TADA! TUNING AUDIO DIFFUSION MODELS THROUGH ACTIVATION STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio diffusion models can synthesize high-fidelity music from text, yet achieving fine-grained control over specific musical attributes remains challenging, as their internal mechanisms for representing high-level concepts are poorly understood. In this work, we use activation patching to demonstrate that recent audio diffusion architectures exhibit a semantic bottleneck, where a small, shared subset of consecutive attention layers controls distinct musical concepts, such as the presence of specific instruments, vocals, or genres. Building on this, we systematically evaluate a broad spectrum of steering paradigms, comparing activation steering against prompt-level, score-space, and weight-space interventions, analyzing the interaction between the steering mechanism and the intervention site. Our new benchmark, supported by an extensive user study, demonstrates that localized activation steering establishes a new state-of-the-art in audio concept modulation. Audio examples are provided at `https://audio-steering.github.io`.

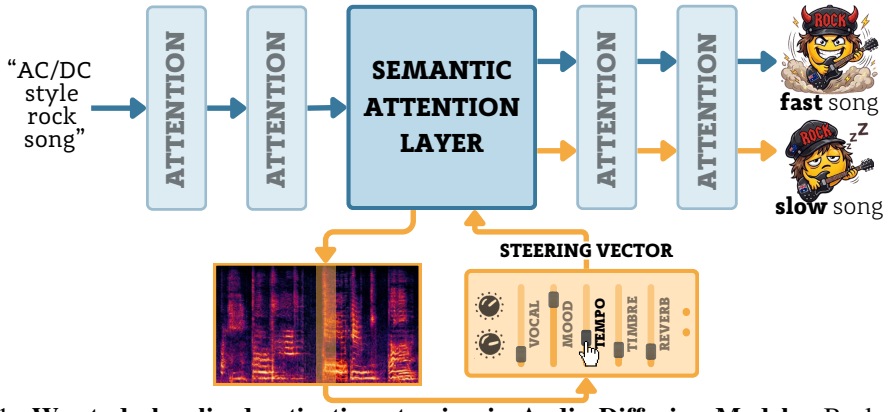


Figure 1: **We study localized activation steering in Audio Diffusion Models.** By localizing functional layers, we enable precise modulation of audio concepts through activation steering.

1 INTRODUCTION

Diffusion-based text-to-music models can synthesize coherent, full-length musical pieces from a single text prompt. Analogous technologies (Suno, Inc., 2026) have established consumer products in routine use among hobbyists, advertisers, podcasters, and game audio designers. Practical creative use of these systems, however, exposes a fundamental limitation of the prompt-based interface: it does not support fine-grained control. A user can request "a fast, lively rock song" and obtain a plausible result, but cannot request a "fast but not too fast tempo", a "slightly lower vocal pitch", or the same arrangement with the more feminine vocal. Re-issuing a perturbed prompt regenerates an entirely different piece. Music is a continuous medium, and creative work with it depends on small adjustments along specific perceptual axes that prompt-based interfaces do not expose.

In language and image generation, this kind of fine-grained control has been achieved by intervening directly inside the model through: manipulations of prompts and their embeddings (Görgün et al., 2026; Baumann et al., 2025), weight adaptations (Gandikota et al., 2024), or activation-space steering (Rimsky et al., 2024; Huben et al., 2024; Gao et al., 2025; Feng et al., 2026). We argue that music's continuous perceptual structure makes it a more natural target for such techniques than in those other domains. Yet, audio adaptations have so far been narrow, focusing on autoregressive music models (Koo et al., 2025; Facchiano et al., 2025; Singh et al., 2026).

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.

In this work, we address these limitations by posing two core questions: *where* and *how* should we intervene in the audio diffusion models to achieve smooth, precise control over semantic concepts?

To answer *where*, we apply activation patching (Meng et al., 2022) across a curated set of counterfactual prompt pairs covering vocals, tempo, mood, instruments, and genres. We evaluate three models spanning the dominant architectural families: AudioLDM2 (Liu et al., 2024) (Diffusion U-Net), Stable Audio Open (Evans et al., 2025) (Diffusion Transformer), and Ace-Step (Gong et al., 2025) (Flow-Matching Transformer). The empirical picture is consistent across all architectures: every model exhibits a 'semantic bottleneck.' The control for high-level musical concepts is densely concentrated in a small, shared subset of cross-attention layers: two or three out of twenty-four in the transformer-based models, and a comparably narrow band in the U-Net.

To answer *how*, we systematically evaluate a wide spectrum of steering paradigms. We primarily focus on activation steering techniques and compare them with prompt-level interventions, score-space modifications, and weight-space techniques that we adapt to the music domain. To understand the dynamics of the discovered bottleneck, we investigate the role of localization by comparing each method's standard global application with a localized variant restricted to the functional layers.

Through a rigorous evaluation protocol grounded in the alignment-preservation trade-off across musical concepts, we uncover a crucial interaction between the steering mechanism and the intervention site. We show that restricting interventions to the semantic bottleneck greatly improves the effectiveness of activation-based techniques. Our localized steering avoids collateral acoustic damage and successfully isolates the conceptual change. Consequently, it establishes a new state-of-the-art in audio concept modulation, significantly outperforming both global activation steering and strong baselines across all other paradigms. These objective metrics are corroborated by an extensive listening study, confirming that localized activation steering yields the most seamless and natural integration of desired musical attributes. Our contributions can be summarized as follows:

1. We identify a *semantic bottleneck* in distinct text-to-music diffusion models: a small, shared subset of cross-attention layers governing the majority of interpretable musical concepts.
2. We introduce a comprehensive suite of steering techniques for text-to-music diffusion models and show that restricting activation steering to selected functional layers establishes a new state-of-the-art in audio concept modulation.
3. We propose an evaluation protocol grounded in the alignment–preservation trade-off, instantiate it as a benchmark over musical concepts, and validate it with a listening study whose preferences are consistent with the automatic metrics.

We release the benchmark, trained artifacts, and our adaptations of all evaluated steering methods at **[hidden for review]** to support further research on controllable music generation.

2 Related work

Interpretability and Steering of Generative Models. Mechanistic interpretability is emerging as a research direction aimed at understanding the internal computations of large-scale architectures. The objective of Causal Mediation Analysis (Pearl, 2001; Meng et al., 2022) is to understand how the model output changes under interventions in its computational graph. Recently, this technique has been applied to the image domain (Basu et al., 2024b;a; Staniszewski et al., 2025; Zarei et al., 2025), thereby uncovering the mechanistic role of DMs' Attention layers. Precisely, Basu et al. (2024b) employ activation patching and Basu et al. (2024a) use prompt injection to localize layers controlling model knowledge in U-Net-based DMs. Staniszewski et al. (2025) combine these techniques to identify layers that control text generated in images with DiTs. Finally, Zarei et al. (2025) show that important attentions in DiTs can be efficiently traced by the magnitudes of their outputs.

On the other hand, the linear representation hypothesis (Park et al., 2024) posits that neural networks encode high-level concepts as linear directions in the activation space. Recently, Rimsky et al. (2024) proposed the Contrastive Activation Addition (CAA) technique, showing that differentiating neural network hidden states from runs with prompts describing opposite concepts yields an activation difference that encodes the semantic change between the prompts. This approach has been widely applied to Large Language Models (LLMs), e.g., for steering model behavior (Chen et al., 2025; Feng et al., 2026) or reducing its toxicity (Rodriguez et al., 2025a).

In vision, Rodriguez et al. (2025a;b) steer generations by training affine maps that represent transformations between activation distributions in distilled text-to-image models. Similarly, difference-based methods have been applied in the text encoder space of diffusion models (Baumann et al., 2025) for concept modulation, or in attention layers (Gaintseva et al., 2026) for concept removal. Beyond activation contrasting, Sparse Autoencoders (SAEs) (Olshausen & Field, 1997b) have been applied to LLMs (Huben et al., 2024; Bricken et al., 2023) to decompose activations into sparse, interpretable features in an unsupervised manner by training an autoencoder to reconstruct them with a sparsity constraint. In the image domain, Surkov et al. (2025) demonstrated that SAEs can capture meaningful features inside DMs, and Cywiński & Deja (2025) ablated features within the SAE latent space for concept unlearning. Apart from activations, Gandikota et al. (2024) steer concepts via trainable low-rank adapters that represent meaningful directions in the weight space.

Interpretability and Control of Audio Models. Recent research has begun to adapt interpretability techniques to the audio domain to understand models and control their internal computations. Initial efforts (Sadov, 2024) identify interpretable circuits in the ASR model via feature discovery techniques. Prior works have also studied text-to-music models, focusing on autoregressive architectures (Music LLMs). In particular, Wei et al. (2024) introduced a dataset to probe whether music foundation models encode specific music-theory concepts, such as intervals and chords. Similarly, Vásquez et al. (2024) probe for the information about instruments and genres. Moving towards control in Music LLMs, Koo et al. (2025) use classifier probes to steer attention heads for specific musical traits, Zhao et al. (2026) adapt Recursive Feature Machines for steering, while Facchiano et al. (2025) use activation patching to manipulate binary attributes. Recently, SAEs have been employed in the audio domain to steer the residual stream of MusicGen (Singh et al., 2026), or map features to acoustic concepts within audio autoencoders (Paek et al., 2025) and speech transformers (Aparin et al., 2026). However, despite the recent high performance of diffusion-based music generators (Gong et al., 2026), interpretability and steering of Audio DMs remain largely underexplored. Recent works predominantly focus on audio editing by manipulating attention maps (Xu et al., 2024; Liu et al., 2025a; Yang et al., 2026), or by combining the inversion process with intermediate latents optimization (Liang et al., 2024; Jia et al., 2025; Zhang et al., 2024; Manor & Michaeli, 2024; Niu et al., 2026; Lee et al., 2026). Notably, Yang et al. (2026) probe self-attention maps within AudioLDM 2 (Liu et al., 2024), achieving better preservation of temporal structure. These methods perform a single instruction-based edit to an existing recording, whereas steering aims to continuously modulate a target concept during generation, yielding generation variants along a concept axis rather than a single track edit. Closest to our work, Ezra et al. (2025) evaluate diffusion score-based methods for steering acoustic concepts with Stable Audio (Evans et al., 2025). In this work, we focus on musical concepts, systematically benchmark our activation steering against their score-based approach, and further analyze it within our localization framework.

3 Background and Methodology

Audio Diffusion Models. Modern audio DMs operate in a compressed latent space and employ a U-Net (Ronneberger et al., 2015) or a Transformer (Vaswani et al., 2017; Dosovitskiy et al., 2021) as their backbone architecture. Unlike text-to-image models processing spatial patches, audio latents are structured as sequences of temporal frames $z_t \in \mathbb{R}^{F \times d}$. Thus, each token f corresponds to a distinct timeframe in the audio. The cross-attention role is to introduce the text information into the hidden states $\mathbf{h} \in \mathbb{R}^{F \times d}$ through the prompt embedding $\mathbf{c} \in \mathbb{R}^{C \times d_c}$. The output of the l-th Cross-Attention block at step t is then

$$\text{CrossAttn}(\mathbf{h}_{l-1}^{(t)}, \mathbf{c}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}\mathbf{W}_O, \tag{1}$$

where $\mathbf{Q} = \mathbf{h}_{l-1}^{(t)}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{c}\mathbf{W}_K$, $\mathbf{V} = \mathbf{c}\mathbf{W}_V$, and $\{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O\}$ are the learned weights. This output is further added to the residual stream as $\mathbf{h}_l^{(t)} = \mathbf{h}_{l-1}^{(t)} + \text{CrossAttn}(\mathbf{h}_{l-1}^{(t)}, \mathbf{c})$.

Activation Patching. To identify which layers control specific musical concepts, we employ activation patching (Meng et al., 2022) illustrated in Fig. 2. For a concept c (e.g., "male voice"), we define a set of counterfactual prompt pairs $(\mathcal{P}_c, \mathcal{P}_{\tilde{c}})$ where $\mathcal{P}_c$ contains the concept and $\mathcal{P}_{\tilde{c}}$ its semantic counterpart.

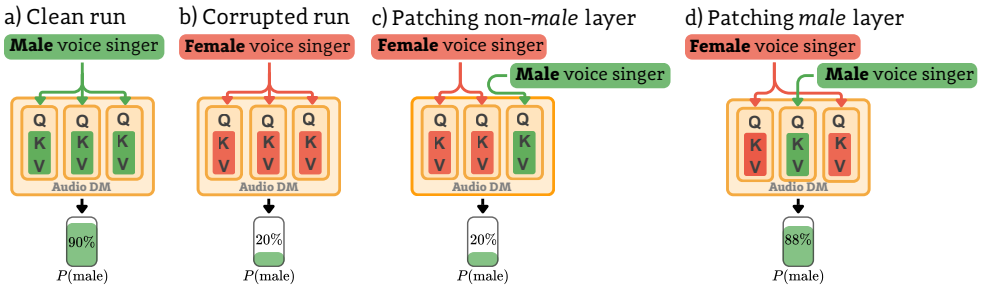


Figure 2: **Layer localization via Activation Patching.** For a target concept, we perform a clean run (a) with a prompt referring to the concept and save the attention key (**K**) and value (**V**) activations. In a corrupted run (b), we infer with a counterfactual prompt. Next (c,d), we run a corrupted run, substituting the l-th layer key and value activations with the saved ones from the clean run. If (d) patching inside l produces audio with the target concept, we identify it as a functional layer.

