sampled equiprobably). Cartesian-I exhibits a non-uniform angular density that concentrates probability mass on the diagonal and thins toward the axes.

Parameterization	**Natural Sampler**	**Induced Density on $\mathcal{C}$**
Polar (Ours)	$a \sim \mathcal{U}([s_{\min}, s_{\max}])$ w.p. p_g, else $a = 1$; $\theta \sim \mathcal{U}([0, \pi/2])$	$p(\theta) \propto 1$ (isotropic in θ)
Cartesian-I	$w_i \sim \mathcal{U}([0, S])$ independently	$p(\theta) \propto \frac{1}{\max(\cos\theta, \sin\theta)^2}$
Cartesian-II	w sampled with Jacobian weight $\frac{1}{\sqrt{w_1^2+w_2^2}}$	Equivalent to Polar (by construction)

Table 5: Cost of one universal mixup term, measured on one NVIDIA L40 with the backbone and loss weights of Appendix B.1 ($B = 64$, 2 s chunks at 16 kHz, AdamW). Top: components isolated on the real code path (forward and forward+backward). Bottom: a complete generator step with exactly one such term added to the reconstruction, adversarial and KL objectives. Percentages are relative to the no-mixup baseline of the same run; each entry is the mean over 15 timed iterations after warm-up, on an otherwise idle GPU. The absolute step delta is stable across repetitions of this measurement to within $\pm 5\%$.

Isolated component (ms)	**forward**	**forward + backward**
Extra decoder pass on z_{mix}	24.7	59.8
Second spectral evaluation (ℓ_{rec} on mix)	3.6	11.1
One $\mathcal{L}_{\text{UMix}}$ (total)	**28.3**	**70.9**
Complete generator step (ms)	**time**	**vs. baseline**
Reconstruction + adversarial + KL	160.9	1.000×
+ one $\mathcal{L}_{\text{UMix}}$	248.3	1.543× (+54.3%)

A.2 Computational Cost of a Single Universal Mixup Loss

A practical consideration for the conical objective is its training cost. Here, we quantify the overhead for the configuration used throughout the paper: a single $\mathcal{L}_{\text{UMix}}$ term Eq. (2) per generator step, evaluated at one random cone point per sample. The forward graph is:

$$z_1 = E_\psi(x_1), \quad z_2 = E_\psi(x_2), \qquad z_{\text{mix}} = a(\cos\theta\, z_1 + \sin\theta\, z_2), \qquad \hat{y} = D_\phi(z_{\text{mix}}), \tag{14}$$

with the mixed waveform target $x_{\text{tgt}} = \text{clip}_1\big(a(\cos\theta\, x_1 + \sin\theta\, x_2)\big)$.

No re-encoding or extra backward pass. Assuming the batch codes $z = E_\psi(x)$ are already computed for the reconstruction loss, evaluating one $\mathcal{L}_{\text{UMix}}$ term requires exactly:

1. **One additional decoder forward pass** on z_{mix}.
2. **One evaluation of** ℓ_{rec} on $(\hat{y},\, x_{\text{tgt}})$.

All other operations (such as batch shifting, sampling (a, θ), and performing affine combinations) incur negligible memory traffic ($O(BPT)$) compared to the decoder's layer-wise cost ($O(BLd^2)$). Notably, the term requires no second encoder pass. Furthermore, because z_{mix} is smooth with respect to z, the mixup gradient backpropagates through the decoder, mixing operation, and encoder within the single `backward()` call already triggered by $\mathcal{L}_{\text{rec}}$ (Appendix A.1). An explicit second backward pass is only necessary for encoder-side mirror objectives (Table 6), which must encode the mixed and endpoint waveforms, roughly doubling the cost:

$$\mathcal{L}_{\text{enc}}(\psi) = \mathbb{E}_{x_1,x_2}\, \mathbb{E}_{a,\theta}\, \big\| E_\psi\big(a(\cos\theta\, x_1 + \sin\theta\, x_2)\big) - a\big(\cos\theta\, E_\psi(x_1) + \sin\theta\, E_\psi(x_2)\big) \big\|^2. \tag{15}$$

The overhead is dominated by the extra decode. At $L = 125$ latent frames, $d = 512$, and six layers, the decoder accounts for about half of a standard autoencoder forward pass, dominated by

Table 6: Generator-step cost of the mixing regularizers on the same backbone, optimizer and batch ($B = 64$, 2 s at 16 kHz, one NVIDIA L40). All single-decode variants add one decoder forward plus one spectral evaluation; only the sampling law of (a, θ) differs. K samples per step (`universal`$\times K$) is what one cone *sample* avoids. The encoder-side cone objective $\mathcal{L}_{\text{enc}}$ is listed for reference: it needs an additional encoder pass over the endpoint and mixed waveforms and costs about twice as much.

Generator step	ms	vs. no-mixup baseline
Reconstruction + adversarial + KL	160.9	1.000×
+ Trig mixup (constant-power arc)	245.4	+52.5%
+ implicit linearity (Torres)	246.9	+53.5%
+ **one universal cone sample (ours)**	**248.3**	**+54.3%**
+ universal × 2 samples	334.5	+107.8%
+ universal × 4 samples	507.1	+215.2%
+ universal × 8 samples	854.4	+431.0%
+ encoder-side cone objective	310.5	+93.0%

$O(L^2 d)$ attention scores and $4Ld^2$ projections. Table 5 isolates these components using the recipe from Appendix B.1. The additional decoder pass constitutes almost the entire overhead, with the second spectral evaluation contributing approximately 13%. (Note that mixed targets depend only on (x_1, x_2, a, θ) and can be precomputed).

Single-sample coverage is the most efficient strategy. Table 6 compares our single universal term against alternative regularizers and multi-sample cone approximations. Two key findings emerge:

- **Matched overhead for latent mixup:** A single $\mathcal{L}_{\text{UMix}}$ costs the same as other single-decode regularizers (+52.5% for Trig mixup, +53.5% for implicit-linearity). The overhead is inherent to latent mixing, not specific to conical sampling, ensuring a fair, wall-clock-matched comparison across models (Appendices C, D.1, A.3).
- **Linear scaling with samples:** Achieving full cone coverage via a single stochastic sample is far cheaper than repeated sampling. Overhead scales linearly (+108% at $K = 2$, +431% at $K = 8$), proving that a single uniform-arc sample per step is the most efficient way to achieve law-invariant interpolation.

Scaling and end-to-end cost. The extra decoder cost, scaling as $O(BL^2 d + BLd^2)$ per layer, is consistent across sequence lengths because the encoder, discriminator, and data pipeline remain fixed. We measure a +53% overhead at 2 s, +54% at 10 s, and +53% at 44.1 kHz ($L = 344$), showing no deterioration in the quadratic attention regime. Practically, for a 300-epoch run (batch 64, 1922 batches/epoch), the single term adds roughly 14 hours total on an NVIDIA L40 (from ≈ 25.8 h to ≈ 39.8 h).

Zero inference cost. Because $\mathcal{L}_{\text{UMix}}$ is strictly a training objective, the final checkpoint retains the exact architecture, parameter count, and per-sample inference cost of the baseline. The resulting algebraic separation (+13.7 dB SI-SDR, Section 4.2) and interpolation robustness (Section 4.3) are achieved with zero inference penalty.

A.3 The Three-Arm Sampler Ablation

The geometric framework of Appendix A.1 establishes two testable hypotheses: (i) under matched cone densities, coordinate parameterization is functionally invariant (Cartesian-II ≡ Polar); (ii) the uncorrected Cartesian-I distribution induces geometric biases detrimental to mixing tasks. We validated both hypotheses by training three autoencoders differing strictly in their loss sampling procedure (identical Transformer backbone, loss weights, optimization recipe, gain anneal schedule, and PRNG seed): `cart_square` (Cartesian-I), `cart_iso` (Cartesian-II, Jacobian-corrected), and the polar reference `univ`. All arms were evaluated on the 600 MUSDB18-HQ pairs under the pipeline of Section 4.3, accompanied by dedicated radial, angular, and compositional probes. To account for

Table 7: Cone-fidelity sweep of the three sampler ablation arms: decoder SI-SDR (dB, ↑) on 600 held-out MUSDB18-HQ stem pairs (2 s clips, 16 kHz, seed 42). Gain sweep: $D_\phi(a\,z)$ vs. $a\,x$ at θ=0 across 400 (track, stem) pairs. Angular sweep: constant-energy mix at a=1. Multi-stem block: zero-shot linear combinations of 3–4 codes (never encountered in training). Bold indicates best performance per row (drop row: smallest fidelity loss). Across all unconstrained multi-source compositions, the polar sampler provides superior generalization ($p \leq 0.03$).

Probe	`cart_square`	`cart_iso`	`univ`
Gain sweep (θ=0, pure gain-scaled reconstruction)			
$a = 0.05$ (near-silence)	6.09	6.28	**10.44**
$a = 0.1$	**13.10**	8.33	12.34
$a = 0.2$	**14.11**	12.75	13.83
$a = 0.35$	14.62	13.83	**14.86**
$a \in [0.5, 3.0]$	14.9–15.3	14.3–14.7	**15.2–15.7**
Angular sweep ($a = 1$)			
$\theta = 0^\circ$ (reconstruction)	17.55	16.85	**17.77**
$\theta = 45^\circ$ (balanced mix)	12.68	12.61	**13.55**
drop θ: $0^\circ \to 45^\circ$	-4.87	-4.24	$\mathbf{-4.22}$
Multi-stem composition (never trained)			
4 stems, quarter mix ($\sum_k w_k$=1)	12.24	12.00	**12.79**
4 stems, full sum (a=2)	12.26	12.13	**12.66**
3 stems, third mix (mean of 4 leave-one-outs)	12.55	12.17	**13.16**

varying reconstruction baselines, we assess the reconstruction-normalized generalization gap (path fidelity minus isolated reconstruction).

(i) Equivalence on the unit shell at matched density. Across the seven standard interpolation laws operating on the unit-gain shell ($a \approx 1$), `cart_iso` is statistically indistinguishable from the polar reference: the normalized gap differs by only -0.24 dB (95% CI $[-0.45, -0.03]$, track-level $p = 0.27$; pair-level $p = 0.024$, $n = 600$). Furthermore, both isotropic formulations trace nearly identical angular degradation profiles: the drop from $\theta = 0^\circ$ to balanced mixing ($\theta = 45^\circ$) is 4.24 dB for `cart_iso` versus 4.22 dB for `univ`. This confirms that along constant-energy mixing trajectories, coordinate choice does not alter representation quality once angular isotropy is preserved.

(ii) Geometric penalties of naive Cartesian sampling. On the normalized gap across all seven interpolation trajectories, Cartesian-I performs consistently worse than Cartesian-II (-0.27 dB, 95% CI $[-0.40, -0.15]$, track-level $p = 9 \times 10^{-3}$). Additionally, `cart_square` incurs a $+0.63$ dB penalty on the constant-power arc despite oversampling the diagonal by $1.57\times$. This occurs because its natural radial density ($\propto a$) depresses coverage of the intermediate unit-gain shell ($a \in [0.707, 1]$) by a factor of $2.3\times$ relative to polar sampling. The raw $+0.36$ dB advantage of Cartesian-I is entirely confined to single-stem reconstruction and vanishes when mixing supervision is engaged.

Radial curriculum dynamics and numerical boundaries. Discrepancies among samplers emerge primarily outside the unit shell, governed by radial curriculum dynamics rather than angular coverage (Table 7):

- **Curriculum exposure vs. retention:** Under the scheduled cosine gain annealing, the sampling radius is restricted toward $(1, 1)$ late in training. Because `cart_square` retains a linear $\propto a$ density profile across $[0, S]$, it maintains incidental exposure to low gains ($a \in [0.1, 0.2]$) longer than `cart_iso`, yielding higher retention at $a = 0.1$ (13.10 dB vs. 8.33 dB, $p = 5 \times 10^{-81}$).
- **Near-silence robustness ($a = 0.05$):** The polar sampler uniquely retains near-silence fidelity (10.44 dB, a lead of more than $+4$ dB over both Cartesian implementations). While the soft-norm bottleneck provides an isotropic denoising basin across all models (Appendix A.5), the polar parameterization samples the radial coordinate directly, avoiding the numerical singularity of the Cartesian Jacobian weight $1/\sqrt{w_1^2 + w_2^2}$ near $a \to 0$.

- **Zero-shot multi-stem composition:** In complex composite mixtures of 3 and 4 sources, `univ` consistently outperforms both Cartesian variants (12.79 dB vs. 12.24 dB and 12.00 dB on 4-stem mixes, all pairwise $p \leq 0.016$).

In summary, while Cartesian-II confirms the theoretical invariance of isotropic mixing at unit gain, native polar sampling provides robust numerical behavior across radial boundaries and multi-source compositions without reliance on corrective Jacobian weighting.

A.4 Probabilistic Interpretation: Soft-Norm as a Constant-Variance VAE

This appendix formalises the probabilistic interpretation introduced in Section 3.4, establishing how the scale-sensitive soft-norm bottleneck can be read as a constant-variance VAE (CV-VAE; Ghosh et al., 2020) at the level of loss terms and deriving its angle-invariant noise floor.

Posterior model and exact KL collapse. We model the encoder as parameterising a conditional Gaussian posterior over latent codes $z \in \mathbb{R}^P$ with deterministic mean $\hat{\mu}(x) \triangleq E_\psi(x)$ and an input-independent isotropic covariance:

$$q(z \mid x) = \mathcal{N}\big(\hat{\mu}(x),\, \sigma_0^2 I_P\big), \qquad z = \hat{\mu}(x) + \sigma_0\, g, \quad g \sim \mathcal{N}(0, I_P), \tag{16}$$

where P is the total latent dimension (channels $\times$ frames), and the perturbation scale is defined as $\sigma_0 \triangleq \alpha\, \bar{\sigma}$, with $\alpha = 0.05$ during training ($\alpha = 10^{-3}$ at inference) scaled by the running code standard deviation $\bar{\sigma}$ tracked via exponential moving average (treated as a detached constant with $\nabla_x \bar{\sigma} = 0$).

Under a standard isotropic prior $p(z) = \mathcal{N}(0, I_P)$, the analytical Kullback–Leibler divergence evaluates to:

$$\begin{aligned} D_{\mathrm{KL}}\big(q(z \mid x) \parallel p(z)\big) &= \frac{1}{2}\Big[\mathrm{Tr}(\sigma_0^2 I_P) + \|\hat{\mu}(x)\|^2 - P - \log\det(\sigma_0^2 I_P)\Big] \\ &= \frac{1}{2}\|\hat{\mu}(x)\|^2 + \underbrace{\frac{P}{2}\big(\sigma_0^2 - 1 - \log\sigma_0^2\big)}_{\text{constant w.r.t. } \hat{\mu}(x)}. \end{aligned} \tag{17}$$

Because the posterior covariance does not depend on the input x, the gradient $\nabla_{\hat{\mu}} D_{\mathrm{KL}}$ reduces strictly to $\hat{\mu}(x)$. Thus, the variational penalty collapses identically to an explicit ℓ_2 weight-decay term on the unquantized latent activations, coinciding with the Regularized Autoencoder framework (RAE; Ghosh et al., 2020). The reconstruction objective is thereby equivalent to a fixed-variance denoising loss (Vincent, 2011):

$$\mathcal{L}_{\mathrm{rec}}^{\mathrm{stoch}} = \mathbb{E}_{g \sim \mathcal{N}(0, I_P)}\big[\ell_{\mathrm{rec}}\big(x,\, D_\phi(\hat{\mu}(x) + \sigma_0 g)\big)\big]. \tag{18}$$

Scope and boundary conditions of the equivalence. We explicitly clarify that this identification holds at the level of objective terms rather than claiming structural equivalence to a fully generative probabilistic graphical model:

1. **Surrogate reconstruction:** The waveform reconstruction loss ℓ_{rec} is a multi-scale STFT spectral surrogate rather than the exact negative log-likelihood under a prescribed decodability prior.
2. **Aggregate axis regularization:** The empirical regularization objective $\mathcal{L}_{\mathrm{reg}}$ enforces normality along aggregated temporal (t) and channel (c) axes:
$$\mathcal{L}_{\mathrm{reg}} = \mathbb{E}_t\big[\mu_t^2 + \sigma_t^2 - \log\sigma_t^2 - 1\big] + 0.4\,\mathbb{E}_c\big[\mu_c^2 + \sigma_c^2 - \log\sigma_c^2 - 1\big], \tag{19}$$
which acts as a composite constraint: the μ^2 terms penalise activation magnitudes in line with RAE-ℓ_2, while the $\sigma^2 - \log\sigma^2$ penalties standardise latent channel statistics in the spirit of β-VAE (Higgins et al., 2017).
3. **Detached tracking:** Noise injection uses tracked scale statistics $\bar{\sigma}$ rather than learned variational variances, and the adversarial loss $\mathcal{L}_{\mathrm{adv}}$ operates outside the ELBO bound.

Subject to these stated assumptions, the reduction of the variational penalty to latent regularization is exact.

Linear compositionality and angle-invariant noise floors. A critical vulnerability of applying mixup to noisy representations is that stochastic perturbations can become anisotropic across mixing ratios. Let two distinct input samples x_1, x_2 produce latent posteriors with independently sampled noise vectors $g_1, g_2 \sim \mathcal{N}(0, I_P)$ ($g_1 \perp g_2$). A mixed latent representation at coordinate (a, θ) on the cone expands as:

$$\begin{aligned} z_{\text{mix}} &= a\cos\theta\, z_1 + a\sin\theta\, z_2 \\ &= a\big(\cos\theta\,\hat{\mu}_1 + \sin\theta\,\hat{\mu}_2\big) + a\,\sigma_0\big(\cos\theta\, g_1 + \sin\theta\, g_2\big). \end{aligned} \tag{20}$$

Because independent Gaussian random vectors are closed under linear combinations, the composite stochastic vector satisfies:

$$a\,\sigma_0\big(\cos\theta\, g_1 + \sin\theta\, g_2\big) \;\sim\; \mathcal{N}\Big(0,\, a^2\sigma_0^2(\cos^2\theta + \sin^2\theta)I_P\Big) \;=\; \mathcal{N}\big(0,\, a^2\sigma_0^2 I_P\big). \tag{21}$$

Taking expectations over the noise distribution yields the total expected second moment:

$$\mathbb{E}_{g_1,g_2}\big[\|z_{\text{mix}}\|^2\big] = a^2\big\|\cos\theta\,\hat{\mu}_1 + \sin\theta\,\hat{\mu}_2\big\|^2 + a^2\sigma_0^2 P. \tag{22}$$

Equation Eq. (22) proves that the stochastic noise floor is strictly isotropic with respect to the mixing angle θ. The decoder is therefore exposed to an identical corruption distribution regardless of whether it reconstructs an isolated stem ($\theta = 0$) or inverts a balanced multi-source combination ($\theta = \pi/4$).

Positioning in the RAE–VAE design space. The full optimization objective is summarized as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}}^{\text{stoch}} + \lambda_{\text{UMix}}\,\mathcal{L}_{\text{UMix}} + \lambda_{\text{reg}}\,\mathcal{L}_{\text{reg}} + \lambda_{\text{adv}}\,\mathcal{L}_{\text{adv}}, \tag{23}$$

where $\mathcal{L}_{\text{UMix}}$ acts as an off-manifold fixed-variance denoising regularizer across the conical support.

Our architecture occupies the intersection of RAE and CV-VAE: whereas standard RAEs and Wasserstein Autoencoders (Tolstikhin et al., 2018) omit stochastic injection in favor of deterministic manifolds, we deliberately maintain isotropic stochastic perturbation. This noise floor establishes a controlled, non-vanishing tolerance basin around the latent manifold, preserving reconstruction stability during out-of-distribution latent subtraction while the polar sampler provides uniform angular coverage of the mixing cone.

A.5 The Role of Noise: Curriculum, Tolerances, and the CV-VAE

To isolate the operational mechanism of the constant-variance bottleneck (Section 3.4), we trained two targeted deviations of the reference architecture under strictly identical conditions (shared transformer backbone, dataset, seed, radial gain schedule, and 300-epoch budget):

- `hetero`: Replaces the constant noise standard deviation σ_0 with an energy-matched, input-dependent law. Each latent code is corrupted by $\sigma_i = \sigma_0(1+\kappa u_i)/\sqrt{1+\kappa^2\text{Var}(u)}$, where $\kappa = 1$ and u_i denotes the empirical batch quantile of the clean code norm $\|z_i\|$. This preserves total average perturbation energy while introducing a path-dependent, $\sim 300\times$ within-pair variance disparity.
- `zero`: Disables additive noise entirely ($\sigma_0 \equiv 0$), reducing the bottleneck to a fully deterministic Regularized Autoencoder (RAE).

Hypothesis 1: θ-invariance mechanism (Falsified). We initially hypothesised that input-dependent noise might allow the decoder to "game" the objective by inferring the mixing angle θ directly from the residual noise magnitude. The `hetero` arm clearly falsifies this conjecture: despite a $\sim 300\times$ within-pair noise variation, the interpolation law-invariance metric $\mathcal{I}_{\text{MS}}$ remains essentially indistinguishable across all three variants (0.00238 for `univ`, 0.00250 for `hetero`, and 0.00249 for `zero`, all within $\sim$5%). The decoder does not exploit the noise amplitude channel. Thus, angular generalization is governed purely by uniform conical sampling $\theta \sim \mathcal{U}([0, \pi/2])$, independent of stochastic injection.

Table 8: Algebraic speech denoising on 10 s LJSpeech mixtures (16 kHz) using known interference latents. Additive noise during training is critical for subtraction algebra: the deterministic `zero` arm suffers relevant degradation, dropping below the unmixed soft-norm baseline despite demonstrating superior isolated reconstruction accuracy.

Arm	Recon Anchor (dB)	Separation (White Noise, dB)	Separation (Bandpass, dB)
`univ (ref)`	$+7.05$	$+8.11$	$\mathbf{+7.80}$
`hetero`	$+7.29$	$\mathbf{+8.22}$	$+7.47$
`zero`	$\mathbf{+7.42}$	$+0.93$	-2.55
`softnorm`	$+7.13$	-3.38	-2.87

Table 9: Gain sweep ($\theta = 0$, SI-SDR dB, 16 kHz MUSDB18-HQ battery) isolating the radial axis. Constant variance governs small-signal decodability: `univ` maintains fidelity at near-silence ($a = 0.05$) and high gains ($a \geq 2.2$), whereas the heteroscedastic and deterministic arms collapse.

Arm	$\mathbf{a = 0.05}$ (near-silence)	$\mathbf{a = 0.1}$	$\mathbf{a = 0.35}$	$\mathbf{a \in [2.2, 3.0]}$
`univ (ref)`	$\mathbf{+10.4}$	$\mathbf{+12.3}$	$\mathbf{+14.9}$	$\mathbf{15.6 \to 15.4}$
`hetero`	-29.6	-22.7	$+12.7$	$15.7 \to 15.4$
`zero`	-49.9	-44.7	-20.8	$13.3 \to \ 7.1$

Hypothesis 2: Off-manifold tolerance basin (Confirmed). The primary necessity of the additive noise floor emerges during out-of-distribution latent arithmetic. When isolating a target stem via algebraic subtraction,

$$z_{\text{target_pred}} = z_{\text{mix}} - z_{\text{interference}}, \tag{24}$$

the resulting representation does not lie strictly on the encoder manifold. Empirical profiling reveals that this subtraction residual exhibits an RMS amplitude $1.13\times$ that of the code itself ($\approx 400\times$ the training perturbation scale σ_0), with approximately half of its energy directed anti-parallel to the target code.

As reported in Table 8, the constant-variance model achieves $+7.80$ dB SI-SDR on bandpass denoising. In contrast, the noise-free `zero` model experiences a notable breakdown to -2.55 dB, despite exhibiting the sharpest isolated reconstruction anchor ($+7.42$ dB). Independent stress tests confirm that the deterministic decoder breaks down under minimal out-of-distribution drift: injected with a 1% relative random perturbation, the reference decodes cleanly at $+7.6$ dB ($+5.0$ dB at 10%), whereas the deterministic arm drops to -2.7 dB at 1%. The continuous perturbation noise expands the decodable basin roughly $400\times$ beyond the nominal noise radius, providing the geometric elasticity required for latent arithmetic.

Hypothesis 3: Constant variance governs radial boundaries and composition. While noise *presence* provides off-manifold tolerance, noise *homoscedasticity* governs radial stability. Along the unit-gain arc, both `hetero` and `univ` successfully preserve the subtraction algebra. However, evaluating purely radial trajectories ($\theta = 0$) exposes significant failures at boundary regimes (Table 9).

The failure modes highlight two distinct mechanisms:

- **Small-signal suppression ($a \leq 0.1$):** Because `hetero` scales noise proportionally with code norm, it strips away the regularization ball precisely as $a \to 0$. When decoding scaled codes $a\,z$, the reference faithfully reproduces amplitude down to $a = 0.05$ (output/target RMS ratio of 1.00). Conversely, the deterministic decoder outputs an unattenuated noise floor, measuring $158\times$ too loud at $a = 0.05$ (heteroscedastic: $5.9\times$).
- **Gauge calibration and multi-stem scaling:** Due to unconstrained bottleneck scaling, the deterministic model converges to an alternate numerical gauge (tracked code scale ~ 112 vs. 0.054 in `univ`; both remain unit-standardized under aggregate KL regularization). Evaluated against relative operating scales, `zero` exhibits the standard trade-off of unregularized autoencoders (Table 10): sharper in-domain reconstruction on MUSDB18-HQ (16.25 dB vs. 15.42 dB for `univ`), but significant failure when composing out-of-

Table 10: Noise-ablation arms on the 16 kHz MUSDB18-HQ battery (SI-SDR dB, ↑; 2 s, seed 42; paired evaluation pipeline). Recon. anchor: decoder's isolated reconstruction averaged across paired stems. Multi-stem composition: zero-shot decoding of $\sum_k w_k z_k$ for 4-stem combinations never encountered in training against the clamped target audio sum (100 tracks; quarter mix: $\sum_k w_k = 1$; full sum: gain $a = 2$). Bold indicates best performance per row.

Metric / Probe	`univ (ref)`	`hetero`	`zero`
Recon. anchor	15.42	15.53	**16.25**
Multi-stem composition (never trained)			
4 stems, quarter mix ($\sum_k w_k$=1)	**12.79**	12.07	12.16
4 stems, full sum (a=2)	**12.66**	12.20	5.07

distribution multi-source combinations (collapsing to 5.07 dB vs. 12.66 dB on 4-stem full-sum mixes, $p = 6 \times 10^{-12}$).

The experimental evidence is consistent with the CV-VAE reading: maintaining a scale-invariant denoising ball across the radial axis safeguards algebraic decodability at boundary regimes where explicit curriculum sampling is sparse or annealed.

A.6 Encoder-Side Cone Alignment and the Distributional Collapse Trade-off

The positive-weight cone framework admits an intuitive encoder-side analogue. Specifically, an explicit encoder homomorphism loss can be formulated to supervise the pre-noise latent codes:

$$\mathcal{L}_{\text{enc}}(\psi) = \mathbb{E}_{x_1,x_2}\mathbb{E}_{a,\theta}\left\| E_\psi\big(a(\cos\theta\, x_1 + \sin\theta\, x_2)\big) - a\big(\cos\theta\, E_\psi(x_1) + \sin\theta\, E_\psi(x_2)\big)\right\|^2, \quad (25)$$

enforcing $E_\psi(w_1x_1 + w_2x_2) \approx w_1E_\psi(x_1) + w_2E_\psi(x_2)$ over clean latent features using the polar cone sampler (Section 3.3). Jointly optimized with the Universal Mixup loss $\mathcal{L}_{\text{UMix}}$, the positive cone would theoretically commute bidirectionally across both E_ψ and D_ϕ.

We also examine a cycle-consistency alternative (`Univ+Cycle`), which enforces algebraic closure indirectly through a latent round trip: the mixed code is decoded by the decoder and re-encoded, and the re-encoding must recover the mixed code,

$$\mathcal{L}_{\text{cycle}}(\psi) = \mathbb{E}_{x_1,x_2,a,\theta}\left\| E_\psi\big(D_\phi(z_\Lambda)\big) - \text{sg}\big[z_\Lambda\big]\right\|^2, \qquad z_\Lambda = a\big(\cos\theta\, E_\psi(x_1) + \sin\theta\, E_\psi(x_2)\big), \quad (26)$$

where $\text{sg}[\cdot]$ denotes the stop-gradient and D_ϕ is held frozen within this term, so gradients flow through the encoder only. The mixture z_Λ uses the same polar cone sampler as $\mathcal{L}_{\text{UMix}}$; unlike $\mathcal{L}_{\text{enc}}$ there is no penalty on the raw code geometry, and hence no direct incentive to collapse the latent onto a low-rank subspace.

The homomorphism versus latent regularity trade-off. Table 11 evaluates this structural intervention across the full diagnostic battery of Section 4.1. Directly penalizing encoder non-linearity succeeds in its immediate objective: `Univ+Enc` lowers encoder homomorphism MSE to 0.039 (a 20-fold reduction relative to decoder-only supervision) and marginally sharpens decoder commutation (0.92×10^{-4} LERP error).

However, this gain comes at the cost of latent space integrity. Across all configurations where the encoder is explicitly regularized for linearity, the latent norm error (`NormErr`, the norm-preservation error of the constant-power mix at α=0.5: $\big|\,\|\text{trigMix}(z_A, z_B)\| - \|z_A\|\,\big|$) surges by an order of magnitude: from 3.40 in the decoder-only reference to 30.31–32.21. The cycle-consistency model (`Univ+Cycle`) achieves near-perfect algebraic linearity (LERP MSE of 0.004), yet collapses the underlying representation geometry (Table 12).

Attempting to reconcile this tension through scheduled annealing (`Univ+Enc+Anneal`, where $\lambda_{\text{enc}} \to 0$ across the final 25% of training alongside doubled KL weighting) successfully recovers the latent-distribution metrics (`NormErr` restored to 3.12), but immediately erases algebraic gains, regressing encoder homomorphism MSE back to 0.486. The trade-off is structural: optimization

Table 11: Encoder alignment ablation on 2 s clips (seed 42; same battery as Table 1, column definitions as Table 14: Enc/Dec MSE are encoder-homomorphism and decoder-commutation errors at α=0.5, Dec MSE in units of 10^{-4}; ΔTrig/ΔLERP = LSD between the decoded and ideal audio mixes at α=0.5; Eff-Rank = participation ratio). `Univ (ref.)` denotes the from-scratch decoder-only reference (`univ_warm`; Appendix B.1). `Univ+Enc` incorporates direct encoder cone supervision Eq. (25); `Univ+Cycle` employs latent round-trip closure Eq. (26); `Univ+Enc+Anneal` anneals $\lambda_{\text{enc}} \to 0$ across the final 25% of epochs while doubling KL weight. `NormErr` = mean absolute norm-preservation error of the constant-power mix at α=0.5. Best per column in **bold**.

	Enc Homomorphism MSE ↓		Dec MSE ($\times 10^{-4}$) ↓					
Configuration	**Trig**	**LERP**	**Trig**	**LERP**	**ΔTrig ↓**	**ΔLERP ↓**	**NormErr ↓**	**Eff-Rank ↑**
`Univ (ref.)`	0.815	0.699	2.18	1.16	3.13	3.15	3.40	78.28
`Univ+Enc`	0.039	0.020	**1.73**	**0.92**	3.05	3.06	30.31	64.21
`Univ+Cycle`	0.148	**0.004**	1.78	0.96	**2.98**	**3.02**	32.21	**98.59**
`Univ+Enc+Anneal`	0.538	0.486	2.39	1.38	3.14	3.16	**3.12**	82.58

Table 12: Latent space health and flow-matching compatibility under encoder alignment (500 clips, 10 s audio duration, 16 kHz, seed 42). **Lat Var**: element-wise latent variance (target $\approx$ 1.0). $|\cos|$: absolute cosine similarity between independent representations (target $\approx$ 0.0 for uncorrelated codes). **NormR**: relative norm ratio $\|\text{trigMix}(z)\|/\|z\|$ evaluated at $\alpha = 0.5$ (ideal = 1.0). Best per column in **bold** (Lat Var and NormR: closest to target).

Configuration	**SI-SDR (dB) ↑**	**MS-STFT ↓**	**Lat Var**	$\lvert\cos\rvert$ **↓**	**NormR → 1**
`Univ (ref.)`	**13.60 ± 9.67**	**0.567 ± 0.58**	**0.971**	**0.190**	**1.068**
`Univ+Cycle`	13.05 ± 8.98	0.661 ± 0.63	0.629	0.721	1.312
`Univ+Enc+Anneal`	13.29 ± 9.66	0.600 ± 0.59	0.969	0.296	1.128

paths that drive encoder MSE below 0.15 routinely trigger `NormErr` > 30, while constraints preserving `NormErr` ≈ 3 regress encoder MSE back to ≈ 0.5.