We generate audio with the target prompt $\mathcal{P}_c$, caching the attention keys $\mathbf{K}_l = \mathbf{c}\mathbf{W}_K$ and the values $\mathbf{V}_l = \mathbf{c}\mathbf{W}_V$ at each layer l. Then, generating audio with the source prompt $\mathcal{P}_{\tilde{c}}$, we patch layer l by substituting its keys and values with the cached ones. Finally, we compare how the intervention affects the similarity between the generated audio and the prompt that describes the concept c.

If the output audio for the patched run exhibits concept c, we identify l as the diffusion model's *functional layer* for c. Precisely, the impact of layer l on concept c is calculated as:

$$\mathrm{I}(l, c) = \frac{\mathrm{s}(l \leftarrow c, l' \leftarrow \tilde{c}) - \mathrm{s}(l \leftarrow \tilde{c}, l' \leftarrow \tilde{c})}{\mathrm{s}(l \leftarrow c, l' \leftarrow c) - \mathrm{s}(l \leftarrow \tilde{c}, l' \leftarrow \tilde{c})}, \tag{2}$$

where $\mathrm{s}(l \leftarrow c_1, l' \leftarrow c_2)$ denotes the audio-text similarity between the concept name and a generation where layer l inputs prompt c_1 while all other layers $l' = L \setminus \{l\}$ receive c_2. The denominator serves as a normalization constant representing the upper bound of similarity gain, ensuring $\mathrm{I}(l, c) \in [0, 1]$. Each layer's impact score is calculated separately for each concept, then aggregated by averaging across concepts: $\mathrm{I}(l) = \frac{1}{|C|}\Sigma_{c \in C}\mathrm{I}(l, c)$. The final set of functional layers ($\mathcal{L}$) consists of ones that score above a predefined threshold τ: $\mathcal{L} = \{l \in L : \mathrm{I}(l) \geq \tau\}$.

Activation Steering. Activation steering intervenes on a model's hidden states (internal representations) at inference time to control its final output. Steering Vector $\mathbf{v}_c$ encoding the concept c is obtained by contrasting model's intermediate activations $\{(h_c^i, h_{\tilde{c}}^i)\}_{i=1}^N$, collected from N paired positive (concept-amplifying, c) and negative (concept-suppressing, $\tilde{c}$) prompts runs. Finally, $\mathbf{v}_c$ is added to layer l outputs as $\mathbf{h}_l' = \mathbf{h}_l + \alpha \cdot \mathbf{v}_c$. The coefficient $\alpha \in \mathbb{R}$ controls both the direction of the modulation (positive values amplify the concept, negative values suppress it) and its strength.

During experiments, we employ diverse steering methods, which differ in how $\mathbf{v}_c$ is constructed. For *Contrastive Activation Addition* (CAA (Rimsky et al., 2024)), we follow Gaintseva et al. (2026) and compute $\mathbf{v}_c$ as the mean difference between positive and negative activations, precisely $\mathbf{v}_\mathbf{c}^{\mathrm{CAA}} = \frac{1}{N}\Sigma_{i=1}^N(\bar{h}_c^i - \bar{h}_{\tilde{c}}^i)$, with $\bar{h}$ denoting cross-attention layer outputs averaged across temporal frames.

In *AUSteer* (Feng et al., 2026), $\mathbf{v}_c$ is forced to be sparse by deriving the ranking activations' dimensions with a signed contrastive score. For each layer ℓ and activation dimension d, AUSteer measures the fraction of contrastive pairs for which the activation difference $h_c^{i,\ell}[d] - h_{\tilde{c}}^{i,\ell}[d]$ agrees in sign. This yields scores $\beta_{\ell,d} \in [-1, 1]$ whose magnitude reflects the directional reliability of the given dimension. A global top-s selection over scores $|\beta_{\ell,d}|$ produces a sparse steering vector at layer ℓ through dimensions $d \in 1 \ldots D$ as $\mathbf{v}_\mathbf{c}^{\mathrm{AUS}}[d] = \mathbb{1}_{|\beta_{\ell,d}| \in \text{top-}s} \cdot \beta_{\ell,d}$. Notably, AUSteer, through its ranking-and-selection mechanism, inherently performs localization.

Finally, to discover interpretable directions within activations, we train TopK *Sparse Autoencoders* (SAEs (Olshausen & Field, 1997a; Gao et al., 2025; Bussmann et al., 2024)). Here, an autoencoder architecture is trained to reconstruct activations with a TopK sparsity constraint. To identify concept-specific features, we use a set of positive and negative activations and perform TF-IDF-like scoring within the autoencoder latent space. The final steering vector for concept c is built by summing SAE's decoder W_{dec} columns corresponding to the top-τ_c scoring features $\mathcal{F}_c$ within SAE latent space, as $\mathbf{v}_\mathbf{c}^{\mathrm{SAE}} = \sum_{j \in \mathcal{F}_c} W_{\mathrm{dec}}[:, j]$. In Appendix I.1.4, we describe these methods in detail.

4 LAYER LOCALIZATION

To understand *where* an intervention enables precise concept steering, we measure the causal importance of individual cross-attention layers in text-to-audio models. We construct a dataset of counterfactual prompt pairs that considers musical concepts: *vocal gender*, *tempo*, *mood*, instruments, and genres. For each contrasting pair $(c, \tilde{c})$, we select captions from the MusicCaps (Agostinelli et al., 2023) containing keywords associated with concept c (e.g., 'female voice') but excluding any terms related to the alternative variant $\tilde{c}$ (e.g., 'male voice'). For genres and instruments, we replace concepts (e.g., 'violin') with alternatives (e.g., 'trumpet'). We select up to 256 such prompts as target ones $\mathcal{P}_c$ and construct alternatives $\mathcal{P}_{\tilde{c}}$ via rule-based replacement of concept-associated terms with their counterparts. We employ GPT-4 (Achiam et al., 2023) to find alternative forms for concepts and propose substitution rules (see examples in App. E), which we manually verify.

Using a constructed dataset, we apply the activation patching procedure described in Section 3 to three state-of-the-art audio diffusion models: AudioLDM2 (Liu et al., 2024), Stable Audio Open (Evans et al., 2025), and Ace-Step (Gong et al., 2025). We generate 10-second (AudioLDM2, Stable Audio Open) and 30-second (Ace-Step) waveforms, using 8 different random seeds per prompt, resulting in 2048 generations per concept. We use MuQ (Zhu et al., 2025) for assessing mood, tempo, instruments, and genres, and CLAP (Wu et al., 2022) for vocal gender.

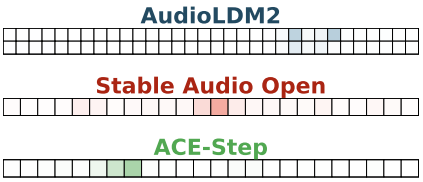


Figure 3: **Average impact scores $\mathrm{I}(l)$ of cross-attention layers (denoted with color intensity in □) within text-to-music DMs.** Singular layers control different concepts: vocal, tempo, mood, instruments, and genres across diverse audio diffusion architectures. Plots for all concepts are provided in Appendix F.

Fig. 3 presents the layer-wise impact scores (aggregated across concepts) for each model. For all three architectures, we observe that a small subset of layers concentrates the control over musical concepts. Using a selection threshold of $\tau = 0.10$, in AudioLDM2 (U-Net), we localize key components in the decoder, specifically the layers 45–51 (7 out of 64). In transformer-based architectures, we observe a high degree of control in the middle layers: in cross-attentions {7,8} for Ace-Step, and {12,13,14} in Stable Audio Open.

These findings show that audio diffusion models develop a **semantic bottleneck** – a set of interpretable, functionally specialized layers shared across various semantic concepts. Additionally, such specialized control is not limited to a single model but appears to be a general property of music DMs. These observations contrast with studies on image generators Basu et al. (2024a); Staniszewski et al. (2025); Zarei et al. (2025), where localized components remain largely disjoint across concept categories. Such surprising regularity opens the door to future studies (see App. C) on layer transfer and interchangeability (Lenc & Vedaldi, 2015) across text-to-audio models.

5 CONTINUOUS AUDIO CONCEPT MODULATION

The high degree of semantic control concentrated in single layers suggests their activations as a natural intervention site: restricting modifications to the semantic bottleneck should isolate the desired change, leaving the rest of the audio intact. To answer the question of *how* to precisely modulate the concept c in the generated track, we adapt a broad set of steering paradigms from other domains to music generation and benchmark them against activation steering localized to the bottleneck.

5.1 BASELINES

In addition to the activation steering methods (Sec. 3), we evaluate techniques operating in other levels: prompts, weights, and diffusion prediction. Detailed descriptions are provided in App. I.1.

Prompt-level interventions. The most direct way to influence a text-to-music model is through its conditioning. For fine-grained control, *PCI* (Görgün et al., 2026) inputs the model with a prompt describing the modulated concept starting from a chosen diffusion timestep, ensuring smooth, natural manipulation in the conditioning space. Alternative approaches alter conditioning embedding itself: *Text Embeddings* (Ezra et al., 2025) add to prompt embeddings a difference between concept-amplifying and concept-suppressing embeddings, while *Token Embeddings* (Baumann et al., 2025) restrict this approach to tokens referring to the modulated subject (e.g., *'instrument'* when adding piano), leaving the rest of the prompt untouched.

Weights-level interventions. *Concept Sliders* (Gandikota et al., 2024) encode concepts directions as low-rank weight updates (Hu et al., 2022) learned during model fine-tuning. During inference, learned weights are further merged with the base model at varying scales, serving as the steering-strength α.

Score-level interventions. *FreeSliders* (Ezra et al., 2025) offer a training-free alternative to Concept Sliders, where direction is encoded as the difference between the diffusion model's noise prediction on contrastive (amplifying and suppressing) prompts and further added to standard DM prediction.

5.2 Evaluation protocol

For robust evaluation, we construct a new benchmark that assesses the modulation of 9 musical concepts. Methods are compared over a range of steering strengths $\alpha \in \{-\alpha_{\text{max}} \dots 0 \dots \alpha_{\text{max}}\}$, with positive (concept-amplifying, $\alpha > 0$) and negative (concept-suppressing, $\alpha < 0$) directions.

Steering effectiveness. No single metric adequately captures the performance of a steering method. Modulation involves an inherent trade-off between conserving the characteristics of the original audio (*preservation*) and increasing the presence of the target concept (*alignment*) – e.g., raising the *tempo* inevitably alters the song structure. An effective intervention is therefore one that, at a matched level of preservation, achieves better alignment. It introduces the desired change, leaving the remaining musical content intact. Inspired by the perception–distortion trade-off (Blau & Michaeli, 2018), we quantify *steering effectiveness* as the area under the alignment–preservation curve, as in Fig. 5.

For audio preservation, we use LPAPS (Iashin & Rahtu, 2021). We measure perceptual distance from the unsteered baseline (when $\alpha = 0$), and define the preservation score for audio generated with strength α as $p(\alpha) = \text{LPAPS}(\alpha_{\text{max}}) - \text{LPAPS}(\alpha)$, where higher values indicate better conservation of the original audio. To compare methods fairly, we calibrate their steering range $(-\alpha_{\text{max}}, \alpha_{\text{max}})$ so that their maximum distortion matches the distance between two audios generated by the diffusion model from different textual prompts. Such a reference represents the natural level of audio change expected from a prompt-level intervention, provides an interpretable reference, and, technically, is obtained by the *PCI* method when steered at maximum strength. In App. H.3, we present an analysis of additional decomposable audio preservation metrics, including harmony, rhythm, and melody.

For alignment, we use the sign-corrected delta $\Delta a(\alpha) = \text{sign}(\alpha) \cdot (a(\alpha) - a(0))$, where $a(\cdot)$ is the similarity between audio steering with given strength and the textual description of the target direction. Similarity is measured with CLAP (Wu et al., 2022) and MuQ (Zhu et al., 2025). Sign correction ensures that both steering directions are on a common axis, with higher values indicating stronger alignment. The final **AUC (↑)** is then defined as $\int_0^{P_{\text{max}}} \Delta a(p)\, dp$ and computed over the α-swept pairs: $(p(\alpha), \Delta a(\alpha))$.

Alternative steering axes. To complete the benchmark, we measure **Smoothness (↓)** (Ezra et al., 2025), which captures how uniformly a method increases alignment along the steering scale. We normalize the alignment deltas to $[0, 1]$, compute consecutive gaps $g_i = \Delta a_{i+1} - \Delta a_i$ ordered by increasing $|\alpha|$, and report $\text{S} = \text{std}(\{g_i\})$. We further assess **Audio Quality (↑)** with Audiobox Aesthetics (Tjandra et al., 2025), averaging the Content Enjoyment (CE), Content Usefulness (CU), Production Complexity (PC), and Production Quality (PQ) scores.