Mechanism: marginal vs. cross-covariance collapse. The root cause of this instability stems from an objective mismatch. The aggregate regularizer $\mathcal{L}_{\text{reg}}$ (Equation 19) constrains only *marginal* moments across coordinate axes: an encoder satisfies zero mean and unit variance along each slice even if its representations collapse onto a highly correlated, anisotropic manifold.

Minimizing the linear homomorphism objective $\mathcal{L}_{\text{enc}}$ finds an easy degenerate solution by mapping diverse acoustic inputs into low-dimensional, co-linear subspaces. This directly undermines the inductive bias required by generative priors (such as Flow Matching): spectral statistics can even look healthy while the cross-sample geometry collapses. `Univ+Cycle` posts the highest Eff-Rank of all four arms (98.6, Table 11), yet its codes for independent inputs are near-parallel ($|\cos| = 0.721$ vs. 0.190 in the reference) and its marginal variance is depressed (0.63 vs. 0.97; Table 12).

The annealed experiment corroborates this diagnosis: releasing algebraic constraints allows the marginals to recover (**Lat Var** $= 0.969$), but directional damage persists ($|\cos| = 0.296$, $p < 10^{-29}$; **NormR** $= 1.128$, $p < 10^{-28}$). Resolving encoder homomorphism without compromising generative priors appears to require regularization over full cross-feature covariance operators rather than decoupled marginals, which supports our decoder-centric formulation as a principled operational design for continuous audio representations.

A.7 Relation to Implicit Linearity

The implicit-linearity regularizer of Torres et al. (2026) supervises two properties in-batch, split by a stochastic mask. With probability $1 - m_{\text{T}}$, an isolated latent code is gain-scaled (*homogeneity*):

$$z' = a\, z_1, \qquad x_{\text{tgt}} = \text{clip}_1(a\, x_1), \tag{IH}$$

corresponding to $\theta = 0$ along the primary axis of the mixing cone $\mathcal{C}$, forcing the decoder to infer output scale from latent norm alone. With probability m_{T}, two latent codes are summed and subsequently gain-scaled (*additivity*):

$$z' = a\,(z_1 + z_2), \qquad x_{\text{tgt}} = \text{clip}_1\big(a\,(x_1 + x_2)\big). \tag{IA}$$

In conical coordinates, this sum expands as $a(z_1 + z_2) = \tilde{a}(\cos\tilde{\theta}\, z_1 + \sin\tilde{\theta}\, z_2)$ with a fixed balance angle $\tilde{\theta} = \pi/4$ and gain $\tilde{a} = a\sqrt{2}$. Consequently, Eq. (IH)–Eq. (IA) supervise exactly *two isolated rays* within $\mathcal{C}$: $\theta \in \{0, \pi/4\}$. Every continuous mixture with intermediate balance $\theta \notin \{0, \pi/4\}$ is strictly out-of-distribution for an implicit-linearity decoder.

Universal Mixup as a strict continuous generalization. Comparing the general conical objective Eq. (2) with the implicit formulations Eq. (IH)–Eq. (IA):

$$\mathcal{L}_{\text{Torres}} = \mathcal{L}_{\text{UMix}}\Big|_{\theta\in\{0,\pi/4\}\text{ discrete}}, \qquad \mathcal{L}_{\text{UMix}} = \mathcal{L}_{\text{Torres}}\Big|_{\theta\sim\mathcal{U}(0,\pi/2)}. \tag{27}$$

UMix substitutes the degenerate two-point distribution on θ with continuous uniform measure over the entire quadrant.

The optimization implications for the backward pass are structural. Backpropagating the loss ℓ to the latent code branches via $g = \nabla_{z_{\text{mix}}}\ell$ (Appendix A.1), the additive branch Eq. (IA) distributes an identical scalar gain a symmetrically across both branches:

$$\nabla_{z_1}\ell\Big|_{\text{add}} = \nabla_{z_2}\ell\Big|_{\text{add}} = a\, g, \tag{4A}$$

yielding a single shared degree of freedom locked at a $1:1$ ratio with zero variance across source balances. In contrast, the continuous cone sampler provides distinct branch gains:

$$\nabla_{z_1}\ell\Big|_{\text{gen}} = a\cos\theta\, g, \qquad \nabla_{z_2}\ell\Big|_{\text{gen}} = a\sin\theta\, g. \tag{4G}$$

While the expected gradient under $\theta \sim \mathcal{U}(0, \pi/2)$ remains isotropic, $\mathbb{E}_\theta[\nabla_{z_1}\ell] = \mathbb{E}_\theta[\nabla_{z_2}\ell] = \mathbb{E}[a]\,\frac{2}{\pi}\,g$, each stochastic gradient step injects non-zero balance variance. Equation Eq. (4A) emerges as the degenerate zero-variance limit of Eq. (4G). UMix therefore exposes the decoder to directional balance gradients entirely absent from implicit linearity.

Empirical evaluation: generalization off the annealed shell. To disentangle the contributions of angular coverage and radial scaling, we evaluate three matched 16 kHz models on held-out interpolation laws across 600 MUSDB18-HQ stem pairs (Table 13):

1. **Implicit (Torres):** Supervised strictly on two rays ($\theta \in \{0, \pi/4\}$) under the cosine gain curriculum $(0, 3) \to (1, 1)$.
2. **Trig mixup:** Supervised over the full angular arc $\theta \in [0, \pi/2]$, but rigidly restricted to the unit-energy circle ($a \equiv 1$, no radial curriculum).
3. **Universal (Ours):** Supervised over the complete (a, θ) conical support under the gain schedule.

Diagnosis of failure modes. The experimental hierarchy directly mirrors the geometric completeness of the training objectives:

- **Unit shell behavior ($a \in [0.71, 1]$):** All three decoders perform robustly on standard convex interpolations (*lerp* and *trig*), demonstrating that discrete ray supervision is sufficient to interpolate along surface boundaries within 2.1 dB of reconstruction anchors. Universal mixup remains superior even on the dedicated training trajectory of the arc model (*trig*: $+0.65$ dB over Trig mixup, $p = 1.6 \times 10^{-8}$; $+0.81$ dB over Implicit, $p = 6.6 \times 10^{-7}$).
- **Cone interior breakdown ($a < 0.71$):** When evaluating attenuated non-convex transitions, performance diverges sharply. The implicit model suffers failure (-13.63 dB on *half-trig* and -5.75 dB on *dip*, with MS-STFT errors rising up to three orders of magnitude above its reconstruction baseline). Trig mixup degrades gracefully but significantly (9.64 dB on *half-trig*; 0.00 dB on *dip*), lagging behind Universal by 4.3 dB to 14.0 dB ($p \leq 1.1 \times 10^{-31}$). In stark contrast, Universal preserves high fidelity across all fades, remaining within 1.5 dB of its isolated reconstruction anchor everywhere.
- **Scale mis-inference mechanism:** Profiling decoded output energy indicates that interior collapse is driven by broken small-signal gain calibration. Along *half-trig*, the implicit

Table 13: Generalization of the annealed shell across held-out interpolation laws on 600 MUSDB18-HQ stem pairs (9 α evaluation points, seed 42; 16 kHz evaluation corpus held out from training). Standard laws (*lerp*, *trig*) inhabit the unit shell $a = \|(w_1, w_2)\| \in [0.71, 1]$. Non-convex interior fades test $a < 0.71$, a domain abandoned by late gain annealing: *half-trig* ($a = 0.5$ const.), *quad* (a dips to 0.354), and *dip* (a dips to 0.129). Values report mean $\pm$ SEM over pairs (SI-SDR in dB, $\uparrow$; MS-STFT magnitude MSE, $\downarrow$). Law-invariance score $\mathcal{I}_{\text{SI-SDR}}$ tracks standard deviation across path means.

	Universal (Ours)		**Trig Mixup**		**Implicit (Torres)**	
Interpolation Law	**SI-SDR**	**MS-STFT**	**SI-SDR**	**MS-STFT**	**SI-SDR**	**MS-STFT**
lerp	**14.34** $\pm$ 0.38	**0.0119**	13.47 $\pm$ 0.36	0.0129	13.12 $\pm$ 0.38	0.0237
trig (a=1)	**14.27** $\pm$ 0.37	**0.0168**	13.62 $\pm$ 0.36	0.0180	13.46 $\pm$ 0.38	0.0179
half-trig (a=0.5)	**14.03** $\pm$ 0.40	**0.0043**	9.64 $\pm$ 0.30	0.0083	-13.63 $\pm$ 0.48	27.5
quad (min a=0.35)	**14.46** $\pm$ 0.40	**0.0079**	10.19 $\pm$ 0.36	0.0147	-4.85 $\pm$ 0.36	23.0
dip (min a=0.13)	**13.96** $\pm$ 0.38	**0.0053**	0.00 $\pm$ 0.33	0.0711	-5.75 $\pm$ 0.36	11.2
Law Invariance $\mathcal{I}_{\text{SI-SDR}}$ ($\downarrow$)	**0.212**	**0.0051**	5.553	0.0260	12.193	12.71
Overall Path Metric	**14.21**	**0.0092**	9.39	0.0250	0.47	12.34
Recon. Anchor	**15.41**	**0.0164**	14.56	0.0175	15.15	0.0167

decoder emits waveforms $15\times$ to $45\times$ too energetic relative to ground truth (surging to $\sim 700\times$ on *dip*), treating attenuated combinations as full-scale codes. Trig mixup exhibits intermediate gain overestimation ($1.4\times$ to $8\times$), whereas Universal faithfully tracks reference RMS within 9% across all evaluation points.

The law-invariance score ranks the paradigms strictly according to their geometric support: $\mathcal{I}_{\text{SI-SDR}} = 0.212$ dB (full cone) $\ll 5.553$ dB (unit arc) $\ll 12.193$ dB (two rays). Dense continuous supervision over (a, θ) is therefore what prevents decoders from hallucinating unit-scale energy in attenuated interior regimes.

B Global Experimental pipeline

B.1 Architecture and Training Details

Backbone architecture. All 16 kHz models are based on a symmetric 6-layer encoder and 6-layer decoder Transformer (Vaswani et al., 2017) with model dimension 512, 8 attention heads, and GLU feed-forward networks (Shazeer, 2020). Input audio is mapped into latent frames via non-overlapping 256-sample 1D convolutional patch embeddings with Rotary Position Embeddings (RoPE) (Su et al., 2024). Activations use LeakyReLU, with two periodic Snake layers (Ziyin et al., 2020) placed at the decoder output. The continuous bottleneck maintains $C = 64$ channels. A 2 s audio segment at 16 kHz (32,000 samples) produces a latent representation of shape $(125, 64)$, corresponding to a $256\times$ temporal downsampling (62.5 Hz frame rate) and an overall $4\times$ scalar dimensionality reduction. The 44.1 kHz model (Sections 4.2–4.3) shares the identical backbone configuration. To ensure linear scale sensitivity, final LayerNorm operations (Ba et al., 2016) at the encoder bottleneck and decoder input are disabled across all mixing-supervised variants.

Soft-norm bottleneck and stochastic floor. The `SoftNormBottleneck` applies a learnable per-channel affine transformation, normalization by an EMA-tracked code standard deviation, additive Gaussian noise ($\sigma_0 = 0.05$ during training, $\sigma_0 = 10^{-3}$ at inference), and the dual-axis moment penalty (Equation 19). Stochastic perturbations are sampled independently at the bottleneck output for each input code with a single shared variance per batch, ensuring an angle-invariant composite noise distribution under conical mixing (Equation 21). For the reference model, this yields an effective per-element noise standard deviation of approximately 3×10^{-3} (the tracked code scale is $\hat{\sigma} \approx 0.054$, so $\sigma_0 \hat{\sigma} \approx 2.7 \times 10^{-3}$).

Optimization and loss objectives. Reconstruction loss ℓ_{rec} is the multi-resolution spectral loss of Parker et al. (2026) across 7 FFT scales ($N_{\text{fft}} \in [32, 2048]$ with hop size $N_{\text{fft}}/4$), combining

spectral contrast (Frobenius-normalized linear magnitude) with adaptive log-magnitude terms. The generator optimization objective follows Equation Eq. (23) with $\lambda_{\text{adv}} = 0.05$ using a PatchGAN discriminator operating on the magnitude STFT (single scale, N_{fft}=1024, hop 256), $\lambda_{\text{KL}} = 1.0$, and $\lambda_{\text{UMix}} = 1.0$ under default cone sampling hyperparameters ($s_{\min} = 0$, $s_{\max} = 3$, $p_g = 0.8$, $\delta = 0.05$, mixing ratio $m = 0.5$, and cosine gain annealing toward $(1, 1)$).

Baseline configurations and checkpoint pipelines. We evaluate three baseline configurations under identical budgets and seeds:

- **Minimal:** Disables soft-norm scaling, noise injection, and mixing supervision.
- **Soft-Norm:** Enables soft-norm normalization and dual-axis KL without mixing supervision.
- **Universal (`univ`):** Trained from scratch under the full conical objective.

The 16 kHz models are trained from random initialisation for 300 epochs. The 44.1 kHz variant is fine-tuned directly from the converged 16 kHz reference (identical weights, verified at checkpoint level) on the same corpus resampled to 44.1 kHz (batch size 64).

Training environment. Models are optimized via AdamW (Loshchilov & Hutter, 2019) (lr $= 10^{-4}$, weight decay 10^{-4}) using a plateau learning-rate scheduler with patience of 15 epochs, batch size 64, and 2 s audio slices. All autoencoder and flow-matching runs share PRNG seed 42, with two deliberate exceptions: the flow-model seed replication and autoencoder-seed robustness studies (Appendix F.3), which train three seeds (42/43/44) to quantify run-to-run variance. Training corpora comprise speech from VCTK (Veaux et al., 2017), acoustic events from FSD50K (Fonseca et al., 2022), and music from FMA (Defferrard et al., 2017), with waveforms peak-normalized to 0.95.

B.2 Evaluation Metrics Definitions

Waveform and spectral fidelity. Reconstruction accuracy and path fidelity are quantified via two complementary criteria:

- **SI-SDR:** Scale-Invariant Signal-to-Distortion Ratio (Le Roux et al., 2019), evaluated on raw time-domain waveforms.
- **MS-STFT:** The magnitude component of the multi-resolution training objective (Parker et al., 2026) (spectral-contrast and adaptive log-magnitude terms, phase terms disabled), evaluated across seven resolutions ($N_{\text{fft}} \in \{32, 64, 128, 256, 512, 1024, 2048\}$, hop size $N_{\text{fft}}/4$, Hann window):

$$\begin{aligned} \text{MS-STFT}(y, \hat{y}) &= \sum_{i=1}^{7} \big(\text{SC}_i + \text{LM}_i\big), \\ \text{SC}_i &= \frac{\big\| |S_i(y)| - |S_i(\hat{y})| \big\|_F}{\big\| |S_i(y)| + |S_i(\hat{y})| \big\|_F + \epsilon}, \\ \text{LM}_i &= \frac{1}{T_i F_i} \sum_{t,f} \Big| \log\Big(\tfrac{|S_i(y)|_{t,f}}{\sigma_i} + 1\Big) - \log\Big(\tfrac{|S_i(\hat{y})|_{t,f}}{\sigma_i} + 1\Big) \Big|. \end{aligned} \tag{28}$$

  where $|S_i(\cdot)|$ is the linear-frequency magnitude spectrogram and $\sigma_i = \big(\text{Var}\,|S_i(y)| + \text{Var}\,|S_i(\hat{y})|\big)^{1/2}$ is the per-clip, per-resolution adaptive scale. Because this metric coincides with the spectral loss component, it is paired with SI-SDR as independent reconstruction benchmarks that enter no training loss.

Homomorphism and commutation metrics. Let E_ψ denote the encoder, D_ϕ the decoder, $z_i = E_\psi(x_i)$, $\alpha \in [0, 1]$, $\bar{\alpha} = 1 - \alpha$, and $\mathcal{A} = \{0.25, 0.5, 1.0, 1.5, 2.0, 3.0\}$. We define the convex linear interpolations in waveform and latent domains as $x_\alpha = \bar{\alpha} x_A + \alpha x_B$ and $z_\alpha = \bar{\alpha} z_A + \alpha z_B$.