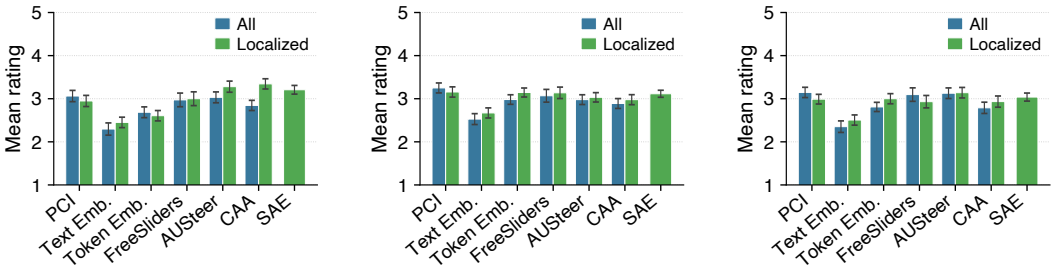
Technical details. We adapt the methods to the Ace-Step (Gong et al., 2025), a state-of-the-art text-to-music model which significantly outperforms AudioLDM2 and Stable Audio Open in the generative capabilities. We generate $N = 100$ thirty-second tracks across 15 positive and 15 negative steering strengths. For activation steering, concept-suppressing vectors are constructed from their respective concept-amplifying vectors through negation. Working with diffusion models, we construct steering vectors for each diffusion step separately. More technical details, including method ablations and hyperparameter analyses, are provided in Appendix I.2.

5.3 Results

Table 1 compares all evaluated steering paradigms, averaged across nine musical concepts, while individual results are provided in App. L. Activation steering restricted to the functional layers resolves the alignment–preservation trade-off most effectively: SAE tops the AUC ranking under both alignment metrics (0.118 MuQ, 0.070 CLAP), with localized CAA and AUSteer completing the top three under MuQ. Under CLAP, Concept Sliders (0.067) is a close runner-up, with localized activation steering variants following closely. The non-localized activation steering, PCI, and FreeSliders form a middle tier, while remaining prompt-level interventions trail. In particular, Text Embeddings achieves a near-zero AUC because its embedding-level distortions quickly fall outside the preservation window.

Table 1: **Audio concept modulation with Ace-Step.** Localized activation steering outperfroms prompt-, score-, and weight-level methods. Results are averaged over 9 musical concepts, steered in both positive and negative directions (see App.M for details). **Best** and second results are highlighted.

Method	AUC (↑)		Smoothness (↓)		Audio Quality (↑)
	MuQ	CLAP	MuQ	CLAP	
PROMPT-LEVEL INTERVENTIONS					
PCI (Görgün et al. (2026))	$0.084_{\pm.019}$	$0.049_{\pm.010}$	$0.185_{\pm.064}$	$0.223_{\pm.066}$	$\mathbf{6.803}_{\pm.016}$
Text Embeddings (Ezra et al. (2025))	$0.014_{\pm.007}$	$0.011_{\pm.004}$	$0.439_{\pm.224}$	$0.681_{\pm.277}$	$6.690_{\pm.027}$
Token Embeddings (Baumann et al. (2025))	$0.035_{\pm.011}$	$0.024_{\pm.006}$	$0.267_{\pm.063}$	$0.211_{\pm.058}$	$6.776_{\pm.023}$
SCORE-SPACE STEERING					
FreeSliders (Ezra et al. (2025))	$0.079_{\pm.012}$	$0.049_{\pm.009}$	$\underline{0.055}_{\pm.007}$	$0.066_{\pm.012}$	$6.790_{\pm.017}$
WEIGHT-SPACE STEERING					
Concept Sliders (Gandikota et al. (2024))	$0.086_{\pm.010}$	$\underline{0.067}_{\pm.011}$	$0.056_{\pm.006}$	$\mathbf{0.045}_{\pm.006}$	$6.762_{\pm.019}$
ACTIVATION STEERING					
AUSteer (Feng et al. (2026))	$0.081_{\pm.020}$	$0.058_{\pm.013}$	$0.212_{\pm.069}$	$0.065_{\pm.011}$	$6.758_{\pm.014}$
CAA (Rimsky et al. (2024))	$0.075_{\pm.016}$	$0.046_{\pm.010}$	$0.092_{\pm.034}$	$0.071_{\pm.015}$	$6.787_{\pm.016}$
LOCALIZED STEERING (OURS)					
AUSteer **(loc.)**	$0.096_{\pm.016}$	$0.063_{\pm.012}$	$\mathbf{0.054}_{\pm.006}$	$\underline{0.050}_{\pm.005}$	$6.775_{\pm.018}$
CAA **(loc.)**	$\underline{0.104}_{\pm.019}$	$0.065_{\pm.011}$	$0.063_{\pm.018}$	$0.052_{\pm.010}$	$\underline{6.792}_{\pm.017}$
SAE **(loc.)**	$\mathbf{0.118}_{\pm.019}$	$\mathbf{0.070}_{\pm.013}$	$0.057_{\pm.007}$	$0.055_{\pm.004}$	$6.770_{\pm.020}$



(a) **Seamless Edit** (b) **Audio Quality** (c) **Smoothness**

Figure 4: **Human evaluation of audio steering methods.** We report mean Likert rating (1–5) for each method in the *standard* (all layers) and *localized* configurations across the three questions.

In smoothness, our methods distribute concept changes uniformly along the intervention scale rather than producing abrupt jumps. The localized activation steering closely matches Concept Sliders and FreeSliders at 0.05–0.07, three to twelve times lower than the prompt-level methods. Audio Quality is stable across the table, with a majority of aesthetic scores remaining within a narrow range, confirming that none of the methods, except Text Embeddings, cause severe perceptual degradation.

Human Evaluation. As the objective metrics only loosely correlate with perceived musical quality, we additionally ran an extensive listening study. We collected 1279 ratings from 32 listeners over 15 trials covering the *piano*, *female vocals*, and *tempo* concepts (see App. K). For each trial, we presented the seven steering methods (PCI, Text Emb., Token Emb., FreeSliders, AUSteer, CAA, SAE) in both the *standard* (all layers) and the *localized* configuration. We asked participants to rate clips on a 1–5 Likert scale along: *Seamless Edit* (was the concept introduced, and without breaking the original audio?), *Audio Quality*, and *Smoothness* (how gradual is the transition with α?).

Fig. 4 reports the mean rating per method. Methods that act on the cross-attention activations (CAA, SAE, AUSteer) consistently lead on *Seamless Edit*, with the localized CAA reaching the highest rating (3.32) followed closely by SAE (3.22) and localized AUSteer (3.24). Embedding-level methods, particularly Text Emb., trail across all three dimensions, mirroring their lower Audio Quality scores in Table 1. Differences between methods are smaller on *Audio Quality* and *Smoothness*, suggesting that listeners perceive most interventions as comparably clean once the concept is properly introduced.

5.4 DOES LOCALIZED STEERING ALWAYS BEAT GLOBAL STEERING?

Steering inside vs. outside the bottleneck. We test our bottleneck hypothesis, using three architecturally distinct backbones: AudioLDM2 (U-Net), Stable Audio Open (diffusion transformer), and ACE-Step (flow transformer), with CAA as a representative activation-steering method. We compare three intervention sites: *all layers*, the *functional layers only*, and an *ablation* setup that steers every layer except the functional ones ($57/64$ in AudioLDM2, $21/24$ in Stable Audio, and $22/24$ in Ace).

As visible in Fig. 5, across all concepts, localized steering achieves the highest alignment at matched preservation level, while the ablation curve flattens close to the unsteered baseline or leads to negative results (female vocal). This indicates that our identified subset acts as a true semantic bottleneck. The residual capacity outside it is insufficient to introduce the target concept on its own. We provide quantitative results for this experiment in App. L.

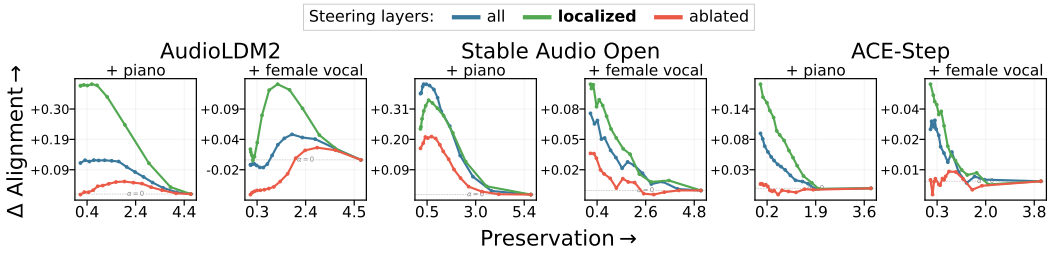


Figure 5: **Localized activation steering outperforms the global approach.** For the same audio preservation level, intervening in functional layers offers higher gain in the alignment. Additionally, omitting functional layers for steering often yields no conceptual gain.

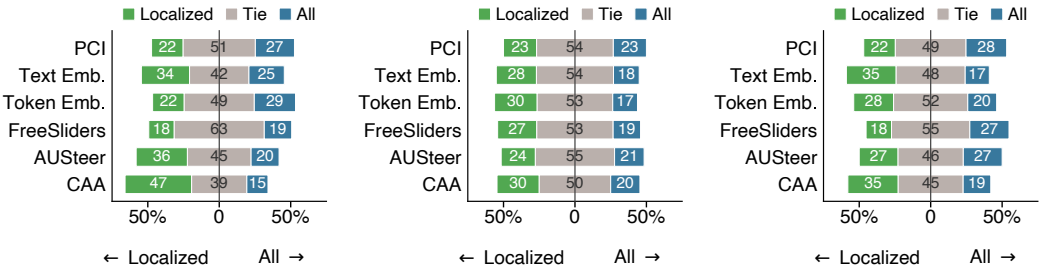
Localization across steering approaches. So far, we have established that localization is the right intervention site for activation steering. Table 2 extends this analysis to other steering paradigms. We report the relative gain (%) achieved by methods when their intervention is restricted to localized layers, with green and red shading marking improvements and regressions.

Localization is not a universal improvement but interacts strongly with the steering paradigm. Activation-space methods benefit substantially without degrading quality: CAA gains (MuQ/CLAP) 39%/40% in AUC and 32%/26% in Smoothness. Notably, constraining AUSteer to our layers significantly improves its Effectiveness (19%/8%) and Smoothness (74%/22%), indicating that its original discriminative localization mechanism prioritizes different layers than those identified by activation patching, potentially leading to suboptimal results. Conversely, weight-space steering degrades sharply with localization: Concept Sliders lose 21%/25% of AUC and 77%/34% on Smoothness. We attribute this behavior to Concept Sliders' methodology of adding new mechanisms to the existing model rather than leveraging its intrinsic capabilities.

Table 2: **Effect of layer localization on each steering method.** We compare the methods' global variant (all layers steering) against the localized one. Gain from localization is expressed as a % change of the mean: **improvement** or **degradation** of steering quality, faded when the absolute gain is below the standard error of the difference. We provide full results in App. M.1.

Method	**AUC** (↑)		**Smoothness** (↓)		**Audio Quality** (↑)
	MuQ	CLAP	MuQ	CLAP	
PCI (Görgün et al. (2026))	$0.084_{\pm.019}$	$0.049_{\pm.010}$	$0.185_{\pm.064}$	$0.223_{\pm.066}$	$6.803_{\pm.016}$
Δ (localization)	−19%	−22%	−50%	+21%	−0.2%
Text Embeddings (Ezra et al. (2025))	$0.014_{\pm.007}$	$0.011_{\pm.004}$	$0.439_{\pm.224}$	$0.681_{\pm.277}$	$6.690_{\pm.027}$
Δ (localization)	+62%	+31%	+25%	+48%	+1.4%
Token Embeddings (Baumann et al. (2025))	$0.035_{\pm.011}$	$0.024_{\pm.006}$	$0.267_{\pm.063}$	$0.211_{\pm.058}$	$6.776_{\pm.023}$
Δ (localization)	−8%	−20%	+18%	−33%	+0.3%
FreeSliders (Ezra et al. (2025))	$0.079_{\pm.012}$	$0.049_{\pm.009}$	$0.055_{\pm.007}$	$0.066_{\pm.012}$	$6.790_{\pm.017}$
Δ (localization)	+1%	−1%	−18%	−13%	−0.0%
Concept Sliders (Gandikota et al. (2024))	$0.086_{\pm.010}$	$0.067_{\pm.011}$	$0.056_{\pm.006}$	$0.045_{\pm.006}$	$6.762_{\pm.019}$
Δ (localization)	−21%	−25%	−77%	−34%	+0.2%
AUSteer (Feng et al. (2026))	$0.081_{\pm.020}$	$0.058_{\pm.013}$	$0.212_{\pm.069}$	$0.065_{\pm.011}$	$6.758_{\pm.014}$
Δ (localization)	+19%	+8%	+74%	+22%	+0.3%
CAA (Rimsky et al. (2024))	$0.075_{\pm.016}$	$0.046_{\pm.010}$	$0.092_{\pm.034}$	$0.071_{\pm.015}$	$6.787_{\pm.016}$
Δ (localization)	+39%	+40%	+32%	+26%	+0.1%

Our listening study (Fig. 6) corroborates these trends in the pairwise perceptual preference test. For *Seamless Edit*, restricting steering to the functional layers yields a clear advantage for the activation-space methods: localized CAA is preferred over the global one (47% versus 15%), as well as for AUSteer and TextEmb. In *Audio Quality*, the localized variant is preferred or tied for every method. On *Smoothness*, the two configurations are statistically close for most methods, with CAA again favoring the localized version.



(a) **Seamless Edit** (b) **Audio Quality** (c) **Smoothness**

Figure 6: **Human preference for localized vs. global setups of steering methods.** We compare user study ratings of both configurations for the same method. Bars present pairs in which the localized variant is preferred (green), the two are tied (grey), or the global steering is preferred (blue).

5.5 Multi-concept steering

In our final experiment, we investigate whether narrowing down the intervention scope to the semantic bottleneck can improve audio modulation when steering the model in multiple directions simultaneously. For example, such an intervention can add two new instruments to a song or manipulate vocal gender while decreasing tempo. In this setting, we compare three activation steering methods: CAA and AUSteer in *localized* and *global* configurations, as well as SAE. In all three cases, the combined vector is obtained by uniformly weighted summation of the single-direction steering vectors, optionally negated for concept-suppressing directions.

Results in Table 3 show that our localization extends to the multi-concept steering setting. Across 9 concept combinations (5 pairs and 4 triples), the three localized methods outperform global steering. This suggests that applying combined steering vectors at every layer accumulates collateral changes that mask the intended ones. Within the localized group, SAE is the most consistent performer (with $5/9$ wins), ahead of localized CAA (with $3/9$ wins). Even when including negated directions in vector summation, restricting the intervention to the semantic bottleneck preserves a clear steering semantics, whereas spreading it across the entire stack substantially weakens it.

Table 3: **Multi-concept steering performance measured with average Area Under LPAPS-MUQ Curve.** Best result per combination of concepts is **bolded**. Single column denotes steering on combinations of concepts, where **P** = Piano, **V** = Violin, **FV** = Female Vocal, **A** = Acoustic Guitar, **J** = Jazz Music, **M** = Brighter Mood, $\overline{\textbf{FV}}$ = Male Vocal, $\overline{\textbf{T}}$ = Slower Tempo, and $\overline{\textbf{M}}$ = Darker Mood. In App. O, we provide more details regarding the experiment.

Method	Steering Two Concepts					Steering Three Concepts			
	P + V	**P + FV**	**A + FV**	**P + $\overline{\text{FV}}$**	**J + $\overline{\text{T}}$**	**P + V + J**	**A + FV + M**	**P + V + $\overline{\text{T}}$**	**A + $\overline{\text{FV}}$ + $\overline{\text{M}}$**
AUSteer	+0.057	+0.001	+0.022	+0.013	+0.090	+0.062	+0.005	+0.048	+0.045
CAA	+0.048	+0.011	+0.028	+0.008	+0.068	+0.047	+0.026	+0.029	+0.032
AUSteer (loc.)	+0.099	+0.018	+0.029	**+0.039**	+0.086	+0.088	+0.021	+0.069	+0.051
CAA (loc.)	+0.094	**+0.034**	**+0.052**	+0.027	+0.103	+0.081	**+0.035**	+0.061	+0.039
SAE (loc.)	**+0.114**	+0.027	+0.041	+0.034	**+0.110**	**+0.100**	+0.018	**+0.081**	**+0.063**

6 Conclusions

In this work, we demonstrated that audio diffusion models exhibit a semantic bottleneck – a small subset of attention layers that govern high-level musical concepts. By identifying this bottleneck via activation patching, we systematically benchmarked activation steering against prompt-, score-, and weight-based baselines, evaluating all paradigms within our localized framework. Supported by both objective metrics and a comprehensive listening study, we showed that restricting activation steering (using AUSteer, Contrastive Activation Addition, and Sparse Autoencoders) to these functional layers outperforms both global baselines and other paradigms in terms of precision and fidelity. This layer-specific approach offers a robust method for fine-grained musical control, overcoming the limitations of standard text prompting.

AI USE STATEMENT

We employed Generative AI tools during the execution of this work, primarily to assist with developing experimental code, dataset preparation, and to support our qualitative analysis of the results. As part of the dataset creation process, as described in Section 4 and App. E, we use LLM to propose alternative phrasings of musical concepts and rule-based substitution rules. However, we manually verified the proposed rules before use.

During the writing process, we did not use generative AI to originate the research question or core idea, develop the mathematical formulation, or propose research hypotheses. We employed an LLM to improve the manuscript's clarity and readability and to refine its structure. All AI-refined text and code changes in this work have been manually verified by us. Thus, as authors, we take full responsibility for the final content of this work, including text, claims, code, artifacts, and visualizations produced with the aid of generative AI.

ETHICS STATEMENT

Our work includes a listening study with human volunteer participants (Appendix K), who were informed of the study's purpose before taking part. The study was conducted anonymously in the browser, and we did not collect any names, contact details, or other personal or sensitive data.

Our experiments use existing research datasets and publicly available pretrained text-to-music models, under their respective licenses. Notably, as we describe in App. B, fine-grained control over music generation can support legitimate creative work, but it can also be misused, e.g., to imitate the style of specific artists or copyrighted material. Our designed steering methods add no additional content-safety mechanisms, so their outputs inherit models' biases and failure modes.

REPRODUCIBILITY STATEMENT

The main part of our manuscript and its Appendix contain full details of our experiments, including a description of the methodology, the dataset creation process, the evaluation protocol, the shared inference configuration, and hyperparameter analyses. The details on the listening study are also documented in App. K.

To ensure reproducibility, all generations use fixed random seeds, and the same seeds are shared across methods and steering strengths, so every method is compared on identical initial noise. The code for localization, steering, and evaluation, together with configurations, prompts, and seeds, is provided in the supplementary material. Trained checkpoints (SAEs and Concept Slider LoRAs) and the computed steering vectors will be released upon publication. The pretrained models and the MusicCaps dataset must be obtained separately under their original terms.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Georgii Aparin, Tasnima Sadekova, Alexey Rukhovich, Assel Yermekova, Laida Kushnareva, Vadim Popov, Kristian Kuznetsov, and Irina Piontkovskaya. AudioSAE: Towards understanding of audio-processing models with sparse AutoEncoders. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3221–3254, Rabat, Morocco, mar 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. URL https://aclanthology.org/2026.eacl-long.149/.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech 2022*, pp. 2278–2282, 2022.

Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Vlad I Morariu, Nanxuan Zhao, Ryan A Rossi, Varun Manjunatha, and Soheil Feizi. On mechanistic knowledge localization in text-to-image generative models. In *Forty-first International Conference on Machine Learning*, 2024a.

Samyadeep Basu, Nanxuan Zhao, Vlad I. Morariu, Soheil Feizi, and Varun Manjunatha. Localizing and editing knowledge in text-to-image generative models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=Qmw9ne6SOQ.

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13231–13241, 2025.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237, 2018.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders, 2024. URL https://arxiv.org/abs/2412.06410.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kXwdL1cWOAi.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Bartosz Cywiński and Kamil Deja. SAeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=6N0GxaKdX9.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

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

Rotem Ezra, Hedi Zisling, Nimrod Berman, Ilan Naiman, Alexey Gorkor, Liran Nochumsohn, Eliya Nachmani, and Omri Azencot. Freesliders: Training-free, modality-agnostic concept sliders for fine-grained diffusion control in images, audio, and video. *arXiv preprint arXiv: 2511.00103*, 2025.

Simone Facchiano, Giorgio Strano, Donato Crisostomi, Irene Tallini, Tommaso Mencattini, Fabio Galasso, and Emanuele Rodolà. Activation patching for interpretable steering in music generation. *arXiv preprint arXiv:2504.04479*, 2025.

Zijian Feng, Tianjiao Li, Zixiao Zhu, Hanzhang Zhou, Junlang Qian, Li Zhang, Chua Jia Jim Deryl, Mak Lee Onn, Gee Wah Ng, and Kezhi Mao. Fine-grained activation steering: Steering less, achieving more. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=guSVafqhrB.

Jaden Fiotto-Kaufman, Alexander Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, et al. Nnsight and ndif: Democratizing access to open-weight foundation model internals. In *International Conference on Learning Representations*, volume 2025, pp. 92337–92370, 2025.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. CASteer: Cross-attention steering for controllable concept erasure. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6D5Odqol1B.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. Ace-step: A step towards music generation foundation model. *arXiv preprint arXiv: 2506.00045*, 2025.

Junmin Gong, Yulin Song, Wenxiao Zhao, Sen Wang, Shengyuan Xu, Jing Guo, and Xuerui Yang. Ace-step 1.5: Pushing the boundaries of open-source music generation. *arXiv preprint arXiv:2602.00744*, 2026.

Ada Görgün, Fawaz Sammani, Nikos Deligiannis, Bernt Schiele, and Jonas Fischer. Temporal concept dynamics in diffusion models via prompt-conditioned interventions. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ABjaSsrYPD.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Vladimir E. Iashin and Esa Rahtu. Taming visually guided sound generation. *British Machine Vision Conference*, 2021. doi: 10.5244/c.35.336.

Yuhang Jia, Yang Chen, Jinghua Zhao, Shiwan Zhao, Wenjia Zeng, Yong Chen, and Yong Qin. Audioeditor: A training-free diffusion-based audio editing framework. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887821.

Junghyun Koo, Gordon Wichern, François G Germain, Sameer Khurana, and Jonathan Le Roux. Smitin: Self-monitored inference-time intervention for generative music transformers. *IEEE Open Journal of Signal Processing*, 2025.

Ching Ho Lee, Javier Nistal, Stefan Lattner, Marco Pasini, and George Fazekas. Diffusion timbre transfer via mutual information guided inpainting. *arXiv preprint arXiv:2601.01294*, 2026.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

Jinhua Liang, Yi Yuan, Dongya Jia, Xiaobin Zhuang, Zhengxi Liu, Yuanzhe Chen, Zhuo Chen, Yuping Wang, and Yuxuan Wang. Audiomorphix: Training-free audio editing with diffusion probabilistic models, 2024. URL https://openreview.net/forum?id=a8dQutiF9E.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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*, 2024.

Huadai Liu, Jialei Wang, Xiangtai Li, Wen Wang, Qian Chen, Rongjie Huang, Yang Liu, Jiayang Xu, Zhou Zhao, and Wei Xue. Melodyedit: Zero-shot music editing with disentangled inversion control. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 10083–10092, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755377. URL https://doi.org/10.1145/3746027.3755377.

Zihan Liu, Shuangrui Ding, Zhixiong Zhang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Songgen: A single stage auto-regressive transformer for text-to-song generation. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=aTyQjpsaB4.

Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using DDPM inversion. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34603–34629. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/manor24a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Xinlei Niu, Kin Wai Cheuk, Jing Zhang, Naoki Murata, Chieh-Hsin Lai, Michele Mancusi, Woosung Choi, Giorgio Fabbro, Wei-Hsiang Liao, Charles Patrick Martin, and Yuki Mitsufuji. Steermusic: enhanced musical consistency for zero-shot text-guided and personalized music editing. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i3.37181. URL https://doi.org/10.1609/aaai.v40i3.37181.

Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37:3311–3325, 1997a. URL https://api.semanticscholar.org/CorpusID:14208692.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997b.

Nathan Paek, Yongyi Zang, Qihui Yang, and Randal Leistikow. Learning interpretable features in audio latent spaces via sparse autoencoders. *arXiv preprint arXiv:2510.23802*, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=UGpGkLzwpP.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pp. 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

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

Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. Mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, volume 10, pp. 2014, 2014.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, marco cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=l2zFn6TIQi.

Pau Rodriguez, Michal Klein, Eleonora Gualdoni, Valentino Maiorca, Arno Blaas, Luca Zappella, marco cuturi, and Xavier Suau. LinEAS: End-to-end learning of activation steering with a distributional loss. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=EBONa3tT3K.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Konstantine Sadov. Feature discovery in audio models a whisper case study, 2024. URL https://builders.mozilla.org/insider-whisper/.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. Sam audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*, 2025.

Nikhil Singh, Manuel Cherep, and Patricia Maes. Discovering and steering interpretable concepts in large generative music models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mGtEoLYr9j.

Łukasz Staniszewski, Bartosz Cywiński, Franziska Boenisch, Kamil Deja, and Adam Dziedzic. Precise parameter localization for textual generation in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gdHtZlaaSo.

Suno, Inc. Suno: AI music generator. https://suno.com/, 2026. Accessed: 2026-05-06.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-step is enough: Sparse autoencoders for text-to-image diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=MBJJ9Wcpg9.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv: 2502.05139*, 2025. URL https://arxiv.org/abs/2502.05139.

Marcel A Vélez Vásquez, Charlotte Pouw, John Ashley Burgoyne, and Willem H Zuidema. Exploring the inner mechanisms of large generative music models. In *ISMIR*, pp. 791–798, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Megan Wei, Michael Freeman, Chris Donahue, and Chen Sun. Do music generation models encode music theory? *International Society for Music Information Retrieval Conference*, 2024. doi: 10.48550/arXiv.2410.00872.