The encoder homomorphism error metrics quantify compliance with linear algebra in latent space:

$$\mathcal{E}^{\text{enc}}_{\text{single}} = \mathbb{E}_{x,a\in\mathcal{A}}\left[\text{MSE}\big(E_\psi(a\,x),\ a\,E_\psi(x)\big)\right], \tag{29}$$

$$\mathcal{E}^{\text{enc}}_{\text{lerp}}(\alpha) = \mathbb{E}_{x_A,x_B}\left[\text{MSE}\big(E_\psi(x_\alpha),\ \bar\alpha\,E_\psi(x_A)+\alpha\,E_\psi(x_B)\big)\right], \tag{30}$$

$$\mathcal{S}^{\text{enc}}_{\text{lerp}} = \mathbb{E}_{x_A,x_B}\left[\text{Var}_{\alpha\in\Omega}\,\text{MSE}\big(E_\psi(x_\alpha),\ \bar\alpha\,E_\psi(x_A)+\alpha\,E_\psi(x_B)\big)\right], \tag{31}$$

where $\mathcal{E}^{\text{enc}}_{\text{lerp}}(0.5)$ evaluates midpoint additivity, $\mathcal{E}^{\text{enc}}_{\text{single}}$ measures scale homogeneity across the discrete grid $\mathcal{A}$, and $\mathcal{S}^{\text{enc}}_{\text{lerp}}$ tracks path variance along the 11-point uniform evaluation grid $\Omega = \{0.0, 0.1, \ldots, 1.0\}$.

The decoder commutation metrics mirror this formulation directly in waveform space:

$$\mathcal{E}^{\text{dec}}_{\text{single}} = \mathbb{E}_{z,a\in\mathcal{A}}\left[\text{MSE}\big(D_\phi(a\,z),\ a\,D_\phi(z)\big)\right], \tag{32}$$

$$\mathcal{E}^{\text{dec}}_{\text{lerp}}(\alpha) = \mathbb{E}_{z_A,z_B}\left[\text{MSE}\big(D_\phi(z_\alpha),\ \bar\alpha\,D_\phi(z_A)+\alpha\,D_\phi(z_B)\big)\right], \tag{33}$$

$$\mathcal{S}^{\text{dec}}_{\text{lerp}} = \mathbb{E}_{z_A,z_B}\left[\text{Var}_{\alpha\in\Omega}\,\text{MSE}\big(D_\phi(z_\alpha),\ \bar\alpha\,D_\phi(z_A)+\alpha\,D_\phi(z_B)\big)\right]. \tag{34}$$

An ideal algebraic autoencoder satisfies $\mathcal{E}^{\text{dec}}_{\text{lerp}} = 0$ and $\mathcal{E}^{\text{enc}}_{\text{lerp}} = 0$ everywhere on the cone. In reporting tables, ΔLERP and ΔTrig denote the log-spectral distance (mean absolute log-magnitude difference in dB, computed on peak-normalized waveforms) between the decoded mix and the ideal audio mix at the path midpoint α=0.5 (end-to-end mix fidelity; lower is better).

Latent geometry and manifold integrity. Structural diagnostics on the latent manifold are computed over an empirical pool of evaluated representations $\mathcal{Z} = \{z_i\}_{i=1}^N \subset \mathbb{R}^P$:

$$\mathcal{N}\ (\text{NormErr}) = \mathbb{E}_{z_A,z_B}\left[\,\big|\,\|c\,z_A + s\,z_B\| - \|z_A\|\,\big|\,\right], \tag{35}$$

$$\mathcal{C}\ (\text{Coverage}) = \mathbb{E}_{z_A,z_B}\left[\min_{z\in\mathcal{Z}\setminus\{z_A,z_B\}} \|\,(c\,z_A + s\,z_B) - z\,\|\right], \tag{36}$$

$$\mathcal{R}\ (\text{Eff-Rank}) = \frac{\left(\sum_i \sigma_i^2\right)^2}{\sum_i \sigma_i^4}, \tag{37}$$

where $c = s = \cos(\pi/4) = 1/\sqrt{2}$ specifies the constant-power arc midpoint. Here:

- $\mathcal{N}$ (**NormErr**) measures the absolute failure of constant-power norm preservation at the cone midpoint: the deviation of $\|\text{trigMix}(z_A, z_B)\|$ from the reference code norm $\|z_A\|$, expressed in raw latent-norm units (lower is better).
- $\mathcal{C}$ (**Coverage**) measures off-manifold distance by computing the Euclidean nearest-neighbour distance between the mixed latent vector and the reference code pool $\mathcal{Z}$.
- $\mathcal{R}$ (**Eff-Rank**) calculates the continuous participation-ratio effective rank (Roy & Vetterli, 2007) of the centred latent covariance matrix with singular values σ_i.

Statistical estimation and uncertainty. Encoder, decoder, and `NormErr` metrics report per-pair means $\pm$ standard deviation over the balanced 1,000-pair evaluation battery. Coverage ($\mathcal{C}$) and Effective Rank ($\mathcal{R}$) are aggregate pool statistics estimated via bootstrapping over independent clips ($B = 500$ iterations over 256 clips). Spectral reconstruction scores (MS-STFT) report empirical means $\pm$ standard error of the mean (SEM) over 2,000 held-out segments.

B.3 Statistical Significance and Robustness pipeline

Experiment 1 (algebraic commutation and geometry). All comparative evaluations are conducted under paired conditions across identical validation sets ($n = 1{,}000$ paired clips). For primary decoder commutation metrics, Table 15 reports paired mean differences, Cohen's effect sizes $d_z = \bar{d}/s_d$, two-sided paired t-test p-values, and minimum sample sizes required for 80% statistical

power ($\alpha = 0.05$). The universal reference separates from both baseline paradigms by multiple standard errors across all decoder commutation criteria ($p < 10^{-70}$; effect sizes $|d_z| \in [0.6, 5.3]$; all auxiliary commutation metrics yield $p \leq 1.1 \times 10^{-5}$).

The solitary exception where a baseline outperforms the reference is the latent norm-preservation metric (`NormErr`), where the Soft-Norm model achieves a marginally lower deviation ($\Delta = -0.72$, $p = 4.8 \times 10^{-10}$). As established in Section 4.1, this difference directly exemplifies the trade-off of regularization without mixing supervision: preserving isolated spherical priors at the expense of a relevant failure in linear decoder commutation.

Experiment 2 (algebraic source separation). On the zero-shot algebraic separation benchmark over 100 MUSDB18-HQ tracks, the performance advantage of $\mathcal{L}_{\text{UMix}}$ ($+13.7$ dB SI-SDR vs. < -10 dB for the external baselines) remains significant under non-parametric paired Wilcoxon signed-rank tests over the 383 common valid (track, stem) trials ($p < 10^{-63}$ against every baseline; the 17 excluded trials have silent ground-truth stems, for which SI-SDR is undefined), confirming that algebraic stem retrieval is not an artifact of outlier tracks.

Experiment 3 (interpolation laws and boundary stability). Across every evaluated continuous mixing law (Table 23), the performance margin between the universal model and the top-performing baseline exceeds the pooled standard error of the mean (SEM) by factors ranging from $19.9\times$ to $40.8\times$. This corresponds to a standardized separation $Z > 19.9$ ($p < 10^{-87}$ under normal asymptotic approximations), indicating that cross-model rankings are robust to sampling variance. The only pairwise statistically indistinguishable comparison occurs between Lin-CAE and DAC on the randomized path under the MS-STFT metric (0.1035 vs. 0.1032; paired Wilcoxon $p = 0.62$); neither baseline achieves usable fidelity on this trajectory.

Experiment 4 (generative modeling with Flow Matching). All primary generative trends replicate consistently across three independent DiT training seeds with paired clip-level evaluations ($n = 128$ clips per seed, 384 overall):

- **Latent arithmetic decodability (*latavg*):** Decoding averaged generated latents yields superior MS-STFT for the universal model across 128/128 clips in every independent seed (per-seed p-values in Table 32; 384/384 overall, cumulative exact sign test $p < 10^{-115}$).
- **Conditional velocity regression:** Flow field velocity prediction errors favor our representation on 126/2, 124/4, and 126/2 clips per seed respectively (per-seed paired tests $p \leq 1.1 \times 10^{-44}$; pooled across seeds, the exact sign test gives $p < 10^{-99}$).
- **Trajectory straightness:** Velocity consistency tests confirm straighter generative trajectories under our geometry (Welch's $t = 9.24$, $p = 1.2 \times 10^{-15}$).

Observed effect sizes surpass seed variance by one to two orders of magnitude: an arithmetic validity margin of $+9.41 \pm 0.06$ dB; a single-step Euler integration lead of $+8.42$ dB (per-arm seed spreads ≤ 0.07 dB); and a velocity regression advantage of $+3.29 \pm 0.12$ dB.

Conversely, we report the null result: when averaging directly in waveform space (*wavavg*), the Soft-Norm flow baseline exhibits a slight edge over Universal ($+0.21 \pm 0.09$ dB SI-SDR). However, per-seed significance is inconsistent ($p = 0.21$, $p = 0.012$, and $p = 1.9 \times 10^{-5}$ across the three runs), and MS-STFT differences remain insignificant across roughly one-third of evaluated clips. We therefore read raw waveform-domain generation quality at full budget as practically equivalent, while the advantage in downstream latent algebraic manipulation is large and consistent (Appendix F.3).

Hierarchical clustering and track correlation. Because stem pairs extracted from the same audio track share underlying musical characteristics (key, tempo, production aesthetics), naive pair-level evaluations risk variance underestimation due to intra-track correlation. To prevent anti-conservative significance claims, the MUSDB18-HQ pair batteries (Experiment 3 and the sampler ablation in Appendix A.3) evaluate clustered track-level degrees of freedom ($n = 100$ independent tracks) alongside pair-level statistics, and all headline claims retain high significance ($p < 0.01$) under these clustered evaluations. Experiment 4 evaluates held-out clips of the VCTK/FSD50K/FMA

Table 14: Comprehensive additivity and latent geometry ablation across pipeline-matched 16 kHz models (2 s audio slices, PRNG seed 42, evaluated over the balanced 1,000-pair battery). **Encoder MSE**: homomorphism error in latent space; **Decoder Commutation**: error in units of 10^{-4}, accompanied by midpoint ($\alpha = 0.5$) mix fidelity deltas (ΔTrig / ΔLERP, in log-spectral distance); **Latent Geometry & Fidelity**: `NormErr` is the mean absolute norm-preservation error of the constant-power mix at α=0.5 (code-norm units, Section B.2), `Eff-Rank` evaluates the participation ratio, and **MS-STFT** is the magnitude component of the training objective (Section B.2). Bold denotes best performance per column.

	Encoder MSE ↓		Decoder Commutation ↓				Latent Geometry & Fidelity		
Configuration	**Trig**	**LERP**	**Trig** ($\times 10^{-4}$)	**LERP** ($\times 10^{-4}$)	**ΔTrig**	**ΔLERP**	**NormErr ↓**	**Eff-Rank ↑**	**MS-STFT ↓**
Minimal	14.153	2.211	56.41	16.56	5.22	4.58	202.59	9.03	**0.534**
Soft-Norm	0.893	0.798	368.45	308.99	6.82	7.28	**2.68**	**107.20**	0.617
Trig Mixup	**0.561**	**0.516**	2.19	1.28	3.25	3.28	3.08	92.70	0.644
Implicit (Torres)	0.587	0.533	2.19	1.34	**3.07**	3.25	3.12	89.43	0.598
Universal (Ours)	0.815	0.699	**2.18**	**1.16**	3.13	**3.15**	3.40	78.28	0.601

Table 15: Paired hypothesis tests of midpoint decoder commutation $\mathcal{E}^{\text{dec}}_{\text{lerp}}(0.5)$ against `Universal (Ours)` across 1,000 paired clips (Δ in units of 10^{-4}; $d_z = \bar{d}/s_d$ represents paired Cohen's effect size; n_{80} denotes sample size required to reach 80% statistical power at $\alpha = 0.05$).

Configuration	$\mathbf{\Delta}$	$\mathbf{d_z}$	**p-value**	$\mathbf{n_{80}}$
Minimal	$+15.169$	-0.612	3.5×10^{-71}	21
Soft-Norm	$+307.420$	-5.257	$< 10^{-100}$	1

validation split rather than MUSDB stems: its comparisons are clip-level paired tests over 128 clips per seed, with three-seed replication as the robustness axis (Appendix F.3).

C Extended Results for Experiment 1 (Additivity)

C.1 Full Additivity and Geometry Ablation

Table 14 extends the core additivity evaluation of Section 4.1 to include the constant-power arc baseline (*Trig mixup*) and the implicit-linearity formulation of Torres et al. (2026).

All mixing-supervised autoencoders suppress decoder commutation error down to $\sim 2 \times 10^{-4}$, a $15\times$ to $250\times$ reduction relative to the configurations without mixing supervision, indicating that cone-directed objectives instill operational additivity in the decoder. While the implicit-linearity and constant-power arc models perform comparably on this in-domain unit-gain test (as both supervise fixed trajectories within the cone), both fail to match the full conical model once evaluated on non-convex or attenuated paths (Appendix A.7, Table 13). Notably, our reference universal model preserves high effective rank (78.28) and controlled latent norm preservation (`NormErr` 3.40), striking the necessary balance between algebraic commutation and generative manifold regularity.

C.2 Why the Final Bottleneck LayerNorm Is Removed

Our architecture disables the final LayerNorm modules at the encoder output and decoder input (Appendix B.1) to preserve latent code magnitude. This choice underwrites the constant-variance VAE reading (Section 3.4) and is inherited by all downstream models. Here, we evaluate this design through an architecture-paired ablation: both arms share the 50.7 M-parameter backbone, training data, conical objective, and seed 42, differing strictly by the presence of the two LayerNorm operations (2,048 parameters, $+0.004\%$).

Trade-off analysis. As reported in Table 16, retaining bottleneck LayerNorm presents a clear structural trade-off. On the encoder side, the normalized arm achieves significantly lower interpolation error ($0.699 \rightarrow 0.051$) and smoother latent paths ($p < 10^{-100}$). Normalization forces the encoder to become approximately gain-invariant ($\cos(E(ax), E(x)) = 0.912$ vs. 0.852; Table 17), which brings the encoder map closer to a formal linear homomorphism.

Table 16: Bottleneck LayerNorm ablation on the Exp. 1 additivity suite (2 s clips, 16 kHz, seed 42, identical batteries as Table 1). Both models share the 50.7 M backbone and conical objective. *Scale-sensitive* disables both bottleneck LayerNorms; *+ final LayerNorm* restores them. Δ is ablation minus reference (positive is worse, except for `Eff-Rank`). p-values are two-sided paired t-tests or paired bootstraps over shared evaluation items. ‡Raw `Conv` gain is a shrinkage artifact; normalized by the intrinsic pool scale, the midpoint sits farther from the pool (Table 17). Best entry per row in **bold**.

Metric	Scale-sensitive (reference)	+ final LayerNorm	Δ	p	d_z
Reconstruction					
MS-STFT ↓	0.601 ± 0.014	0.640 ± 0.015	$+0.039$	$< 10^{-100}$	$+0.54$
Encoder					
single ↓	1.375 ± 0.228	1.254 ± 0.171	-0.122	7.8×10^{-67}	-0.59
lerp ↓	0.699 ± 0.123	0.051 ± 0.084	-0.648	$< 10^{-100}$	-4.43
smooth ↓	$1.3\times10^{-1} \pm 9.1\times10^{-2}$	$4.9\times10^{-3} \pm 2.3\times10^{-2}$	-1.2×10^{-1}	$< 10^{-100}$	-1.39
Decoder ($\times10^{-4}$)					
single ↓	28.19 ± 70.42	29.98 ± 74.40	$+1.79$	1.7×10^{-10}	$+0.20$
lerp ↓	1.11 ± 2.01	1.16 ± 2.30	$+0.05$	3.4×10^{-2}	$+0.07$
smooth ↓	$8.1\times10^{-5} \pm 6.0\times10^{-4}$	$1.0\times10^{-4} \pm 7.9\times10^{-4}$	$+2.2\times10^{-5}$	1.6×10^{-2}	$+0.08$
Latent Space					
NormErr ↓	3.40 ± 3.14	30.71 ± 5.94	$+27.30$	$< 10^{-100}$	$+3.94$
Conv‡ ↓	105.93 ± 0.46	46.25 ± 0.33	-59.68	$< 2\times10^{-3}$	–
Eff-Rank ↑	78.28 ± 3.45	31.71 ± 4.98	-46.56	$< 2\times10^{-3}$	–

Table 17: Latent-geometry and scale-response diagnostics for Table 16. LayerNorm forces magnitude information into code directions, causing code collinearity, variance collapse, and mixture norm inflation.