Yusong Wu, K. Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and S. Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. doi: 10.1109/ICASSP49357.2023.10095969.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=N8Oj1XhtYZ.

Manjie Xu, Chenxing Li, Duzhen Zhang, Dan Su, Wei Liang, and Dong Yu. Prompt-guided precise audio editing with diffusion models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=kQ1dwuheR0.

Yi Yang, Haowen Li, Tianxiang Li, Boyu Cao, Xiaohan Zhang, Liqun Chen, and Qi Liu. Melodia: Training-free music editing guided by attention probing in diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(3):2209–2217, Mar. 2026. doi: 10.1609/aaai.v40i3.37204. URL https://ojs.aaai.org/index.php/AAAI/article/view/37204.

Arman Zarei, Samyadeep Basu, Keivan Rezaei, Zihao Lin, Sayan Nag, and Soheil Feizi. Localizing knowledge in diffusion transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=SiBVbL7rsX.

Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Musicmagus: Zero-shot text-to-music editing via diffusion models. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 7805–7813. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/864. URL https://doi.org/10.24963/ijcai.2024/864.

Daniel Zhao, Daniel Beaglehole, Julian McAuley, Taylor Berg-Kirkpatrick, and Zachary Novack. Steering autoregressive music generation with recursive feature machines. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NaHzPMaCY9.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. Muq: Self-supervised music representation learning with mel residual vector quantization. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3653–3664, 2025. doi: 10.1109/TASLPRO.2025.3602320.

Appendix

TABLE OF CONTENTS

A Limitations **17**

B Broader impact **17**

C Future work **17**

D Compute resources **18**

E Dataset construction for layer localization **18**

F Layer localization **20**

F.1 Per-concept impact maps 20

F.2 Parameter analysis 22

F.3 Computational cost 22

G Prompts: steering vectors construction and evaluation **23**

G.1 Contrastive prompts for steering vectors and SAE features 23

G.2 Benchmark prompts 23

H Evaluation metrics **24**

H.1 Audio-text alignment prompts 24

H.2 Alternative concept alignment measures 24

H.3 Alternative audio preservation metrics 25

H.4 Generation quality over steering range 27

I Steering methods **28**

I.1 Descriptions 28

I.2 Implementation and hyperparameter details 30

J Studied audio diffusion architectures **35**

K Details on user study **36**

L Additional results for localized activation steering **39**

L.1 Localized activation steering 39

L.2 Quantitative methods' configurations comparison 41

M Audio concept steering benchmark **42**

M.1 Methods comparison across concepts 42

M.2 Alignment–preservation curves 45

N Spectrograms **48**

O Details on multi-concept steering **49**

A LIMITATIONS

Architectural scope. Our localization study and findings about the existence of "semantic bottleneck" are limited to audio diffusion architectures: AudioLDM2 (Liu et al., 2024), Stable Audio Open (Evans et al., 2025), and Ace-Step (Gong et al., 2025). We believe a similar study could be conducted on Music-LLMs, such as MusicGen. The evaluation protocol we propose can be easily applied to benchmark audio modulation with autoregressive text-to-music models. We further note that our comparison is conducted on a single model, Ace-Step. However, in Fig. 5 and App. L, we show that our central finding – restricting the intervention to the semantic bottleneck drives effective steering – generalizes to all three models.

Intervention site. We localize functional components at the level of cross-attention layers. The rationale is that those modules are the only ones that incorporate text conditioning into generated audio. We do not study activations in residual streams, self-attention, or feed-forward layers and leave this for follow-up work. We believe that the proposed methodology for localization and steering can be easily adapted to other layer types.

Concept coverage. Our steering benchmark considers nine directions: piano, violin, acoustic/electric guitar, mood, tempo, vocal gender, vocal style, jazz/rock, and classical/electronic genres. The localization analysis covers additional instruments (drums, flute, maracas, trumpet), and genres (jazz, techno, reggae). Together, they span the dimensions commonly used in music-generation evaluation, but remain a finite slice of musical structure. We emphasize that our choice of concept depends on the generative model's intrinsic capabilities, as it is impossible to perform text-based localization of concepts that the model cannot generate when prompted.

Evaluation metrics. Across steering paradigms, we calibrate the maximum steering strength using an LPAPS preservation cutoff. Hence, the reported AUC values inherit the choice of LPAPS as a perceptual distance measurement. In App. H.3, we show that our findings generalize to alternative preservation metrics. Additionally, to measure the presence of target concepts in generated audio, we employ CLAP (Wu et al., 2022) and MuQ (Zhu et al., 2025). We emphasize that such models may exhibit their own biases and rate certain features more confidently than others. To ensure robust evaluation, during experiments, we report both (CLAP, MuQ) metrics values, include a user study, and, in the App. H, show that our evaluation protocol is aligned with external music evaluators, such as tempo detectors or vocal gender classifiers. We note, however, that inference-time steering applied with excessive strength can push activations off the training manifold, degrading generation quality. In App. H.4 we analyze this effect and show that, within the calibrated steering range, quality remains stable, with localized steering avoiding the monotone degradation exhibited by global interventions.

B BROADER IMPACT

Our work targets a long-standing limitation of audio generators that rely on prompting. Localized activation steering provides a continuous control surface without changing the underlying composition, which is the kind of fine-grained adjustment musicians or composers may need in practice.

We demonstrate that semantic musical attributes are concentrated in a small, shared subset of layers. This advances scientific understanding of audio diffusion models, which until now have largely been treated as black boxes, providing a foundation for downstream interpretability work.

The capability our work adds is fine-grained *control* over existing models, not new generative capacity. The standard concerns associated with text-to-music systems, such as training-data copyright, may apply to the underlying generators we analyze, but are not amplified by our approach.

C FUTURE WORK

We discover semantic bottlenecks independently in three text-to-music models, showing that a narrow, contiguous band of functional layers recurs across diverse architectures. While we observe this structural regularity, we do not study whether layer indices can be transferred across architectures. An open question is whether the bottlenecks of different models are representationally compatible. A potential future direction is to study the stitching (Lenc & Vedaldi, 2015) of the semantic bottlenecks across audio DMs to test the transfer of steering vectors or layers.

D COMPUTE RESOURCES

For the experiments, we used a scientific cluster consisting of 110 nodes with CrayOS operating system. Each node is powered by 288 CPU cores, stemming from 4 NVIDIA Grace processors, each with 72 cores and a clock speed of 3.1 GHz. The nodes are equipped with substantial memory, featuring 480 GB of RAM per node. For GPU acceleration, each node in the cluster consists of 4 NVIDIA GH200 96GB GPUs with 120 GB of RAM and 72 CPUs per GPU.

E DATASET CONSTRUCTION FOR LAYER LOCALIZATION

Table 4 presents examples of counterfactual prompt pairs used in our localization experiments. Each pair consists of an original prompt $\mathcal{P}_c$ from the MusicCaps dataset and a modified prompt $\mathcal{P}_{\tilde{c}}$ where concept-associated terms are replaced with their counterparts. Table 5 lists the keywords used to filter MusicCaps captions for each concept and the corresponding replacement terms used to generate counterfactual prompts.

Table 4: **Examples of counterfactual prompt pairs.** For each concept category, we show an original prompt from MusicCaps and its counterfactual version with the target concept replaced. Modified terms are highlighted in **bold**.

Concept	**Original Prompt $\mathcal{P}_c$**	**Counterfactual Prompt $\mathcal{P}_{\tilde{c}}$**
Vocal Gender	*"The low quality recording features a ballad song that contains sustained strings, mellow piano melody, and soft **female** vocal singing over it. It sounds sad and soulful."*	*"The low quality recording features a ballad song that contains sustained strings, mellow piano melody, and soft **male** vocal singing over it. It sounds sad and soulful."*
Tempo	*"This is a country music piece. There is a fiddle playing the main melody. The acoustic guitar and electric guitar are playing gently. The song has a **slow** tempo. The atmosphere is sentimental."*	*"This is a country music piece. There is a fiddle playing the main melody. The acoustic guitar and electric guitar are playing gently. The song has a **fast** tempo. The atmosphere is sentimental."*
Mood	*"This is a Hindustani classical music piece. There is a harmonium playing the main tune. A bansuri joins in to play, supporting a melody. The rhythmic background consists of tabla percussion and electronic drums. The atmosphere is **joyful**."*	*"This is a Hindustani classical music piece. There is a harmonium playing the main tune. A bansuri joins in to play, supporting a melody. The rhythmic background consists of tabla percussion and electronic drums. The atmosphere is **sorrowful**."*
Instrument	*"This folk song features a male voice singing the main melody in an emotional mood. This is accompanied by an accordion playing fills in the background. A **violin** plays a droning melody."*	*"This folk song features a male voice singing the main melody in an emotional mood. This is accompanied by an accordion playing fills in the background. A **trumpet** plays a droning melody."*
Genre	*"The low quality recording features a **reggae**/dub song that consists of a flat male vocal singing over punchy 808 bass, punchy snare, shimmering hi hats and groovy piano chords. It sounds energetic, groovy and the recording is noisy and in mono."*	*"The low quality recording features a **metal**/dub song that consists of a flat male vocal singing over punchy 808 bass, punchy snare, shimmering hi hats and groovy piano chords. It sounds energetic, groovy and the recording is noisy and in mono."*

Table 5: **Keywords for dataset construction.** For each concept, we list the examples of keywords proposed by an LLM (GPT-4o) to filter captions from MusicCaps (for a given concept, we select prompts containing these terms) and the replacement keywords used to generate counterfactual prompts.

Concept	Example Filter Keywords	Example Replacements
female	female, woman, girlish	female → male woman → man girlish → boyish
male	male, man	man → woman his → her himself → herself
slow	slow, slowed, slower, slowly, slow-paced, leisurely	slow → fast slower → faster slow-paced → fast-paced
fast	fast, rapid, dynamic, quick, energetic, groovy, lively	fast → slow quickly → slowly lively → slowly
happy	happy, joyful, upbeat, uplifting, cheerful	joyful → sorrowful cheerful → gloomy warm → cold
sad	sad, depressing, dejected	sad → happy lonely → connected emotional → lighthearted
drums	drums, drum, percussion	drums → saxophone percussionist → saxophonist
flute	flute, flutes	flute → maracas flutist → maracas player
maracas	maracas, maraca	maracas → flute maraca → flute
trumpet	trumpet, trumpets	trumpet → violin trumpeter → violinist
violin	violin, fiddle	violin → trumpet string → $\emptyset$
jazz	jazz, blues	jazz → opera band → choir
reggae	reggae, jamaican, tropical, beach	reggae → metal chill → heavy beach → concert

F LAYER LOCALIZATION

In this section, we describe our localization experiment and its results in detail. In F.1, we provide full results for all three architectures. In F.2, we analyze how the choice of prompts and the selection threshold τ affect the results. Finally, in F.3, we quantify the computational cost of the localization experiment.

F.1 PER-CONCEPT IMPACT MAPS

Figs. 7 to 9 present the impact scores $\mathrm{I}(l, c)$ (see Eq. (2)) of every cross-attention layer on probed concepts across all three architectures: AudioLDM2, Stable Audio Open, and Ace-Step. In the figures, we provide the aggregated (across concepts) impact scores $\mathrm{I}(l) = \frac{1}{|C|}\sum_c$, and highlight (orange) the layers that are selected to carry the control, i.e., the layers that score above the selection threshold τ.

Score maps make the *semantic bottleneck* visible per concept. In AudioLDM2 (Fig. 7), control over vocal gender, tempo, mood, instruments and genres concentrates in layers 45–51 (out of 64), in Stable Audio Open (Fig. 8) in layers 12–14 (out of 24), and in Ace-Step (Fig. 9) in layers 7–8 (out of 24). The aggregate row is strongly peaked across all architectures. In AudioLDM2: $\mathrm{I}(\texttt{attn51}) = 0.36$, $\mathrm{I}(\texttt{attn45}) = 0.33$, $\mathrm{I}(\texttt{attn46}) = 0.15$, while other score below 0.09. In Stable Audio Open: $\mathrm{I}(\texttt{attn13}) = 0.51$, $\mathrm{I}(\texttt{attn12}) = 0.22$, and $\mathrm{I}(\texttt{attn14}) = 0.12$, while other score below 0.10. In Ace-Step: $\mathrm{I}(\texttt{attn8}) = 0.49$, $\mathrm{I}(\texttt{attn7}) = 0.29$, and $\mathrm{I}(\texttt{attn6}) = 0.09$, while other score below 0.04.

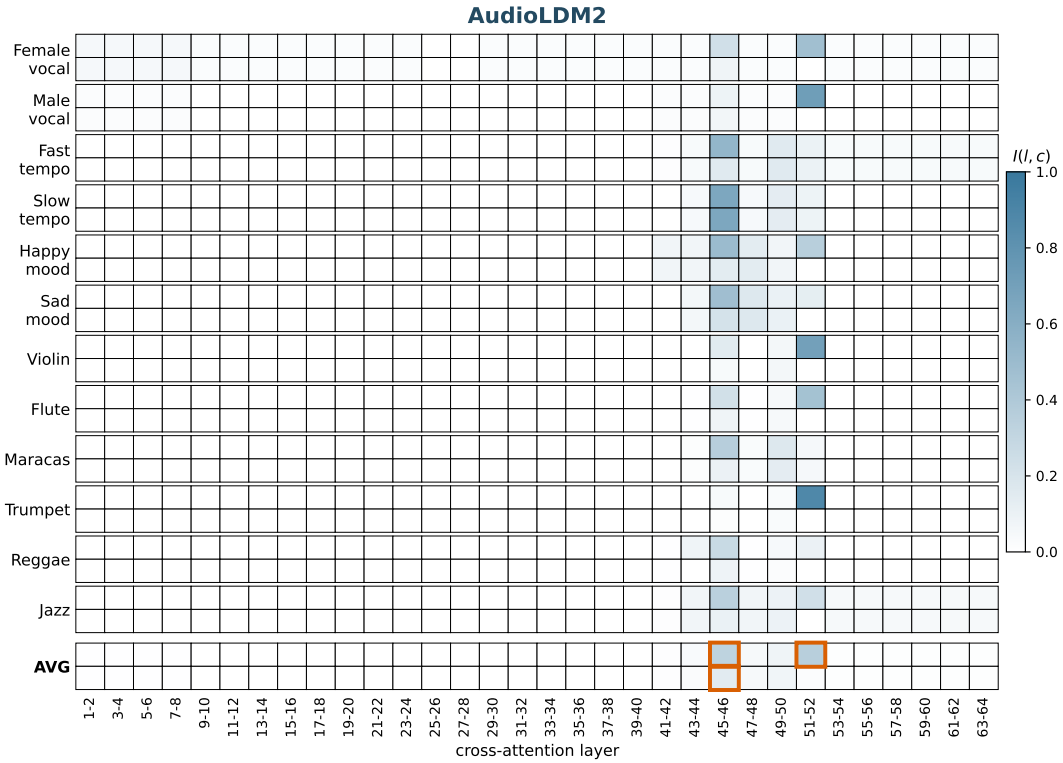


Figure 7: **Impact of every cross-attention layer on every concept in AudioLDM2 (Liu et al., 2024).** Each cell color intensity denotes layer l impact on concept c as $\mathrm{I}(l, c)$ (Eq. (2)). Last row denotes the aggregated impact $\mathrm{I}(l)$ of layer l averaged over all concepts. The final selected layers (where $\mathrm{I}(l) \geq \tau = 0.10$) are outlined in orange.

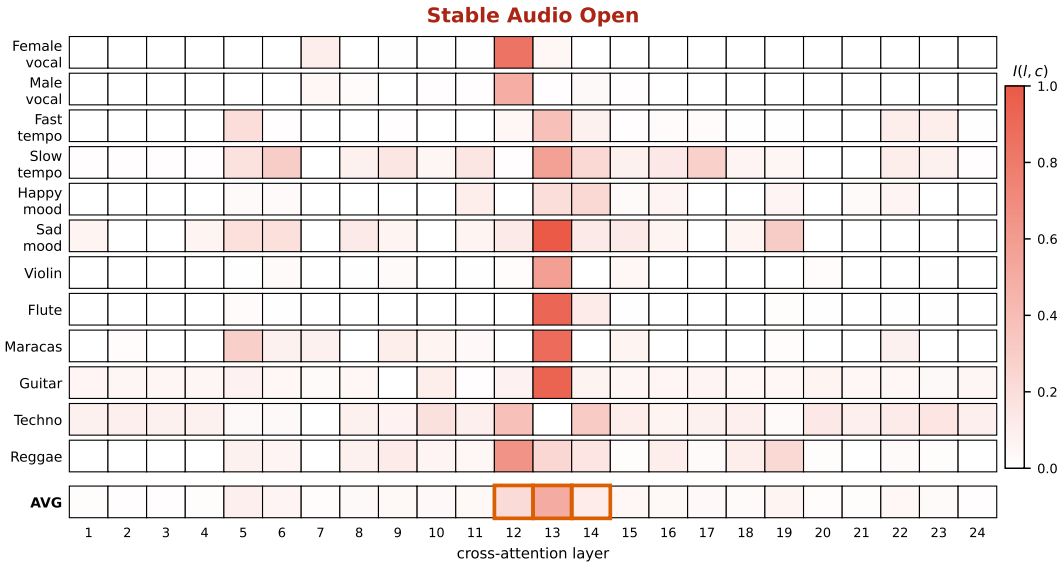


Figure 8: **Impact of every cross-attention layer on every concept in Stable Audio Open (Evans et al., 2025).** Each cell color intensity denotes layer l impact on concept c as $\mathrm{I}(l, c)$ (Eq. (2)). Last row denotes the aggregated impact $\mathrm{I}(l)$ of layer l averaged over all concepts. The final selected layers (where $\mathrm{I}(l) \geq \tau = 0.10$) are outlined in orange.

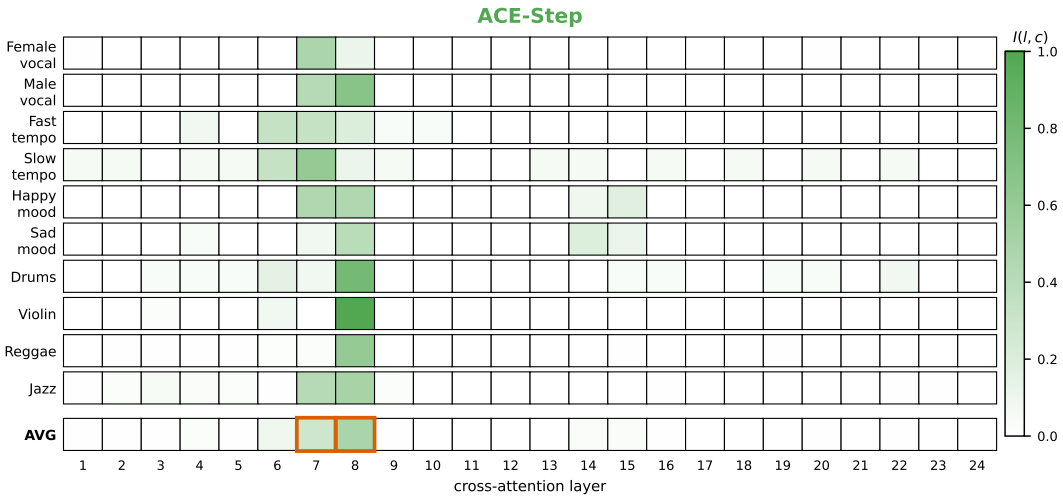


Figure 9: **Impact of every cross-attention layer on every concept in Ace-Step (Gong et al., 2025).** Each cell color intensity denotes layer l impact on concept c as $\mathrm{I}(l, c)$ (Eq. (2)). Last row denotes the aggregated impact $\mathrm{I}(l)$ of layer l averaged over all concepts. The final selected layers (where $\mathrm{I}(l) \geq \tau = 0.10$) are outlined in orange.

F.2 PARAMETER ANALYSIS

Prompt count. In our experiments, to ensure robust results, we run the localization experiment on up to 256 counterfactual prompt pairs per concept (230 on average).

In Table 6, we show that exact localization for ACE-Step is achievable with far fewer prompts. We report the Jaccard similarity between the sets of layers localized using the entire (full-budget) tracing dataset and a random subset of it. As shown, ACE's semantic bottleneck $\{7, 8\}$ can be recovered exactly using only 34 prompt pairs per concept.

Table 6: **Impact of number of prompts on ACE-Step localization results.** J(top-j) denotes Jaccard similarity between sets of j highest scoring layers selected by full-budget localization and a random subset of prompts. Results are averaged across 200 random subsamples.

Prompts / concept	**Budget %**	J**(top-1)**	J**(top-2)**	J**(top-3)**	J**(top-5)**
4	2%	0.54	0.82	0.66	0.45
12	5%	0.65	0.98	0.81	0.56
17	8%	0.67	0.99	0.90	0.63
23	10%	0.60	0.99	0.92	0.67
34	15%	0.65	**1.00**	0.97	0.74
46	20%	0.69	**1.00**	0.99	0.80
69	30%	0.72	**1.00**	**1.00**	0.87
115	50%	0.89	**1.00**	**1.00**	0.98
230	100%	**1.00**	**1.00**	**1.00**	**1.00**

Threshold sensitivity. In Table 7, we analyze the impact of the selection threshold τ on localization results. We show that Ace-Step selection $\{7, 8\}$ is unchanged over $\tau \in [0.10, 0.20]$, and AudioLDM2 keeps layers 45 and 51 over the whole range. In our experiments, we set $\tau = 0.10$.

Table 7: **Selected semantic bottleneck in Audio Diffusion Models as a function of selection threshold** τ. Setting used in our experiments is **bolded**.

τ	**AudioLDM2**	**Stable Audio Open**	**Ace-Step**
0.05	45, 46, 49, 50, 51	5, 6, 12, 13, 14, 19	6, 7, 8
0.075	45, 46, 49, 50, 51	5, 12, 13, 14	6, 7, 8
0.10	**45, 46, 51**	**12, 13, 14**	**7, 8**
0.15	45, 51	12, 13	7, 8
0.20	45, 51	12, 13	7, 8
0.30	45, 51	13	8

F.3 COMPUTATIONAL COST

Localizing C concepts in a model with L cross-attention layers costs

$$N_{\text{gen}} = C \times (L + 2) \times P \times S \tag{3}$$

generations, where P is the number of counterfactual prompt pairs per concept, and S is the number of noise seeds per prompt (in our experiments, we set $S = 8$). For Ace-Step, where $L = 24$, $C = 10$, and $P = 34$, localization involves 70.720 generations. On the NVIDIA GH200 GPU, where, on average (with batch size 32), a single generation takes 1.68 GPU seconds, the entire localization process takes roughly 33 GPU hours.

G PROMPTS: STEERING VECTORS CONSTRUCTION AND EVALUATION

This section details the prompts used in the steering experiments.

G.1 CONTRASTIVE PROMPTS FOR STEERING VECTORS AND SAE FEATURES

To compute steering vectors and identify concept-specific SAE features, we generate audio from contrastive prompt pairs. For each concept, we construct positive prompts $\mathcal{P}_c$ containing the target attribute and negative prompts $\mathcal{P}_{\tilde{c}}$ with the contrasting attribute. Table 8 shows the prompt templates used for each concept.

Table 8: **Contrastive prompt templates for steering.** The {base} placeholder is filled with diverse musical descriptions (e.g., "a song", "a jazz piece", "electronic music").

Concept	**Positive Prompt $\mathcal{P}_c$**	**Negative Prompt $\mathcal{P}_{\tilde{c}}$**
Piano	"{base}, with piano"	"{base}"
Violin	"{base}, with violin"	"{base}"
Guitar Type	"{base}, with acoustic guitar"	"{base}, with electric guitar"
Mood	"happy song, {base}"	"sad song, {base}"
Tempo	"fast song, {base}"	"slow song, {base}"
Vocal Gender	"{base}, with female vocal"	"{base}, with male vocal"
Vocal Style	"{base}, with rap vocal"	"{base}, with sing vocal"
Classical Genre	"classical song, {base}"	"electronic song, {base}"
Jazz Genre	"jazz song, {base}"	"rock song, {base}"

The base prompts span 50 diverse musical styles and genres, i.a.: "a song", "a melody", "music", "a tune", "a track", "instrumental music", "a pop song", "a rock song", "a jazz piece", "a classical piece", "electronic music", "acoustic music", "orchestral music", "hip hop music", "country music", "blues music", "folk music", "reggae music", "ambient music", "lofi music", "a ballad", "a love song", "energetic music", "calm music", "dramatic music", among others.

G.2 BENCHMARK PROMPTS

For steering evaluation, we use 100 diverse prompts (separate from those used to calculate steering vectors) that span a wide range of musical styles. Here, we provide examples:

- **Pop & Rock**: 'Upbeat indie pop track with jangly guitars and handclaps', '90s grunge with fuzzy guitars, angst-filled dynamics'
- **Electronic**: 'Dark synthwave anthem with pulsing bass, retro analog synths', 'Techno industrial with distorted kicks, metallic textures'
- **Jazz & Blues**: 'Melancholic jazz ballad with smooth saxophone, walking bassline', 'Delta blues with slide guitar, stomping rhythm'
- **World Music**: 'Traditional Irish jig with fiddle, tin whistle, bodhran drums', 'Afrobeat groove with polyrhythmic percussion, brass stabs'
- **Classical**: 'Romantic piano nocturne with expressive dynamics', 'Epic orchestral score with full brass, timpani rolls'
- **Hip Hop & R&B**: 'Boom bap with punchy drums, scratched samples', 'Neo-soul with warm keys, silky bassline'
- **Country & Folk**: 'Acoustic folk song with fingerpicked guitar, gentle harmonies', 'Bluegrass breakdown with banjo rolls, fiddle solo'

H EVALUATION METRICS

H.1 AUDIO-TEXT ALIGNMENT PROMPTS

To measure how the generated audio exhibits target concepts, we compute audio-text similarity between steered generations and concept-specific text queries using CLAP (Wu et al., 2022) and MuQ (Zhu et al., 2025) models. In Table 9, we present the evaluation prompts used for computing the similarity scores.