Diagnostic	Scale-sensitive	+ final LayerNorm	Target
Decoder gain response (log–log slope)	0.954 ± 0.032	0.940 ± 0.042	ideal 1.000
Encoder gain sensitivity $\cos(E(ax), E(x))$	0.852	0.912	ideal < 1 (magnitude-carrying)
Intrinsic pool scale (NN distance)	106.5	33.6	unit of `Conv`
Relative midpoint coverage	0.995 ± 0.230	1.376 ± 0.369	ideal → low
Per-channel code variance	0.985	0.203	ideal ≈ 1.0
Pair-wise code $\lvert\cos\rvert$	0.072	0.800	ideal ≈ 0
Constant-power norm ratio	1.007 ± 0.048	1.339 ± 0.071	ideal ≈ 1.0

However, this gain comes at severe cost to all downstream objectives. Decoder commutation degrades across all criteria ($p \leq 0.034$), acoustic reconstruction deteriorates (MS-STFT $0.601 \rightarrow 0.640$, $+6.5\%$), and the constant-power norm error (`NormErr`) expands ninefold ($3.40 \rightarrow 30.71$, $d_z = +3.94$). Concurrently, the latent space loses over half its effective dimensionality (`Eff-Rank` drops from 78.28 to 31.71). The apparent reduction in raw nearest-neighbor distance (`Conv`, $105.93 \rightarrow 46.25$) is an artifact of manifold shrinkage: the normalized pool contracts by $3.2\times$, and relative to its own intrinsic scale, midpoints sit farther from the data manifold ($0.995 \rightarrow 1.376$).

Mechanism: distributional collapse. Table 17 reveals the underlying mechanism. The scale-sensitive model preserves unit per-channel variance (0.985), orthogonal codes ($\lvert\cos\rvert = 0.072$),

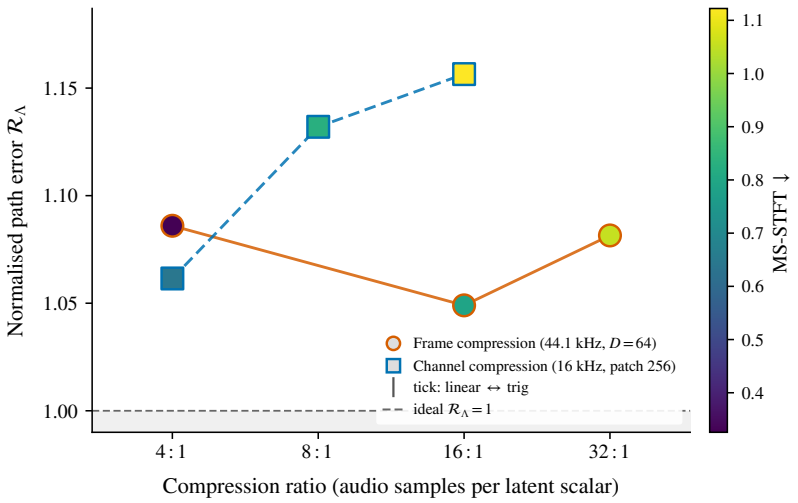


Figure 5: Pareto curve between compression ratio (frame and channel), normalized path error and reconstruction quality.

and faithful constant-power norms (1.007 ± 0.048). Conversely, the LayerNorm bottleneck suffers directional collapse: channel variance falls to 0.203, codes become highly collinear ($|\cos| = 0.800$), and mixture norms inflate by 34% (1.339 ± 0.071).

Because the decoder must still infer loudness under normalized inputs, it encodes scale into activation directions rather than magnitudes. This directional competition distorts the latent geometry, violating the near-orthogonality required by the angle-invariant noise floor (Section 3.4) and replicating the exact collapse pattern seen in encoder-regularized variants (Appendix A.6).

Justification of the design choice. While omitting the final LayerNorm slightly degrades encoder interpolation metrics, it decisively protects decoder commutation, reconstruction fidelity, and manifold dispersion. Because downstream source separation, algebraic editing, and rectified-flow generative priors rely strictly on decodability and an uncollapsed latent space, resolving this trade-off in favor of a scale-sensitive bottleneck is structurally necessary.

C.3 Bottleneck Compression and Scaling Ablation

The algebraic additivity established in Section 4.1 was validated at a nominal $4\times$ scalar compression factor. A fundamental question is whether linear commutation survives under *aggressive compression regimes*, that is, whether additivity is an intrinsic invariant induced by $\mathcal{L}_{\text{UMix}}$ or merely a consequence of over-parameterized latent capacity.

To isolate this behavior, we evaluate two independent compression axes while strictly holding all other architectural hyperparameters invariant (symmetric 6+6 Transformer layers, width 512, AdamW optimizer, 300-epoch budget, 2 s audio slices, PRNG seed 42):

- **Temporal compression sweep** (44.1 kHz, fixed channel dimension $D = 64$): Patch sizes scale across $\{256, 1024, 2048\}$ audio samples. With non-overlapping patch embeddings, the latent frame rate decreases as $172.3 \rightarrow 43.1 \rightarrow 21.5$ Hz (corresponding to frame durations of 5.8, 23.2, and 46.4 ms), increasing scalar compression from $4.0{:}1 \rightarrow 16.0{:}1 \rightarrow 32.0{:}1$. For a 2 s audio segment, this achieves up to an $8\times$ reduction in sequence length ($344 \rightarrow 86 \rightarrow 43$ latent tokens).
- **Channel compression sweep** (16 kHz, fixed temporal patch size of 256 samples): Bottleneck channel dimension narrows across $D \in \{64, 32, 16\}$ at a constant frame rate of 62.5 Hz (125 frames per 2 s clip), scaling scalar compression from $4.0{:}1 \rightarrow 8.0{:}1 \rightarrow 16.0{:}1$.

Table 18: Compression-factor ablation across temporal and channel axes (1,000 paired audio clips, seed 42). **Reconstruction**: single-source test fidelity (MS-STFT and SI-SDR in dB). **Ratio** $\mathcal{R}_\Lambda$: normalized path error (Equation 38) along linear and constant-power (trigonometric) trajectories; values report mean ± SEM over pairs (↓, target → 1.00). Bold indicates optimal algebraic preservation per compression axis.

				Single-Source Recon		Normalized Path Ratio $\mathcal{R}_\Lambda$ ↓	
Configuration	**Patch**	**Frame Rate**	**Compression**	**MS-STFT ↓**	**SI-SDR (dB) ↑**	**Linear**	**Trig**
Temporal compression sweep (44.1 kHz audio, fixed $D = 64$ channels)							
Reference (4.0:1)	256	172.3 Hz	4.0:1	0.3257	23.98	1.083 ± 0.006	1.089 ± 0.007
Compressed (16.0:1)	1024	43.1 Hz	16.0:1	0.7916	9.91	**1.048** ± 0.005	**1.050** ± 0.005
Aggressive (32.0:1)	2048	21.5 Hz	32.0:1	1.0519	6.28	1.080 ± 0.005	1.083 ± 0.006
Channel compression sweep (16 kHz audio, fixed patch size 256 samples)							
Reference (64 channels)	256	62.5 Hz	4.0:1	0.6425	14.36	**1.060** ± 0.005	**1.063** ± 0.006
Compressed (32 channels)	256	62.5 Hz	8.0:1	0.8295	9.38	1.130 ± 0.006	1.134 ± 0.007
Aggressive (16 channels)	256	62.5 Hz	16.0:1	1.1222	6.16	1.152 ± 0.007	1.161 ± 0.007

Invariant diagnostic metric: Normalized Path Error ($\mathcal{R}_\Lambda$). Directly evaluating raw interpolation distortion across compression levels conflates two distinct phenomena: degradation in baseline reconstruction fidelity and degradation in linear algebraic commutation. To eliminate this confound, we evaluate the *normalized path error ratio* $\mathcal{R}_\Lambda$, computed pairwise for each evaluation pair (x_A, x_B):

$$\mathcal{R}_\Lambda = \frac{\frac{1}{|\Omega|}\sum_{\alpha\in\Omega} d\big(D_\phi(z_\Lambda(\alpha)),\, x_\Lambda(\alpha)\big)}{\frac{1}{2}\Big[d\big(D_\phi(E_\psi(x_A)),\, x_A\big) + d\big(D_\phi(E_\psi(x_B)),\, x_B\big)\Big]}, \tag{38}$$

where $d(\cdot,\cdot)$ denotes the 7-resolution MS-STFT magnitude loss, $z_\Lambda(\alpha)$ and $x_\Lambda(\alpha)$ are the latent and waveform mixing trajectories for law Λ, and $\Omega = \{0.0, 0.125, \ldots, 1.0\}$ represents the 9-point evaluation grid.

Because both numerator and denominator are evaluated on the *same model* over the *same audio pairs*, $\mathcal{R}_\Lambda$ isolates algebraic penalty relative to self-reconstruction: an ideal value of $\mathcal{R}_\Lambda \approx 1.0$ indicates that decoding a latent mixture incurs zero excess penalty over reconstructing isolated single sources.

Algebraic robustness across extreme compression. As reported in Table 18 and visualized in Figure 5, across an $8\times$ dynamic range in temporal downsampling and a $4\times$ reduction in channel capacity, the normalized path error remains tightly bounded: $\mathcal{R}_\Lambda \in [1.048, 1.161]$ across all models and both evaluated mixing laws (linear, constant-power). While increasing compression inevitably degrades absolute reconstruction (MS-STFT rises from 0.33 to 1.05 temporally and 0.64 to 1.12 along channels; SI-SDR drops commensurately from ~ 24 dB to 6 dB), the *excess cost* of decoding algebraic mixtures never exceeds 17% of self-reconstruction on either law. This indicates that algebraic decodability is a property induced by the conical mixup objective itself, rather than an artifact of generous bottleneck capacity.

Asymmetric inductive biases across temporal and channel axes. The empirical findings reveal a distinct structural asymmetry between the two compression modalities:

- **Temporal compression preserves algebraic fidelity:** Lengthening patch windows up to 1024 samples (16.0:1 scalar compression) not only maintains additivity but marginally improves the relative ratio ($\mathcal{R}_\Lambda$ drops to 1.048–1.050). Even at aggressive downsampling (patch 2048, emitting only 21.5 frames/s), the ratio remains low (1.080). Because temporal compression reduces the sequence length that the decoder transformer must invert while preserving full per-frame latent capacity ($D = 64$), conical supervision is concentrated across fewer spatial tokens without restricting geometric expressivity.
- **Channel compression restricts conical expressivity:** Conversely, restricting latent channel capacity from $64 \to 16$ causes a monotonic deterioration in algebraic preservation ($\mathcal{R}_\Lambda$ rises from $1.060 \to 1.152$). Narrowing bottleneck width directly constrains the linear degrees of freedom available to span multi-source conical subspaces, making continuous algebraic superposition harder to satisfy simultaneously across diverse acoustic sources.

Table 19: Zero-shot, supervised algebraic source separation on held-out MUSDB18-HQ clips (44.1 kHz; $n = 383$ valid trials), including the *Trig mixup* control: complete per-stem breakdown. Values report mean ± SEM over valid trials. Best results in **bold**, second best underlined.

	Overall ($n = 383$)		**Drums** ($n = 98$)		**Bass** ($n = 94$)	
Model	**SI-SDR (dB)** ↑	**MS-STFT** ↓	**SI-SDR (dB)** ↑	**MS-STFT** ↓	**SI-SDR (dB)** ↑	**MS-STFT** ↓
Universal (Ours)	$\mathbf{+13.71 \pm 0.73}$	$\mathbf{0.0128 \pm 0.0008}$	$\mathbf{+10.04 \pm 1.29}$	$\mathbf{0.0181 \pm 0.0020}$	$\mathbf{+19.01 \pm 1.38}$	$\mathbf{0.0100 \pm 0.0017}$
Trig mixup	$\underline{+12.31 \pm 0.78}$	$\underline{0.3197 \pm 0.0929}$	$\underline{+8.57 \pm 1.52}$	$\underline{0.4173 \pm 0.2193}$	$\underline{+18.00 \pm 1.45}$	$\underline{0.1053 \pm 0.0510}$
SAME-L	-10.65 ± 0.64	2.0746 ± 0.0612	-10.59 ± 1.39	1.7937 ± 0.0807	-5.53 ± 1.06	1.5149 ± 0.0731
Lin-CAE	-14.88 ± 1.04	0.8962 ± 0.0732	-17.27 ± 2.28	1.1389 ± 0.2044	-22.26 ± 2.38	1.4228 ± 0.1754
DAC	-16.62 ± 0.48	82.7153 ± 0.4656	-19.73 ± 0.78	81.1339 ± 0.9915	-15.88 ± 0.76	79.6617 ± 0.8676

	Other ($n = 100$)		**Vocals** ($n = 91$)	
Model	**SI-SDR (dB)** ↑	**MS-STFT** ↓	**SI-SDR (dB)** ↑	**MS-STFT** ↓
Universal (Ours)	$\mathbf{+16.25 \pm 0.65}$	$\mathbf{0.0111 \pm 0.0012}$	$\mathbf{+9.42 \pm 2.08}$	$\mathbf{0.0118 \pm 0.0013}$
Trig mixup	$\underline{+13.77 \pm 1.06}$	$\underline{0.3193 \pm 0.1418}$	$\underline{+8.85 \pm 1.98}$	$\underline{0.4366 \pm 0.2655}$
SAME-L	-12.01 ± 0.66	2.7596 ± 0.1609	-14.50 ± 1.68	2.2026 ± 0.1092
Lin-CAE	-8.21 ± 1.26	0.5497 ± 0.0578	-12.02 ± 2.03	0.4718 ± 0.0541
DAC	-12.69 ± 0.33	86.4470 ± 0.8674	-18.34 ± 1.54	83.4721 ± 0.8395

For practical codec deployment and downstream generative modeling (e.g., Diffusion or Flow Matching Transformers), this suggests an actionable design principle: compression budgets should be allocated preferentially to temporal downsampling, maximizing sequence length reduction for generative efficiency while preserving the full channel dimensionality required for algebraic manipulation.

D Extended Results for Experiment 2 (Source Separation)

D.1 MUSDB18-HQ Setup, Full Stem Breakdowns, and Reconstruction Anchors

Setup and trial pipeline. Evaluations are conducted on 2 s monophonic center crops sampled at 44.1 kHz across 100 tracks from the MUSDB18-HQ corpus (partitioned into 70 train and 30 test tracks via PRNG seed 42). Because none of the evaluated autoencoders encountered MUSDB audio during training, this benchmark evaluates strictly zero-shot out-of-domain transfer.

Every (track, stem) pair yields an isolated reconstruction trial $D_\phi(E_\psi(x))$ and an algebraic separation trial evaluating target stem isolation via latent subtraction:

$$z_{\text{target}} = z_{\text{mix}} - \sum_{k \neq \text{target}} z_k. \tag{39}$$

Trials featuring silent or zero-energy ground-truth stems make SI-SDR mathematically undefined and are omitted, leaving $n = 383$ valid trials out of 400 total combinations (98 Drums, 94 Bass, 100 Other, and 91 Vocals). Evaluated external baselines comprise SAME-L (Parker et al., 2026), the pretrained linear CAE of Torres et al. (2026) (Lin-CAE, trained with the implicit linearity loss), and DAC (Kumar et al., 2023), each decoding identical chunks through its native pipeline.

Reconstruction anchors. Table 20 reports each model's single-audio coding of the isolated stems on the identical trial set. The universal model achieves the sharpest isolated reconstruction across every stem category ($+23.86$ dB SI-SDR overall, MS-STFT 0.0039), indicating that its algebraic separation superiority is not an artifact of an under-regularized autoencoder baseline.

The constant-power-arc model control. Supervising exclusively on the unit-norm arc ($\theta \sim \mathcal{U}(0, \pi/2)$ at $a \equiv 1$, *Trig mixup*) supports latent subtraction to a meaningful degree ($+12.31 \pm 0.78$ dB SI-SDR, Table 19). However, extending supervision across the continuous cone provides an extra $+1.4$ dB of SI-SDR while driving spectral magnitude error down by a factor of $25\times$ (0.0128 vs. 0.3197). This directly measures the contribution of radial coverage: subtraction residuals deviate radially from the unit shell, requiring decodability across the conical interior.

Baseline failure mechanisms. The collapse of the baselines stems from distinct architectural bottlenecks. SAME-L reconstructs isolated stems accurately (Table 20) but breaks the linear vector

Table 20: Single-audio reconstruction anchors on the identical MUSDB18-HQ stems and trial set ($n = 383$ valid trials): each model reconstructs the isolated stem $D_\phi(E_\psi(x))$. Values report mean $\pm$ SEM over pairs. Best in **bold**, second best underlined.

	Overall ($n = 383$)		**Drums** ($n = 98$)		**Bass** ($n = 94$)	
Model	**SI-SDR (dB)** ↑	**MS-STFT** ↓	**SI-SDR (dB)** ↑	**MS-STFT** ↓	**SI-SDR (dB)** ↑	**MS-STFT** ↓
Universal (Ours)	$\mathbf{23.86 \pm 0.88}$	$\mathbf{0.0039 \pm 0.0004}$	$\mathbf{13.35 \pm 1.31}$	$\mathbf{0.0094 \pm 0.0012}$	$\mathbf{40.34 \pm 1.42}$	$\mathbf{0.0000 \pm 0.0000}$
Trig mixup	$\underline{22.51 \pm 0.85}$	$\underline{0.0048 \pm 0.0005}$	$\underline{13.17 \pm 1.27}$	$\underline{0.0108 \pm 0.0015}$	$\underline{36.77 \pm 1.48}$	$\underline{0.0001 \pm 0.0000}$
SAME-L	14.02 ± 0.70	0.0215 ± 0.0017	9.84 ± 1.13	0.0218 ± 0.0025	26.64 ± 1.13	0.0011 ± 0.0002
Lin-CAE	1.00 ± 0.54	0.1408 ± 0.0077	0.65 ± 1.17	0.1242 ± 0.0108	7.58 ± 0.83	0.0797 ± 0.0096
DAC	9.71 ± 0.52	0.0351 ± 0.0020	7.32 ± 1.14	0.0349 ± 0.0031	16.63 ± 0.74	0.0121 ± 0.0012

	Other ($n = 100$)		**Vocals** ($n = 91$)	
Model	**SI-SDR (dB)** ↑	**MS-STFT** ↓	**SI-SDR (dB)** ↑	**MS-STFT** ↓
Universal (Ours)	$\mathbf{26.78 \pm 0.77}$	$\mathbf{0.0014 \pm 0.0002}$	$\mathbf{14.95 \pm 1.90}$	$\mathbf{0.0046 \pm 0.0006}$
Trig mixup	$\underline{25.95 \pm 0.75}$	$\underline{0.0016 \pm 0.0003}$	$\underline{14.07 \pm 1.89}$	$\underline{0.0065 \pm 0.0012}$
SAME-L	11.53 ± 0.57	0.0365 ± 0.0045	8.21 ± 1.74	0.0258 ± 0.0032
Lin-CAE	-0.82 ± 0.48	0.1845 ± 0.0169	-3.41 ± 1.35	0.1740 ± 0.0198
DAC	8.57 ± 0.40	0.0507 ± 0.0048	6.39 ± 1.35	0.0419 ± 0.0044

identity $z_{\text{mix}} = \sum_k z_k$ due to bottleneck LayerNorm. DAC fails completely across all stems (MS-STFT ≈ 80), consistent with residual vector quantization codebooks being ill-suited to representing continuous subtraction vectors. Lin-CAE, the pretrained implicit-linearity model of Torres et al. (2026), exhibits pronounced metric disagreement (achieving the lowest MS-STFT among external baselines, yet trailing SAME-L in SI-SDR), consistent with its stochastic one-step decode introducing waveform phase cancellations that spectral magnitude MSE ignores but time-domain projections penalize. In contrast, the agreement between both metrics for the Universal model is consistent with genuine waveform-level recovery.

D.2 Extended Generalization on Zero-Shot Speech Denoising (LJSpeech)

To further stress-test the algebraic latent space, we extended the zero-shot separation study along three simultaneous extrapolation axes: sequence duration (scaling from the 2 s training length to 10 s evaluation chunks), acoustic domain (evaluating on single-speaker speech from LJSpeech, completely held out from training), and interference family (generalizing from natural music/environmental stems to synthetic FFT-shaped stationary noise).

Experimental setup and noise conditions. We evaluated 100 trials of 10 s audio chunks at 44.1 kHz across ten distinct acoustic conditions (five noise spectra at two SNRs). Stationary interference was synthesized via deterministic FFT filtering under five profiles: full-band white noise, low-pass ($\leq$ 300 Hz), band-pass matching the telephony speech band (300–3400 Hz), high-pass (≥ 3.4 kHz), and pink ($1/f$) noise. These were superposed as natural linear sums at 0 dB and $+6$ dB SNR. Target speech was retrieved via supervised algebraic subtraction: $\hat{x} = D_\phi(z_{\text{mix}} - z_{\text{noise}})$.

Results and cross-corpus robustness. As shown in Table 21, the Universal model preserves algebraic validity across $5\times$ the training duration on the unseen speech corpus, achieving an overall SI-SDR of $+16.19$ dB. This exceeds the best external baseline (SAME-L) by nearly 13 dB while reducing spectral MS-STFT error by a factor of $24\times$.

Table 21: Zero-shot algebraic speech denoising on unseen 10 s LJSpeech chunks at 44.1 kHz (pooled across 10 conditions $\times$ 100 trials). The Universal model generalizes across sequence length and corpus, while baselines degrade severely.

Model	**SI-SDR (dB)** ↑	**MS-STFT** ↓	**Recon Anchor (dB)** ↑
Universal (Ours)	$\mathbf{+16.19 \pm 0.10}$	$\mathbf{0.038}$	$\mathbf{+20.43}$
SAME-L	$+3.41 \pm 0.05$	0.930	$+10.25$
Lin-CAE	$+0.93 \pm 0.04$	1.730	$+1.26$
DAC	-3.02 ± 0.15	10.400	$+8.64$

The model's single-source reconstruction anchor on LJSpeech remains high ($+20.43$ dB), indicating that denoising performance is supported by high autoencoding fidelity.

Table 22 provides the per-condition breakdown. Universal outperforms all baselines in every condition×SNR configuration on both metrics: its most challenging condition ($+10.66$ dB on high-pass noise at 0 dB SNR) still surpasses the highest score attained by any baseline across the entire benchmark (SAME-L at $+4.78$ dB). The baseline degradation patterns mirror those on MUSDB: DAC collapses under vector quantization lattice mismatch (-11.28 dB on low-pass noise), Lin-CAE remains trapped near $+1$ dB, consistent with stochastic decode phase error, and SAME-L achieves partial separation only on narrow out-of-band interference.

Table 22: Per-condition algebraic speech denoising on 10 s LJSpeech chunks (SI-SDR in dB vs. clean speech; 100 trials per cell under deterministic noise seeds; SEM ≤ 0.15 dB across all cells). Low-pass (*lp300*), high-pass (*hp3400*), and pink columns are pooled across 0 dB and $+6$ dB SNR; white and speech-band columns report individual SNRs. **no-op**: zero-subtraction baseline evaluating undecoded mixture $D_\phi(z_{\text{mix}})$ as an unseparated control. Best in **bold**, second best underlined.

Model	lp300	hp3400	pink	white (0 / +6 dB)	speech-band (0 / +6 dB)	no-op
Universal (Ours)	$\mathbf{+20.31}$	$\mathbf{+12.50}$	$\mathbf{+17.40}$	$\mathbf{+14.45}$ / $\mathbf{+17.13}$	$\mathbf{+13.80}$ / $\mathbf{+16.16}$	$+3.30$
SAME-L	$\underline{+2.08}$	$\underline{+4.28}$	$\underline{+4.20}$	$\underline{+4.02}$ / $\underline{+4.78}$	$\underline{+1.26}$ / $\underline{+2.95}$	$+1.90$
Lin-CAE	$+0.33$	$+1.50$	$+0.97$	$+0.31$ / $+0.83$	$+1.05$ / $+1.50$	-2.93
DAC	-11.28	-4.98	$+1.21$	$+1.18$ / $+0.17$	-0.75 / -0.66	$+1.55$

E Extended Results for Experiment 3 (Interpolation Laws)

E.1 Law Definitions and Geometric Characterization

To assess whether the algebraic decodability of $\mathcal{L}_{\text{UMix}}$ generalizes beyond the specific parameterizations encountered during training, we evaluate continuous interpolation across seven distinct mixing laws Λ. All laws satisfy non-negative branch coefficients $w_1(\alpha), w_2(\alpha) \geq 0$ for all $\alpha \in [0, 1]$, so every evaluated path lies within the positive-weight cone $\mathcal{C}(z_A, z_B)$.

The laws defined in Table 23 span diverse curvature regimes:

- **In-distribution reference paths:** *lerp* provides linear interpolation, while *trig* coincides with the unit-gain arc of the polar sampler $\mathcal{S}_{\text{polar}}$.
- **Perceptual fades:** *logfade* is the equal-power fade standard of audio workstations; *sqrt* is its constant-energy counterpart.
- **Asymmetric fades:** *parabolic* traverses the mixing angles with a quadratic skew, while *exp* concentrates its transition near the path endpoints.
- **Stochastic trajectory probe:** the *random* law draws, per stem pair, a fade $w_1(\alpha) = m(1-\alpha)^p + (1-m)\alpha^p$ with $m \sim \mathcal{U}(0, 1)$ and $p \sim \mathcal{U}(0.3, 3)$ from a shared deterministic PRNG stream, so all models decode the same held-out paths.

E.2 Full Continuous Law Battery Results

Table 24 presents the full diagnostic battery evaluating path-mean audio fidelity $D_\phi(z_\Lambda(\alpha))$ against ground-truth audio mixtures $x_\Lambda(\alpha)$ across 600 MUSDB18-HQ stem pairs at 44.1 kHz (evaluated over 9 uniform points $\alpha \in \{0.0, 0.125, \ldots, 1.0\}$).

Empirical analysis across paradigms. The comparison shows a clear separation between the model families:

- **Universal conical invariance:** the Universal model performs near-constantly across all seven trajectories (21.47–22.46 dB SI-SDR), with an overall path mean (22.03 dB) within 1.7 dB of its isolated reconstruction anchor (23.75 dB). This small spread across non-linear and stochastic fades is consistent with continuous conical supervision inducing law-invariant decoding rather than memorization of specific mixup profiles.

Table 23: Continuous interpolation laws evaluated across 600 MUSDB18-HQ stem pairs. Normalization factors $Z_{\log}(\alpha)$ and $Z_{\exp}(\alpha)$ ensure unit endpoint scaling. While *lerp* and *trig* test in-distribution boundary consistency, the remaining five laws represent completely held-out geometric trajectories.

Law	$\mathbf{w_1}(\alpha)$	$\mathbf{w_2}(\alpha)$	Geometric Character
lerp	$1-\alpha$	α	Linear convex combination
trig	$\cos(\pi\alpha/2)$	$\sin(\pi\alpha/2)$	Constant-power circular arc
logfade	$-\log_{10}(1-\alpha+\epsilon)/Z_{\log}$	$-\log_{10}(\alpha+\epsilon)/Z_{\log}$	Equal-power acoustic standard
sqrt	$\sqrt{1-\alpha}$	$\sqrt{\alpha}$	Constant-energy trajectory
parabolic	$1-\alpha^2$	α^2	Asymmetric quadratic fade
exp	$e^{-3\alpha}/Z_{\exp}$	$e^{-3(1-\alpha)}/Z_{\exp}$	Endpoint-concentrated transition
random	$m(1-\alpha)^p+(1-m)\alpha^p$	$1-w_1(\alpha)$	Stochastic fade ($m\sim\mathcal{U}(0,1)$, $p\sim\mathcal{U}(0.3,3)$)

Table 24: Held-out interpolation law battery on 600 MUSDB18-HQ stem pairs (44.1 kHz, seed 42): path-mean fidelity of decoded latent mixtures $D_\phi(z_\Lambda(\alpha))$ against target mixtures $x_\Lambda(\alpha)$. Values report mean ± SEM across 600 pairs (SI-SDR in dB, ↑; MS-STFT magnitude MSE, ↓). Best results in **bold**, second best underlined. The Universal model is the best of the compared models on every law and both metrics.

	Universal (Ours)		**SAME-L**	
Interpolation Law	**SI-SDR (dB) ↑**	**MS-STFT ↓**	**SI-SDR (dB) ↑**	**MS-STFT ↓**
lerp	$\mathbf{22.12\pm0.43}$	$\mathbf{0.0027\pm0.0002}$	9.93 ± 0.31	0.0342 ± 0.0011
trig	$\mathbf{21.98\pm0.44}$	$\mathbf{0.0038\pm0.0002}$	8.33 ± 0.30	0.0871 ± 0.0021
logfade	$\mathbf{22.15\pm0.44}$	$\mathbf{0.0028\pm0.0002}$	10.18 ± 0.32	0.0308 ± 0.0010
sqrt	$\mathbf{21.71\pm0.44}$	$\mathbf{0.0039\pm0.0002}$	7.61 ± 0.31	0.1063 ± 0.0025
parabolic	$\mathbf{22.46\pm0.47}$	$\mathbf{0.0032\pm0.0002}$	10.80 ± 0.34	0.0297 ± 0.0010
exp	$\mathbf{22.30\pm0.44}$	$\mathbf{0.0027\pm0.0002}$	9.81 ± 0.31	0.0313 ± 0.0010
random	$\mathbf{21.47\pm0.43}$	$\mathbf{0.0021\pm0.0001}$	8.43 ± 0.30	0.0405 ± 0.0012

	Lin-CAE		**DAC**	
Interpolation Law	**SI-SDR (dB) ↑**	**MS-STFT ↓**	**SI-SDR (dB) ↑**	**MS-STFT ↓**
lerp	0.37 ± 0.30	0.1109 ± 0.0036	4.24 ± 0.29	0.0829 ± 0.0023
trig	0.27 ± 0.29	0.1651 ± 0.0054	3.57 ± 0.28	0.2413 ± 0.0062
logfade	0.23 ± 0.30	0.1149 ± 0.0038	4.71 ± 0.34	0.0709 ± 0.0020
sqrt	0.07 ± 0.29	0.1712 ± 0.0055	2.98 ± 0.26	0.2956 ± 0.0078
parabolic	0.82 ± 0.31	0.1105 ± 0.0036	5.45 ± 0.34	0.0697 ± 0.0019
exp	0.47 ± 0.29	0.1118 ± 0.0037	4.13 ± 0.35	0.0729 ± 0.0020
random	-0.16 ± 0.30	0.1035 ± 0.0036	2.14 ± 0.33	0.1032 ± 0.0031

- **Curvature sensitivity in SAME-L:** SAME-L performs adequately on the linear path (9.93 dB on *lerp*) but drops on curved transitions (7.61 dB on *sqrt* and 8.33 dB on *trig*), with MS-STFT error on *sqrt* about 3× its *lerp* value.
- **Discrete and stochastic representations:** DAC remains at 2.14–5.45 dB SI-SDR, consistent with residual vector quantization representing continuous fractional mixtures poorly. Lin-CAE stays near 0 dB across all trajectories, consistent with the phase-incoherent single-step decode that magnitude MSE tolerates but SI-SDR penalises.

Degenerate vs. functional law invariance. Table 25 summarizes aggregate path invariance. The low $\mathcal{I}_{\text{SI-SDR}} = 0.311$ dB posted by Lin-CAE is a degenerate case: the model performs poorly on all transitions (overall 0.30 dB), so its cross-path variance is small by construction. The Universal model, by contrast, combines low dispersion ($\mathcal{I}_{\text{MS-STFT}} = 0.0006$, the lowest of all models; $\mathcal{I}_{\text{SI-SDR}} = 0.340$ dB, below every model except the degenerate Lin-CAE) with the highest absolute fidelity of the group (22.03 dB SI-SDR): law-invariance backed by usable decoding, rather than uniform failure.

Table 25: Law-invariance scores and reconstruction anchors (600 MUSDB18-HQ pairs, 44.1 kHz). $\mathcal{I}$: path-invariance metric computed as the sample standard deviation of path means across the seven evaluated laws (↓, Section 4.3). Overall: grand mean over all interpolation trajectories. Recon Anchor: isolated stem reconstruction. Best results in **bold**.

	Law-Invariance $\mathcal{I}$ ↓		Overall Path Metric		Recon Anchor	
Model	**SI-SDR**	**MS-STFT**	**SI-SDR ↑**	**MS-STFT ↓**	**SI-SDR ↑**	**MS-STFT ↓**
Universal (Ours)	**0.340**	**0.0006**	**22.03**	**0.0030**	**23.75**	**0.0037**
SAME-L	1.172	0.0316	9.30	0.0514	13.92	0.0206
Lin-CAE†	0.311	0.0285	0.30	0.1268	0.82	0.1353
DAC	1.103	0.0940	3.89	0.1338	9.60	0.0336

†Lin-CAE exhibits an apparently low $\mathcal{I}_{\text{SI-SDR}}$ variance score (0.311) due to uniformly poor performance near ≈ 0 dB across all laws, rather than genuine algebraic invariance.

F Extended Results for Experiment 4 (Rectified Flow)

F.1 Flow Matching Setup and Latent Space Standardization

Latent space parity and energy calibration. Both generative models operate over 16 kHz continuous latent representations produced by our 6+6 layer transformer autoencoders (256-sample patch windows, $C = 64$ bottleneck channels). A 2 s audio segment yields a latent tensor of shape $(125, 64)$ (125 tokens of 64 channels), while a 10 s sequence expands to $(625, 64)$. The experimental comparison isolates the geometric inductive bias by benchmarking our conical reference (`univ`) against the unmixed soft-norm baseline (`softnorm`) under strictly matched capacities and budgets.