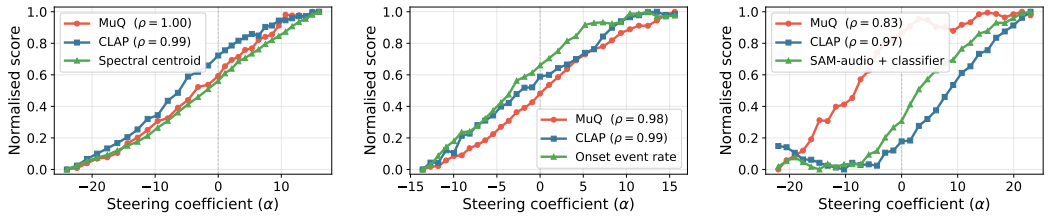
Table 9: **Audio-text alignment queries for audio-text alignment measurement.** These prompts are used to compute similarity scores between generated audio and target concepts.

Concept	Alignment query
Piano	"a piano song"
Violin	"a song with violin"
Guitar Type	"a song with acoustic guitar"
Mood	"a cheerful track"
Tempo	"a fast track"
Vocal Gender	"female vocal singing"
Vocal Style	"rap vocal"
Classical Genre	"a classical song"
Jazz Genre	"a jazz song"

H.2 ALTERNATIVE CONCEPT ALIGNMENT MEASURES

To assess whether text-audio alignment models properly score concepts in generated audio, we compare them with external evaluators that do not rely on contrastive text–audio alignment. We re-evaluate the SAE-steered generations across the three concepts (*mood*, *tempo*, *vocal gender*) using measurements from prior works.

For *mood*, inspired by Facchiano et al. (2025), we compute the **Spectral Centroid**, the magnitude-weighted mean frequency of the spectrum, which serves as a standard proxy for spectral brightness. Brighter timbres tend to correlate with more cheerful content and darker timbres with sadder content. For *tempo*, we follow Zhao et al. (2026) and compute the **onset event rate** (events/s), defined as the number of `librosa` onset peaks divided by clip duration, which captures rhythmic density. Finally, for *vocal gender*, we first **extract a vocal stem** from each song with `SAM Audio` (Shi et al., 2025), conditioned on the text prompt "singing". We then run an off-the-shelf **gender classifier**, a version of `Wav2Vec2-XLS-R` (Babu et al., 2022) fine-tuned for speaker gender classification, and report $P(\text{female} \mid \text{stem})$ averaged over the generated clips at each α.



(a) **Mood** (darker→brigher) (b) **Tempo** (slow→fast) (c) **Vocal gender** (male→female)

Figure 10: **Text-audio alignment models track the presence of concepts in generated audios.** We compare alignment metrics from our evaluation benchmark with external acoustic evaluators on *mood*, *tempo*, and *vocal gender* concepts. We plot the per-α averages of MuQ, CLAP, and the external evaluator, each normalized to $[0, 1]$. As shown, CLAP and MuQ track real audio properties. In legend, we report the Pearson correlation ρ between each text-based metric and the evaluator.

As presented in Fig. 10, across all three concepts, the text-alignment models score similarly to the independent acoustic descriptors. Quantitatively, in legend, the Pearson correlation (ρ) is very high for *mood* ($\rho_{\text{MuQ}} = 0.998$, $\rho_{\text{CLAP}} = 0.987$ vs. spectral centroid) and for *tempo* ($\rho_{\text{MuQ}} = 0.982$, $\rho_{\text{CLAP}} = 0.986$ vs. onset event rate). For *vocal gender*, CLAP tracks the SAM-Audio + classifier signal almost as tightly as it does for the other two concepts ($\rho_{\text{CLAP}} = 0.965$), while MuQ shows a weaker but still strong correlation ($\rho_{\text{MuQ}} = 0.830$). Overall, these correlations support the view that CLAP and MuQ are faithful evaluators of the target concepts across the full α range.

H.3 ALTERNATIVE AUDIO PRESERVATION METRICS

Our evaluation methodology heavily depends on the choice of preservation metric, which, in our main experiments, is LPAPS. However, as shown in Table 10, our localized activation steering achieves the best steering performance across other preservation metrics.

We employ measurements from the Musical Information Retrieval (MIR) domain. We alter our evaluation protocol (Section 5.2), replacing only the preservation axis with four decomposed measures, each expressed as a distance metric ($f_{\text{dist}} = 1 - f_{\text{sim}}$). We assess similarity (f_{sim}) between two audio tracks in terms of their:

- **Harmony** – the mean cosine similarity between corresponding frames of their `librosa`-computed CQT chromagrams.
- **Rhythm** – the beat F-measure calculated with `mir_eval` (Raffel et al., 2014) between their `librosa`-tracked beat sequences. Here, we exclude the *tempo* concept.
- **Melody** – the raw pitch accuracy calculated with `mir_eval` (Raffel et al., 2014) between their pYIN-estimated fundamental frequencies (f_0 contours).
- **Structure** – the Pearson correlation, clamped to $[0, 1]$, between self-similarity matrices of their `librosa`-computed CQT chromagrams.

We emphasize that these metrics do not rely on trained models and therefore share neither training data nor perceptual priors with LPAPS. As presented, localized activation steering offers the best steering performance under 7 out of 8 additional metrics. In Fig. 11 we compare methods under the new preservation metrics across concepts.

Table 10: **Localized steering offers best steering performance across diverse preservation metrics.** We report AUC-MuQ and AUC-CLAP under LPAPS and four MIR preservation axes, measuring harmony, rhythm, melody, and structure. Results are averaged over the nine benchmark concepts (except *tempo* on the Rhythm axis) and across both steering directions. **Best** and <u>second</u> results are highlighted.

Method	LPAPS (↑)		Harmony (↑)		Rhythm (↑)		Melody (↑)		Structure (↑)	
	MuQ	CLAP	MuQ	CLAP	MuQ	CLAP	MuQ	CLAP	MuQ	CLAP
PCI	0.021	0.012	0.026	0.015	0.020	0.013	0.015	0.009	0.020	0.012
PCI (loc.)	0.017	0.009	0.022	0.012	0.015	0.009	0.011	0.006	0.015	0.009
Text Embeddings	0.004	0.003	0.005	0.003	0.003	0.003	0.002	0.002	0.003	0.003
Text Embeddings (loc.)	0.006	0.004	0.007	0.004	0.004	0.003	0.004	0.002	0.005	0.003
Token Embeddings	0.009	0.006	0.010	0.007	0.008	0.006	0.006	0.005	0.008	0.006
Token Embeddings (loc.)	0.008	0.005	0.010	0.006	0.007	0.004	0.005	0.004	0.007	0.005
FreeSliders	0.020	0.012	0.023	0.015	0.017	0.011	0.015	0.010	0.019	0.012
FreeSliders (loc.)	0.020	0.012	0.024	0.015	0.018	0.012	0.015	0.010	0.019	0.012
Concept Sliders	0.022	<u>0.017</u>	0.023	0.018	0.019	<u>0.016</u>	0.016	<u>0.013</u>	0.021	**0.017**
Concept Sliders (loc.)	0.017	0.013	0.019	0.014	0.015	0.012	0.013	0.010	0.017	0.012
AUSteer	0.020	0.015	0.022	0.016	0.018	0.014	0.015	0.011	0.019	0.014
CAA	0.019	0.012	0.022	0.014	0.015	0.011	0.014	0.009	0.018	0.011
AUSteer (loc.)	0.024	0.016	0.027	0.017	0.022	0.016	0.018	0.012	0.023	0.015
CAA (loc.)	<u>0.026</u>	0.016	<u>0.031</u>	**0.019**	<u>0.023</u>	0.016	<u>0.019</u>	0.013	<u>0.025</u>	0.016
SAE (loc.)	**0.030**	**0.018**	**0.032**	<u>0.018</u>	**0.028**	**0.018**	**0.022**	**0.013**	**0.028**	<u>0.016</u>

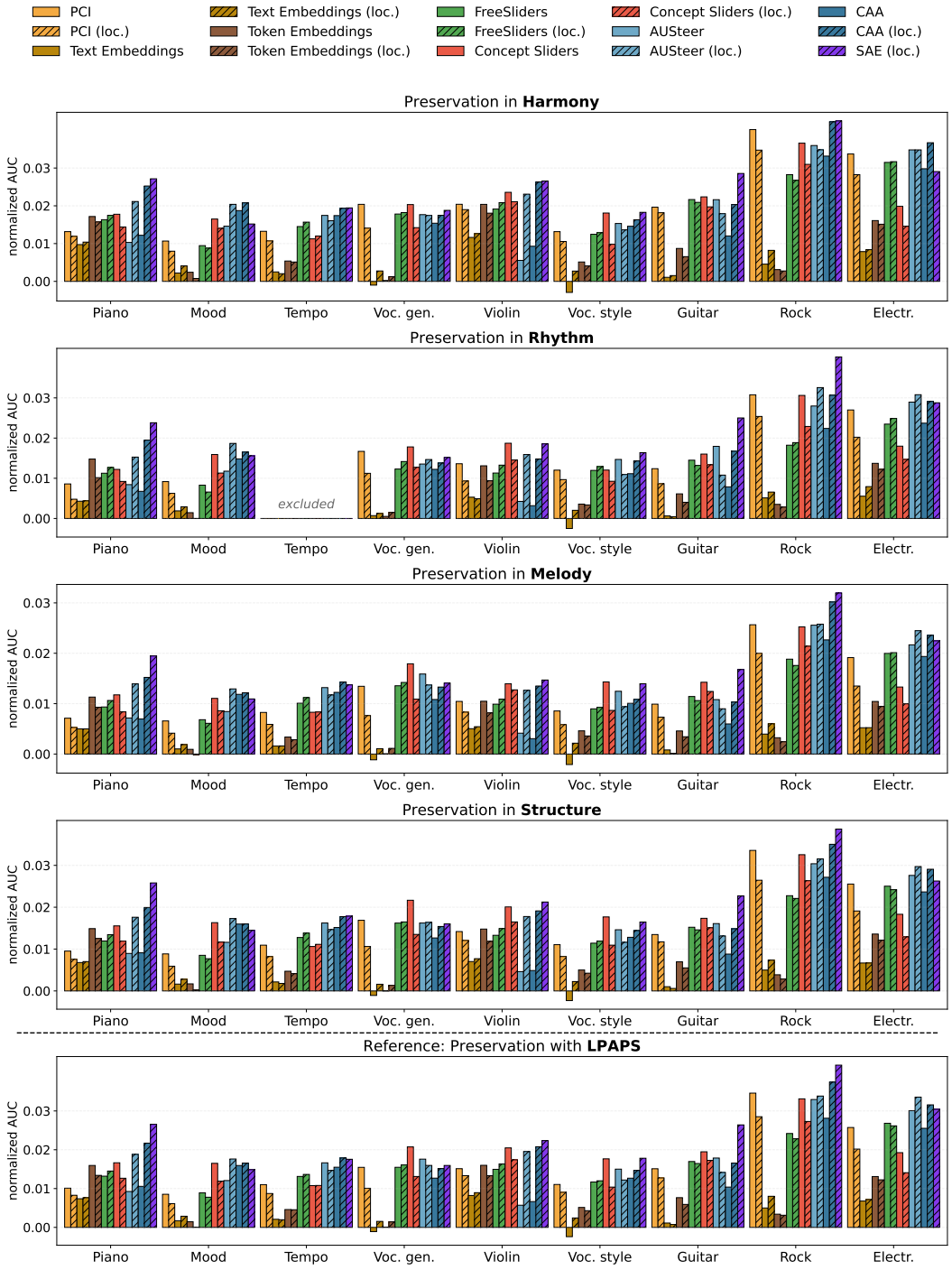


Figure 11: **Per-concept steering methods comparison across different preservation metrics: Harmony, Rhythm, Melody, Structure.** We report preservation–alignment AUC per concept, averaged over both alignment metrics: MuQ and CLAP. Hatched bars represent localized variants of the methods. Localized activation steering (CAA, AUSteer, SAE) leads on nearly every concept and axis.

H.4 GENERATION QUALITY OVER STEERING RANGE

Activation steering intervenes in the model's hidden states, propagating to subsequent layers a signal that was never encountered during training. A natural concern is that such interventions push activations off the training manifold, producing distortions and disrupting the quality of generated audio, even when the target concept is successfully modulated. In this section, we discuss this problem both theoretically and experimentally.

First, the common architectural design of transformers softens the train–inference activation mismatch. We intervene at the output level of the attention block, which, in all evaluated backbones, is followed by a normalization layer: RMSNorm in Ace-Step and LayerNorm in both AudioLDM2 and Stable Audio Open. One beneficial aspect of these layers is that they renormalize the perturbed activations back toward the statistics observed during training, before the next block consumes them.

Second, during our evaluation, the steering ranges are bounded by the distortion induced by prompt replacement, which keeps all methods within the model's natural operating regime. Thus, any off-distribution damage would surface directly as an audio quality drop in Table 1. Under calibrated conditions, activation steering does not produce tracks of worse quality than other baselines. To further quantify this phenomenon, we explicitly analyze the degradation of quality as a function of steering strength.