Importantly, both latent manifolds exhibit empirical unit variance across feature dimensions. The mean per-element regression target energy matches within 0.4% on 2 s sequences (0.9962 for `univ` vs. 1.0000 for `softnorm`) and within 0.04% on 10 s sequences (1.9557 vs. 1.9565). This close energy parity ensures that raw velocity regression errors and Rectified Flow objectives are directly commensurable across manifolds without requiring artificial post-hoc rescaling.

Diffusion Transformer (DiT) architecture and optimization. The generative backbone processes representations as temporal sequences of 64-dimensional feature vectors. An input linear projection concatenates the time-perturbed target z_t with the conditional mixture representation z_{mix}, projecting the combined features to hidden dimension $d = 384$. The network body comprises 6 DiT blocks (16.49 M parameters) utilizing unmasked multi-head self-attention (8 attention heads, MLP expansion ratio of 4) and adaptive layer normalization (adaLN) driven by continuous time conditioning.

Attention is full (unmasked) over the time axis and no positional embeddings are used: temporal structure is conveyed only through the linear input projection, a pilot-scale limitation held identical for both models. Optimization is conducted via AdamW (Loshchilov & Hutter, 2019) (initial learning rate 10^{-4}, weight decay 10^{-4}, gradient clipping at 1.0) under a OneCycle learning-rate schedule (Smith & Topin, 2018) across 50 epochs with batch size 32.

The conditional velocity field f_θ is trained via the standard Rectified Flow objective (Liu et al., 2023b; Albergo et al., 2025):

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],\,\epsilon\sim\mathcal{N}(0,I)}\Big\| f_\theta\big(z_t, t, z_{\text{mix}}\big) - \big(\epsilon - z_{\text{data}}\big)\Big\|^2, \qquad z_t = (1-t)\, z_{\text{data}} + t\,\epsilon, \quad (40)$$

where data is mapped at $t = 0$ and prior Gaussian noise at $t = 1$, with generative inference performed via forward Euler integration from $t = 1 \to 0$. Batches are paired via minibatch optimal transport (OT) coupling (Pooladian et al., 2023; Tong et al., 2024) using 20 Sinkhorn iterations. Deterministic regression baselines share the identical DiT backbone (16.5 M parameters), directly optimizing $\|f_\theta(z_{\text{mix}}) - z_{\text{voice}}\|^2$ with the temporal conditioning index fixed at $t = 0$.

Corpus, mixture dynamics, and evaluation splits. Training leverages 50,000 synthetic audio mixtures pairing speech stems from VCTK (Veaux et al., 2017) with acoustic interference drawn from FSD50K (Fonseca et al., 2022) and music from FMA (Defferrard et al., 2017). Sources are

Table 26: Euler step discretization sweep across sequence durations ($n = 128$ paired held-out clips, 6-draw waveform averaging). 2 s results report mean ± std across 3 independent DiT training seeds (p-values pooled over 384 clip-seed evaluations). 10 s results evaluate continuous full-length sequences. Bold indicates best performance per step count.

	2 s Audio Slices (3 Seeds)			10 s Continuous Audio (1 Run)		
Steps N	`univ (Ours)`	`softnorm`	**Δ**	`univ (Ours)`	`softnorm`	**Δ**
1	$\mathbf{9.07 \pm 0.04}$	0.65 ± 0.06	`univ` **+8.42 dB** ($p < 10^{-70}$)	**6.09**	−3.60	`univ` **+9.69 dB**
2	$\mathbf{9.64 \pm 0.10}$	6.91 ± 0.39	`univ` **+2.74 dB** ($p < 10^{-50}$)	**7.40**	4.47	`univ` **+2.94 dB**
4	$\mathbf{9.95 \pm 0.07}$	9.70 ± 0.11	`univ` **+0.25 dB** ($p = 9.9 \times 10^{-7}$)	**7.63**	7.58	tied ($p = 0.69$)
8	10.05 ± 0.05	$\mathbf{10.17 \pm 0.08}$	`softnorm` +0.12 dB	7.75	**8.43**	`softnorm` +0.68 dB
16	10.03 ± 0.03	$\mathbf{10.21 \pm 0.08}$	`softnorm` +0.18 dB	7.71	**8.63**	`softnorm` +0.92 dB
32	9.98 ± 0.03	$\mathbf{10.19 \pm 0.08}$	`softnorm` +0.20 dB	7.65	**8.66**	`softnorm` +1.02 dB

combined via natural linear waveform summation without post-hoc peak or RMS normalization. Preserving raw algebraic addition matters for latent homomorphic mapping: empirical profiling shows that unnormalized sums achieve $+16.15$ dB separation under latent subtraction, whereas artificial peak normalization and RMS normalization degrade algebraic recovery to $+9.96$ dB and $+5.82$ dB, respectively.

Validation splits remain strictly disjoint from training distributions, with test splits drawn exclusively from the 10% of stems held out from autoencoder exposure. Out-of-domain music mixtures (such as MUSDB18-HQ) are excluded from this generative experiment to prevent baseline autoencoding distortion (6.62 dB vs. 16.70 dB in-distribution) from confounding downstream generative trajectory evaluation. Performance is reported across 3 independent PRNG seeds for 2 s audio slices and replicated on 10 s continuous sequences, assessing trajectory straightness, single-step integration, and latent algebraic decodability.

F.2 Comprehensive Step Sweeps, Trajectory Geometry, and Scale Controls

Inference acceleration and the few-step regime. Table 26 and Table 27 examine generative trajectory integration across discretization scales $N \in \{1, 2, 4, 8, 16, 32\}$ using forward Euler steps. The conical inductive bias delivers a large, statistically robust acceleration in few-step inference:

- **Single-step generation ($N = 1$):** Evaluating single-step velocity integration ($t = 1 \to 0$ in a single forward pass) yields a large gap in favor of our conical geometry: $+8.42$ dB SI-SDR at 2 s (9.07 dB vs. 0.65 dB, paired t-test $p < 10^{-70}$) and $+9.69$ dB at 10 s (6.09 dB vs. -3.60 dB). The effect size exceeds cross-seed standard deviations by over two orders of magnitude: the conical velocity field is measurably near-straight ($S = 0.987$, Table 31), so few-step Euler integration already resolves it, whereas the baseline field requires many corrective steps.
- **Rapid trajectory convergence ($N \leq 4$):** On the conical manifold, generative sampling plateaus by $N = 2$–4 steps (9.95 dB at 2 s; 7.63 dB at 10 s). In contrast, the baseline requires 8 to 16 steps to reach comparable fidelity, a practical $4\times$ to $8\times$ step-budget reduction at matched quality.
- **Statistical equivalence at asymptotic budgets ($N \geq 8$):** At high evaluation budgets, the nominal difference between architectures narrows to $+0.12 - +0.20$ dB in favor of the baseline at 2 s (larger at 10 s). Importantly, as established in Appendix B.3, the 2 s margin is a statistical null result: per-seed paired evaluations yield inconsistent significance ($p = 0.21$ in Seed 42, $p = 0.012$ in Seed 43), with spectral metrics remaining statistically indistinguishable across more than one-third of evaluated clips. We therefore treat asymptotic multi-step generation at 2 s as practically equivalent, contrasted against a large, highly significant advantage ($p < 10^{-70}$) in the few-step and latent-arithmetic regimes.

Latent algebra under generative sampling (*latavg*). A key property of an algebraic latent space is whether codes generated independently by a prior model can be manipulated linearly prior to decoding. Table 28 and Table 29 benchmark the estimator ladder across 10 s continuous sequences:

Table 27: Single-draw Euler step sweep at 2 s (SI-SDR in dB, mean ± std across 3 independent DiT seeds) comparing generative flow sampling against deterministic single-step regression baselines (`regr.`), evaluated on the training-time 32-clip validation subset: internally matched and seed-replicated, but not directly comparable in absolute value to the n=128 grids of Table 26. Regressor metrics are independent of step count by construction.

Steps N	`univ RF (Ours)`	`softnorm RF`	`univ regr. (Ours)`	`softnorm regr.`
1	$\mathbf{9.40 \pm 0.06}$	1.02 ± 0.07	$\mathbf{7.94 \pm 0.05}$	4.51 ± 0.19
2	$\mathbf{9.91 \pm 0.12}$	5.61 ± 0.48	$\mathbf{7.94 \pm 0.05}$	4.51 ± 0.19
4	$\mathbf{9.86 \pm 0.05}$	8.76 ± 0.02	$\mathbf{7.94 \pm 0.05}$	4.51 ± 0.19
8	$\mathbf{9.62 \pm 0.08}$	9.15 ± 0.12	$\mathbf{7.94 \pm 0.05}$	4.51 ± 0.19
16	$\mathbf{9.32 \pm 0.09}$	8.99 ± 0.14	$\mathbf{7.94 \pm 0.05}$	4.51 ± 0.19
32	$\mathbf{9.07 \pm 0.09}$	8.82 ± 0.15	$\mathbf{7.94 \pm 0.05}$	4.51 ± 0.19

Table 28: Replication of the full estimator ladder on 10 s full-length audio ($n = 256$, 6 sampling draws). Decodability of averaged generated codes (*latavg*) succeeds exclusively on the conical representation.

Estimator	`univ (Ours)`	`softnorm`
Reconstruction ceiling $D_\phi(E_\psi(x))$	16.80	16.18
Supervised algebraic subtraction ($z_{\text{mix}} - z_{\text{other}}$)	$\mathbf{+15.81}$	-9.88
Mixture passthrough (*no-op*)	-0.64	-0.61
Flow Matching, waveform averaging (*wavavg*)	8.01	**8.94**
Flow Matching, latent code averaging (*latavg*)	$\mathbf{+8.01}$	$\mathbf{-2.98}$
Flow Matching, isolated single draw	5.78	6.84

- **Linear code superposition (*latavg*):** Decoding the algebraic mean of 6 independently generated latent samples ($\bar{z} = \frac{1}{K}\sum_k z^{(k)}$) achieves $+8.01$ dB SI-SDR on `univ`, matching waveform-domain sample averaging (*wavavg*, $+8.01$ dB). On the `softnorm` baseline, latent averaging falls to -2.98 dB (11.0 dB below the conical model), indicating that the unconstrained soft-norm representation does not support linear operations on generated latents at this scale.

- **Single forward-pass conditional regression:** When predicting conditional clean targets via a single forward-pass regressor (Table 29), the conical model reaches $+6.14$ dB SI-SDR (outperforming a full single-draw sampling trajectory of its own flow model, $+5.71$ dB). The baseline regressor reaches only -1.92 dB, below the mixture passthrough (-0.87 dB): the unconstrained decoder does not map these MSE-optimal interior codes to valid audio without conical regularization.

Geometric mechanism: trajectory straightness and optimization landscape. The origin of these operational advantages is revealed by direct geometric and objective profiling:

- **Regression loss convergence:** As documented in Table 30, under strictly matched target energies ($< 0.04\%$ difference), DiT models trained over our conical manifold converge to lower validation losses: $\mathcal{L}_{\text{RF}} = 0.1315$ vs. 0.2106 at 2 s (38% reduction) and 0.0938 vs. 0.1627 at 10 s (42% reduction). The structured geometry provides a smoother, more coherent velocity field for neural regression.

- **Empirical trajectory curvature:** Direct curvature profiling across 200-step Euler integration paths (Table 31) shows that the conical velocity field is measurably straighter: trajectory straightness achieves $S = 0.9869$ (against a theoretical ideal of 1.0), while mean chord deviation is reduced by 27% (0.0323 vs. 0.0440). Straighter velocity trajectories reduce ODE truncation error, consistent with the large single-step integration advantage.

Table 29: Deterministic conditional-mean estimation benchmark at 10 s ($n = 192$ paired clips; identical backbone, budget, and latent representations across models). Regressor models predict targets in a single forward pass without numerical ODE integration.

Diagnostic / Estimator	`univ (Ours)`	`softnorm`
Reconstruction ceiling / Supervised subtraction / *no-op*	16.79 / **+15.85** / −0.92	16.22 / −9.91 / −0.87
Deterministic Regressor (1 forward pass, SI-SDR dB) ↑	**+6.14**	**−1.92**
Regressor spectral error (MS-STFT) ↓	**1.830**	4.717
Regressor latent relative error $\|z - \hat{z}\|/\|z\|$ ↓	**0.834**	0.970
Generative Flow Matching reference (multi-step ODE)		
Flow Matching, 1 sampling draw (SI-SDR dB)	5.71	6.86
Flow Matching, 6 draws averaged (*wavavg*, SI-SDR dB)	7.92	8.95
Flow Matching spectral error (MS-STFT, 6 draws)	1.494	1.304
Flow Matching latent relative error (6 draws)	**0.916**	0.955

Table 30: Rectified Flow validation loss under strictly matched target energies. Conical regularization lowers velocity fitting error by $\approx 40\%$ across sequence lengths.

Optimization Quantity	`univ (Ours)`	`softnorm`
Best validation RF loss (2 s, 3 seeds)	$\mathbf{0.1315 \pm 0.0004}$	0.2106 ± 0.0007
Regression-target energy per element (2 s)	0.9962	1.0000
Best validation RF loss (10 s, single run)	**0.0938**	0.1627
OT-coupled RF target energy per element (10 s)	1.9557	1.9565

F.3 Seed Replication and Autoencoder-Seed Robustness

Generative DiT seed replication. Table 32 reports the individual seed breakdowns across three independently initialized and trained DiT flow matching models ($n = 128$ held-out clips per seed, 384 paired evaluations overall). Across all primary axes of evaluation, the observed effects exceed seed variance by one to two orders of magnitude:

- **Latent arithmetic decodability (*latavg*):** Decoding the latent mean of generated representations achieves an average of $+9.96 \pm 0.03$ dB SI-SDR on `univ`, whereas the baseline collapses to $+0.54 \pm 0.04$ dB, showing a separation margin of $+9.41 \pm 0.06$ dB that is over $150\times$ larger than cross-seed variability. Pairwise spectral win ratios favor our conical model across 128/128 clips in every single run ($p \leq 5.2 \times 10^{-82}$ per seed, cumulative exact sign test $p < 10^{-115}$).
- **Deterministic conditional regression:** Single-step conditional estimation converges to $+6.71 \pm 0.06$ dB on `univ` vs. $+3.42 \pm 0.08$ dB on `softnorm` (a $+3.29 \pm 0.12$ dB lead, with pairwise win rates of 126/2, 124/4, and 126/2 per seed, $p \leq 1.1 \times 10^{-44}$).
- **Waveform-domain averaging (*wavavg*):** Conversely, when multi-sample averaging is conducted post-decoding directly on raw audio waveforms, the baseline displays an edge of $+0.21 \pm 0.09$ dB SI-SDR (9.98 ± 0.02 dB vs. 10.20 ± 0.08 dB). Notably, this small margin exhibits high cross-seed instability ($p = 0.21$ in Seed 42, $p = 0.012$ in Seed 43, and $p = 1.9 \times 10^{-5}$ in Seed 44), with spectral loss differences failing to reach significance on approximately one-third of evaluated clips. We therefore treat unconstrained unconditional synthesis at full budget as at parity, while the advantage in downstream latent arithmetic remains two orders of magnitude larger than seed variability.

Autoencoder-level seed robustness. Table 33 extends the invariance analysis to the autoencoder representation itself, indicating that the emergent algebraic decodability is a property of the conical optimization objective rather than a favorable initialization. Across three independently trained autoencoder runs (`univ`, seeds 42, 43, 44), algebraic separation performance remains tightly clustered between 14.39 dB and 15.99 dB SI-SDR (a narrow 1.6 dB total spread). In every instance, the algebraic separation result remains within 0.72–1.44 dB of the autoencoder's own isolated reconstruction ceiling.

Table 31: Direct geometric curvature measurements of generative flow trajectories (200-step Euler integration, 64 traces per model). Straightness $S = \|\Delta z\| / \int \|v\|\, dt \in [0, 1]$ evaluates path efficiency; path/chord ratio and chord deviations quantify curvature away from straight linear trajectories.

Trajectory Curvature Diagnostic	`univ (Ours)`	`softnorm`	**Ideal Value**
Trajectory straightness S (↑)	$\mathbf{0.9869}$	0.9806	1.0000
Path length / Chord length (↓)	$\mathbf{1.0133}$	1.0198	1.0000
Mean chord deviation (↓)	$\mathbf{0.0323}$	0.0440	0.0000
Maximum chord deviation (↓)	$\mathbf{0.0523}$	0.0648	0.0000

Table 32: Cross-seed replication of conditional generative modeling across three independent DiT training runs (evaluating 128 paired clips per seed, 384 overall). SI-SDR reported in dB. Win rates count per-clip MS-STFT victories; the *latavg* and regressor blocks report two-sided paired t-tests on per-clip MS-STFT, while the *wavavg* block reports two-sided paired t-tests on per-clip SI-SDR.

DiT Seed	`univ latavg` ↑	`softnorm latavg` ↑	**MS-STFT Wins (u / s)**
Latent code averaging (latavg, 6 draws)			
Seed 42	9.99	0.50	128 / 0 ($p = 1.7 \times 10^{-82}$)
Seed 43	9.93	0.53	128 / 0 ($p = 4.6 \times 10^{-83}$)
Seed 44	9.94	0.59	128 / 0 ($p = 5.2 \times 10^{-82}$)
Mean ± Std	$\mathbf{9.96 \pm 0.03}$	$\mathbf{0.54 \pm 0.04}$	**384 / 0** ($p < 10^{-115}$)
Single forward-pass conditional regressor			
Seed 42	6.80	3.38	126 / 2 ($p = 4.5 \times 10^{-50}$)
Seed 43	6.67	3.54	124 / 4 ($p = 5.8 \times 10^{-46}$)
Seed 44	6.68	3.35	126 / 2 ($p = 1.1 \times 10^{-44}$)
Mean ± Std	$\mathbf{6.71 \pm 0.06}$	$\mathbf{3.42 \pm 0.08}$	**376 / 8** ($p < 10^{-99}$)
Waveform-domain sample averaging (wavavg, null result baseline)			
Seed 42	10.02	10.12	Paired $p = 0.21$ (insignificant)
Seed 43	9.96	10.17	Paired $p = 0.012$
Seed 44	9.97	10.31	Paired $p = 1.9 \times 10^{-5}$
Mean ± Std	$\mathbf{9.98 \pm 0.02}$	$\mathbf{10.20 \pm 0.08}$	Cross-seed spread ≤ 0.08 dB

By contrast, the unregularized soft-norm autoencoder reaches only -9.64 dB SI-SDR, falling 25.62 dB below its reconstruction ceiling. The performance gap between geometric paradigms (> 24 dB) exceeds intra-recipe seed variance by a factor of $16\times$, indicating that algebraic decodability is an intrinsic property of conical mixup supervision.

F.4 Validation of the Conditional Mean Estimator

Section 4.4 reveals a clear functional contrast: while the conditional expectation $\mathbb{E}[z_{\text{voice}} \mid z_{\text{mix}}]$ can be directly and accurately decoded into audio on the conical manifold, the identical estimation yields unusable audio on the standard unconstrained soft-norm baseline. This section isolates the mechanism behind this disparity, indicating that the baseline's failure is geometric rather than an artifact of under-fitting, model capacity, or optimization failure.

The latent additivity mechanism. The decodability of the conditional expectation relies entirely on the decoder's operational linearity. Measured directly on representations without involving any downstream neural network (Table 34), the conical decoder inverts the vector sum of independent stem codes $D_\phi(z_{\text{voice}} + z_{\text{other}})$ with high fidelity ($+10.90$ dB SI-SDR), achieving near-parity with decoding the monolithic encoded mixture $D_\phi(z_{\text{mix}})$ ($+10.93$ dB).

In stark contrast, the soft-norm decoder reaches only -3.95 dB SI-SDR when presented with the sum of component codes, despite decoding the single encoded mixture with high fidelity ($+10.40$ dB). Because the conditional expectation under quadratic loss naturally averages overlapping acoustic hypotheses, it produces interior points within the latent convex hull. When a decoder is not explicitly

Table 33: Autoencoder-level seed robustness evaluated across 60 in-distribution mixture trials. All models share identical backbone architectures, training budgets, and datasets. Algebraic decodability is robust across initial PRNG seeds.

Autoencoder Configuration	Recon Ceiling ↑	No-Op	Algebraic Separation ↑	Gap to Ceiling ↓
`univ`, Seed 42	16.70	−0.33	**+15.79**	**0.91**
`univ`, Seed 43	16.71	−0.30	**+15.99**	**0.72**
`univ`, Seed 44	15.83	−0.60	**+14.39**	**1.44**
`softnorm` baseline (Seed 42)	15.98	−0.29	−9.64	25.62

Table 34: Geometric mechanism of latent commutativity measured without generative models. Both decoders invert monolithic mixture codes $D_\phi(z_{\text{mix}})$ with high fidelity (control row), but only the cone-supervised decoder inverts the algebraic sum of component codes $D_\phi(z_{\text{voice}} + z_{\text{other}})$.

Geometric Diagnostic	`univ (Ours)`	`softnorm`
Normalized decoder additivity error: $\|D(z_a{+}z_b) - D(z_a) - D(z_b)\|/\|D(z_a{+}z_b)\|$ (↓)	**0.102**	0.835
Scale homogeneity error, $\alpha = 0.5$ (↓)	**0.044**	0.751
Component sum decoding $D_\phi(z_{\text{voice}} + z_{\text{other}})$ (SI-SDR dB) ↑	**+10.90**	−3.95
Direct mixture decoding $D_\phi(z_{\text{mix}})$ (SI-SDR dB) *(Control)*	+10.93	+10.40
Local first-order sensitivity $\|\nabla D_\phi(z)\|_F$	0.70	0.32

regularized to commute over conical combinations, these interior estimates can fall outside its valid inversion region, degrading the output waveform.

Exclusion of under-fitting: Bayes-linear floors. To verify that the baseline regressor is not merely under-fitting latent targets in an MSE sense, we benchmark performance against optimal analytical linear MMSE estimators (Bayes-linear floors) fitted on 20,000 training mixtures (Table 35). Both neural regressors comfortably surpass every computable linear floor, including full unconstrained and channel-wise Wiener filters (Ephraim & Malah, 1984).

The soft-norm neural regressor outperforms its optimal linear Wiener benchmark by 17% at 2 s and 14% at 10 s (0.6442 vs. 0.7794 and 0.7386 vs. 0.8535). Thus, the baseline regressor successfully solves the conditional latent regression task in MSE terms. Furthermore, the Bayes-linear floor itself is substantially lower on the conical manifold (0.7082 vs. 0.7794 at 2 s; 0.6846 vs. 0.8535 at 10 s), indicating that conical regularization yields a more linearly predictable latent space.

Capacity scaling test: a $7.8\times$ parameter expansion. To test whether massive overparameterization can overcome this decodability barrier, we scaled the soft-norm regressor by a factor of $7.8\times$ (expanding the DiT body from 16.5 M to 129 M parameters) under identical optimization and validation pipelines (Table 36).

Scaling capacity reduces best-validation latent MSE only marginally ($0.6462 \to 0.6422$) and yields a final decoded audio quality of $+5.08$ dB SI-SDR. Despite consuming nearly an order of magnitude more parameters, this scaled regressor remains 5.25 dB inferior to the 6-draw average of a modest 16.5 M generative flow model ($+10.33$ dB, $p = 1.4 \times 10^{-69}$). Notably, for both the base and wide soft-norm regressors, the checkpoints corresponding to optimal latent MSE decode to negative SI-SDR values (-0.44 dB and -0.17 dB), falling below the mixture passthrough floor ($+2.95$ dB). This indicates that minimizing latent Euclidean MSE on an unconstrained manifold does not translate into decoded audio fidelity, supporting the practical importance of conical decoder supervision for decodable conditional estimates.

F.5 Methodological Diagnostics, Geometric Inductive Biases, and Outlook

Structured error tolerance vs. isotropic decoder sensitivity. Direct waveform-level comparisons across distinct representation spaces require careful calibration against decoder sensitivity.

Table 35: Verification that neural regressors do not under-fit latent targets: both models sit strictly below optimal Bayes-linear floors at all sequence lengths. Values report MSE per element (↓). The regressor row represents the optimal validation checkpoint across base training runs.

	2 s Audio Slices		10 s Continuous Audio	
Estimator	`univ (Ours)`	`softnorm`	`univ (Ours)`	`softnorm`
Mixture identity passthrough (*no-op*)	0.9924	1.1086	1.2184	1.3065
Channel-wise Wiener filter	0.7082	0.7794	0.8104	0.8575
Ridge regression on PCA-640 features	0.7828	0.8961	0.6937	0.8697
Full cross-channel Wiener filter (tuned)	0.7496	0.8467	0.6846	0.8535
Trained Neural Regressor (Best Ckpt)	**0.4957**	**0.6442**	**0.6077**	**0.7386**
Neural gain over best linear floor	**−30%**	**−17%**	**−11%**	**−14%**

Table 36: Capacity scaling ablation on the unconstrained soft-norm latent: scaling network capacity by $7.8\times$ (129 M parameters) fails to resolve the decodability barrier. Evaluated across 192 paired validation clips (Reference anchors: mixture passthrough $+2.95$ dB; reconstruction ceiling $+16.47$ dB; supervised subtraction -4.63 dB).

Diagnostic / Metric	**Base Regressor (16.5 M)**	**Wide Regressor (129 M)**	**Flow Matching (16.5 M, 6 Draws)**
Best val. latent MSE (↓)	0.6462	**0.6422**	
Final val. latent MSE (↓)	1.0321	**1.0427**	
SI-SDR, final ckpt (dB) ↑	+3.37	**+5.08**	**+10.33**
MS-STFT, final ckpt (↓)	2.727	**2.221**	**1.188**
SI-SDR at best-val MSE (dB)	−0.44	**−0.17**	
Final latent rel. error (↓)	0.9923	0.9993	(*latavg*: +0.54 dB)

Under isotropic, unstructured Gaussian noise, the cone-supervised decoder exhibits an error gradient $1.85\times$ steeper than the soft-norm decoder (local first-order sensitivity ratio of $2.2\times$), as unconstrained decoders tend to smooth out random off-manifold directions.

Importantly, this sensitivity disparity does not account for downstream generative performance. Evaluating generative flow trajectories reveals that model velocity estimation errors are highly *structured*: along these learned low-dimensional pathways, structured error is $4.8\times$ less damaging than isotropic noise (producing an MS-STFT distortion of only 1.49 for `univ`, compared to 7.14 under random perturbations of equivalent norm $\epsilon = 0.40$). Within this structured regime, both decoders display comparable tolerance (1.49 vs. 1.30). More importantly, sensitivity cannot explain large directional leads such as the $+8.4$ dB single-step integration advantage of `univ`, since higher sensitivity would penalize rather than favor the conical manifold.

Geometric attribution across multi-metric autoencoder baselines. While the reference and baseline autoencoders share identical depth, width, and optimization budgets, their training objectives induce differing global geometry alongside additivity (Table 1). Our ablation suite (Appendix C.1) identifies $\mathcal{L}_{\text{UMix}}$ as the driver of decoder linearity and flow trajectory straightness. Cross-seed replication across three independent autoencoder initializations (Table 33) shows that this algebraic capacity is robust ($16\times$ larger than the inter-seed spread), indicating that the geometric structure is a property of conical supervision rather than of a particular initialization.

Calibration and objective alignment in deterministic estimation. While generative flow models are largely invariant to checkpoint selection along their training trajectories, deterministic regressors display distinct objective alignments. Because Euclidean MSE in an unconstrained latent space does not correspond to perceptual audio fidelity, early-stopped soft-norm regressors selected via optimal validation MSE decode below the unseparated mixture passthrough baseline (-0.53 dB vs. $+2.58$ dB at 2 s). In contrast, the conical manifold aligns minimum Euclidean distance with

valid acoustic superposition, so that optimizing standard regression objectives translates directly into improved decoded audio.

Trajectory geometry across integration budgets. Evaluating the full 2×2 grid crossing training and evaluation chunk lengths at 32 Euler steps confirms the structural basis of the multi-step crossover:

- **2 s-trained models:** Evaluated at 32 steps, `softnorm` achieves $+0.20$ dB at 2 s ($p = 9.4 \times 10^{-6}$) and $+0.64$ dB at 10 s (significant in every seed; per-seed p from 4.9×10^{-7} to 4.2×10^{-4}).
- **10 s-trained models:** Evaluated at 32 steps, `softnorm` achieves $+1.83$ dB at 2 s ($p = 7.1 \times 10^{-17}$) and $+1.02$ dB at 10 s ($p = 2.2 \times 10^{-15}$).

This behavior reflects the complementary geometric inductive biases of the two spaces. The conical velocity field is measurably straighter (Table 31), concentrating trajectory resolution within the initial $N = 2$–4 steps and enabling fast, low-latency generation. The unregularized baseline follows more curved ODE trajectories that require multiple corrective integration steps to resolve, leaving it weak in low-compute or single-step settings (0.65 dB at $N = 1$) and leaving latent averaging unusable (0.54 dB on `latavg`).

Trajectory straightness and fast-sampling convergence. On the conical manifold, generative sampling rapidly reaches its best single-draw fidelity by $N = 2$ Euler steps (9.91 dB at 2 s), followed by a slow decline toward $N = 32$ (9.07 dB; Table 27). The same shape appears under higher-order integrators (midpoint RK2), which makes it slightly worse at every step count, so the decline is not a Euler-discretisation artefact; we report the shape as an open observation of the conical field rather than an explained property. The practical consequence is that the conical latent's advantage is concentrated in the few-step regime ($N \leq 4$), where it reduces or removes the need for multi-step ODE solvers.

A hypothesis for the single-draw decline. Three measured facts constrain the explanation: (i) midpoint (RK2) integration is worse at every step count, so the effect is a property of the learned field rather than of Euler discretisation; (ii) averaging six decoded draws almost entirely removes it (the *wavavg* curve drifts by only 0.07 dB where the single-draw curve drops 0.84 dB), so the refined endpoints differ mainly by per-draw variance, not by a systematic displacement that would survive waveform averaging; and (iii) the baseline shows the same pattern after its own plateau ($9.15 \rightarrow 8.82$ dB from N=8 to 32), only later and milder. Our working hypothesis is that finer integration trades truncation error against exposure to the learned field's residual error. The conical latent is pinned close to the unit-Gaussian prior in its aggregate statistics by construction (the moment penalty of Eq. (19); per-element variance ≈ 0.97 and near-orthogonal independent codes, Table 12), so the useful signal of the learned velocity is concentrated in its near-linear principal part, which one or two Euler steps already capture (the few-step regime rectified flows are designed for, with path straightness the controlling quantity (Liu et al., 2023b)) and whose near-straightness is precisely the few-step advantage. The nonlinear residual of the learned field carries little useful transport but non-negligible error, and each additional step integrates more of it into the sample: refining the discretisation converges to the *learned* field's ODE rather than to the data distribution, and sampler analyses in the diffusion setting likewise separate model error from discretization error, with perturbation and truncation acting as regularizers (Karras et al., 2022). RK2 tracks the learned field more faithfully and is accordingly worse, and the residual is approximately zero-mean across draws, so the multi-draw average is unaffected. The same exposure exists in the baseline but is masked until $N \approx 8$ by its larger truncation error. The hypothesis is testable by decoding $z(t)$ along the trajectory: per-draw error should accumulate gradually after the first steps and average out across draws. To our knowledge, the specific step-count decline we measure has not been documented for nearly-straight learned fields; running this profiling experiment would establish it, and we state the mechanism as conjecture. None of our claims rests on the shape of the field: they concern the few-step regime ($N \leq 4$), where the conical advantage is largest and statistically robust.

Scope, inductive biases, and future directions. This work establishes that autoencoder bottleneck representations can be regularized into continuous linear cones without sacrificing single-

source reconstruction fidelity. In downstream generative modeling, this provides distinct functional capabilities:

1. **Few-step and real-time generation:** Rather than targeting asymptotic high-budget sampling where unconstrained iterative ODEs eventually catch up, conical flow fields are designed for accelerated, few-step synthesis ($N \leq 4$), where velocity straightness yields large quality gains at fixed low step budgets ($+8.42$ dB at $N = 1$).
2. **Decodable latent algebra:** Conical supervision enables direct linear combination and conditional expectation decodability within the latent space ($\bar{z} = \frac{1}{K} \sum_k z^{(k)}$), addressing a known limitation of unconstrained generative representations, where averaging latent states degrades output quality.
3. **Foundation for algebraic audio synthesis:** Extending conical supervision to large-scale unconditional generative foundations (e.g., text-to-audio diffusion models evaluated via distribution metrics such as FAD (Kilgour et al., 2019)) presents an immediate next step. Enforcing continuous conical commutation suggests a path toward controllable multi-track mixing, zero-shot stem editing, and arithmetic prompt composition directly inside the latent dynamics of generative transformers.