We measure the Production Quality (PQ) score from Audiobox Aesthetics (Tjandra et al., 2025), which targets acoustic "clarity and fidelity" rather than content and therefore serves as a good probe of off-manifold degradation. The unsteered baseline scores PQ $= 7.40$ on the 1–10 scale. Table 11 reports the change in PQ relative to this baseline for the activation-steering methods in their global and localized configurations, as a function of the methods' steering strength ($\alpha/\alpha_{\max}$). We report the average across the benchmark concepts in the concept-adding direction. We exclude the *mood* concept from this experiment, as amplifying it directly biases the quality score assigned by the external model.

The overall effect of steering on quality is small throughout the calibrated range. Even at maximum strength, the PQ change stays within ± 0.16, roughly 2% of the baseline. However, we expect PQ to degrade more noticeably at strengths well beyond this range. Although the magnitudes are small, the trends differ qualitatively by intervention site. The localized variants show almost no degradation at any strength (at full α: CAA $+0.10$, AUSteer -0.01, SAE $+0.03$), whereas both global variants decline with α. This suggests that injecting the change only into the layers that carry semantic control achieves the intended shift with a far smaller perturbation of the network's internal state, and less of the representation is pushed out of distribution.

Table 11: **Change in Production Quality (PQ) under increasing strength in activation steering methods.** We report ΔPQ relative to the unsteered baseline (PQ $= 7.40$) at increasing steering strengths. Localized variants show little-to-no degradation at any strength, whereas global methods decline as steering strength (α) increases.

$\alpha/\alpha_{\max}$	CAA	CAA (loc.)	AUSteer	AUSteer (loc.)	SAE (loc.)
0.0	0.000	0.000	0.000	0.000	0.000
0.2	-0.028	$+0.024$	$+0.008$	$+0.030$	$+0.046$
0.4	-0.045	$+0.041$	-0.043	$+0.043$	$+0.067$
0.6	-0.051	$+0.049$	-0.090	$+0.015$	$+0.045$
0.8	-0.055	$+0.059$	-0.137	-0.001	$+0.034$
1.0	-0.047	$+0.096$	-0.154	-0.010	$+0.030$

I STEERING METHODS

I.1 DESCRIPTIONS

We adapt seven steering paradigms from prior work to Ace-Step and group them by the space in which they intervene: prompt-level (PCI Görgün et al. (2026), Text Embeddings Ezra et al. (2025), Token Embeddings Baumann et al. (2025)), weight-space (Concept Sliders Gandikota et al. (2024)), score-space (FreeSliders Ezra et al. (2025)), and activation-space (CAA Rimsky et al. (2024), AUSteer Feng et al. (2026), SAEs Olshausen & Field (1997b)).

I.1.1 PROMPT-LEVEL INTERVENTIONS

These methods modify the model's text conditioning, leaving its weights and internal activations untouched.

Prompt-Conditioned Intervention (PCI) (Görgün et al., 2026) modifies the conditioning prompt during diffusion process. For each test prompt, a triple $(\mathcal{P}_n, \mathcal{P}_c, \mathcal{P}_{\tilde{c}})$ of neutral, positive (concept-promoting), and negative (concept-suppressing) prompts is used. The denoising trajectory uses $\mathcal{P}_n$ for the first $T - |\alpha|$ inference steps and the directional prompt ($\mathcal{P}_c$ or $\mathcal{P}_{\tilde{c}}$) for the remaining $|\alpha|$ steps. Thus, the integer $\alpha \in [-T, \ldots, 0, \ldots, T]$ encodes both *direction* (sign) and *strength* (number of switched steps) for modulation. PCI's maximum perceptual distortion, when applying directional prompt in all diffusion steps, serves as the reference point used to calibrate $\alpha_{\max}$ for all other methods (see Section 5.2).

Text Embeddings (Ezra et al., 2025) method interpolates in the text-encoder hidden-state space. For each test (neutral) prompt, we build the positive and negative prompts, which are constructed by adding prefixes and suffixes from App. G. Next, all three prompts are encoded with the frozen T5 encoder to obtain embeddings $\mathbf{e}_n, \mathbf{e}_c, \mathbf{e}_{\tilde{c}} \in \mathbb{R}^{L \times d}$ (with L denoting prompt length in tokens) and form the conditioning used by the model during inference $\mathbf{e}'(\alpha) = \mathbf{e}_n + \alpha \cdot \big(\mathbf{e}_c - \mathbf{e}_{\tilde{c}}\big)$, with steering strength α. Within each triple, the prompts are constructed so that all three have the same token count and the same padding after being converted to embeddings.

Token Embeddings (Baumann et al., 2025) restrict the perturbation to the embeddings of the *target-concept token* in the prompt rather than the full embedding sequence. For each concept, a target token (e.g., `instrument` for the piano, `vocal` for vocal-gender, `song` for tempo/mood/genre) is selected. A direction vector $\mathbf{d}_c \in \mathbb{R}^d$ is precomputed by averaging $(\mathbf{e}_c^{(i)} - \mathbf{e}_{\tilde{c}}^{(i)})$ over the 50 steering base prompts of Sec. G, where $\mathbf{e}_c^{(i)}$ is the encoder hidden state at the concept-token position i. At inference time, the test prompt is unchanged, and $\alpha \cdot \mathbf{d}_c^{(i)}$ is added only to the hidden state at the concept-token position. Compared to Text Embeddings, this leaves the rest of the prompt structurally identical and isolates the edit to the encoding of the named subject.

I.1.2 WEIGHT-BASED STEERING

Concept Sliders (Gandikota et al., 2024) encode a steering direction as a low-rank weight delta. LoRA adapters (Hu et al., 2022) ($\Delta\theta$) of rank r are attached to the attention projections (W^Q,W^K,W^O,W^V) of every transformer block in the model and trained with the enhanced-direction objective

$$\mathcal{L}_{\text{CS}} = \Big\| \boldsymbol{\epsilon}_{\theta+\Delta\theta}(x_t, \mathcal{P}_n) - \big[\boldsymbol{\epsilon}_\theta(x_t, \mathcal{P}_n) + \eta \cdot \big(\boldsymbol{\epsilon}_\theta(x_t, \mathcal{P}_c) - \boldsymbol{\epsilon}_\theta(x_t, \mathcal{P}_{\tilde{c}})\big)\big] \Big\|_2^2, \tag{4}$$

where η is a guidance hyperparameter at training time. At inference time, trained LoRAs are merged into the base weights with a scalar α, i.e. $\theta' = \theta + \alpha \cdot \Delta\theta$. Negative α inverts the trained direction (suppressing the concept).

I.1.3 SCORE-BASED STEERING

FreeSliders (Ezra et al., 2025) method, a training-free version of Concept Sliders, edits the noise prediction directly using a contrastive direction in score (diffusion prediction) space. Given the neutral, positive, and negative prompts triple, three conditional forward passes per denoising step are run and combined as

$$\boldsymbol{\epsilon}'(x_t, \alpha) = \boldsymbol{\epsilon}_{\text{cfg}}(x_t, \mathcal{P}_n) + \alpha \cdot \big(\boldsymbol{\epsilon}_\theta(x_t, \mathcal{P}_c) - \boldsymbol{\epsilon}_\theta(x_t, \mathcal{P}_{\tilde{c}})\big), \tag{5}$$

where $\boldsymbol{\epsilon}_{\text{cfg}}(x_t, \mathcal{P}_n)$ is the standard CFG-guided velocity for the neutral prompt and $\boldsymbol{\epsilon}_\theta(x_t, \mathcal{P}_c)$ denotes a single conditional pass.

I.1.4 ACTIVATION STEERING

Activation-steering methods leave the prompt and the weights unchanged and intervene in the outputs of the cross-attention layer l, $\mathbf{h}_l \in \mathbb{R}^{T \times d}$, during the forward pass. With classifier-free guidance, we always apply the intervention only to the conditional pass, leaving the unconditional pass untouched. In our experiments, we calculate steering vectors for each diffusion step separately.

Contrastive Activation Addition (CAA). We compute steering vectors in Cross-Attention (C-A) layers' outputs following CASteer (Gaintseva et al., 2026). Given N contrastive prompt pairs $\{(\mathcal{P}_c^{(i)}, \mathcal{P}_{\tilde{c}}^{(i)})\}_{i=1}^N$ for concept c, we collect activations and compute the steering vector $\mathbf{v}_c^{\text{CAA}}$ as:

$$\mathbf{v}_c^{\text{CAA}} = \frac{\mathbf{v}_c}{\|\mathbf{v}_c\|_2}, \text{where} \quad \mathbf{v}_c = \frac{1}{N}\sum_{i=1}^{N}\left(\bar{\mathbf{h}}_c^{(i)} - \bar{\mathbf{h}}_{\tilde{c}}^{(i)}\right), \tag{6}$$

with $\bar{\mathbf{h}}$ denoting C-A outputs averaged across temporal frames. During inference, we intervene by modifying activations at functional layers as $\mathbf{h}_l' = \mathbf{h}_l + \alpha \cdot \mathbf{v}_c^{\text{CAA}}$, where $\alpha \in \mathbb{R}$ controls steering strength, with positive values adding and negative values removing the modulated concept. Steering vectors are computed separately for each diffusion step.

AUSteer. AUSteer (Feng et al., 2026) replaces the dense mean-difference of CAA with a sparse, momentum-based selection of discriminative dimensions. Given the same contrastive activation pairs $\{(\mathbf{h}_c^{(i)}, \mathbf{h}_{\tilde{c}}^{(i)})\}_{i=1}^N$ at cross-attention layer l and diffusion step t, per-dimension momentum is computed across all N pairs and all temporal frames $f \in 1, \ldots, F$ as $m_j^{(i,f)} = h_{c,j}^{(i,f)} - h_{\tilde{c},j}^{(i,f)}$, and each dimension j is scored by the consistency of the sign of m_j over the S pooled samples (pairs $\times$ timeframes):

$$r_j^+ = \frac{1}{S}\sum_{i,f}[m_j^{(i,f)} > 0], \quad r_j^- = \frac{1}{S}\sum_{i,f}[m_j^{(i,f)} < 0], \quad \beta_j = \text{sign}(r_j^+ - r_j^-) \cdot \max(r_j^+, r_j^-). \tag{7}$$

The dimension-wise score $|\beta_j|$ thus lies in $[0, 1]$ and is large only where the sign is stable across the dataset. Next, a *global* top-s selection is performed across all active layers at each diffusion step, zeroing out β outside the top-s, and obtaining the sparse vector $\mathbf{v}_{c,l,t}^{\text{AUSteer}}$.

While the original AUSteer formulation (Feng et al., 2026) employs scores to scale single activation dimensions

$$\mathbf{h}_l' = \mathbf{h}_l \odot \big(\mathbf{1} + \alpha \cdot \mathbf{v}_{c,l,t}^{\text{AUSteer}}\big), \tag{8}$$

when adapted to diffusion models, our preliminary experiments resulted in zero gain in concept steering (see Table 15). Thus, we evaluate a CAA-style *additive* formulation that uses the sparse β vector as a steering direction added to the activation:

$$\mathbf{h}_l' = \mathbf{h}_l + \alpha \cdot \mathbf{v}_{c,l,t}^{\text{AUSteer}}. \tag{9}$$

Compared to CAA, AUSteer (a) selects *which* dimensions to edit using a discriminative criterion rather than averaging, (b) allocates a fixed global top-s budget, encouraging sparsity, and (c) implicitly performs a soft (discriminative) layer selection, where layers which dimensions never enter the global top-s are not steered, even when nominally included in the set of layers.

Interestingly, in Table 16, we show that the discriminative localization performed with AUSteer does not necessarily result in the same layer selection as our Activation Patching, often leading to suboptimal results compared with our approach.

Sparse Autoencoders (SAEs). We train a TopK SAEs (Gao et al., 2025; Bussmann et al., 2024) on the layers localized with activation patching. The SAE with encoder weights $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{md \times d}$, decoder weights $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{d \times md}$, and bias $\mathbf{b}_{\text{pre}} \in \mathbb{R}^d$ maps activations $\mathbf{h} \in \mathbb{R}^d$ to a sparse code $\mathbf{f} \in \mathbb{R}^{md}$ (with m being the expansion factor):

$$\mathbf{f} = \text{TopK}\left(\mathbf{W}_{\text{enc}}(\mathbf{h} - \mathbf{b}_{\text{pre}})\right), \tag{10}$$

and further reconstructs them:

$$\hat{\mathbf{h}} = \mathbf{W}_{\text{dec}}\mathbf{f} + \mathbf{b}_{\text{pre}}. \tag{11}$$

The TopK$(\cdot)$ operation retains only the k largest activations in the autoencoder's latent space, zeroing out the rest. The SAE is trained to minimize reconstruction error $\|\mathbf{h} - \hat{\mathbf{h}}\|_2^2$. To identify concept-specific features, we use two contrastive sets of N prompts $(\mathcal{P}_c, \mathcal{P}_{\tilde{c}})$ for concept c to compute importance scores using a TF-IDF-based criterion:

$$\text{score}(j, c) = \underbrace{\mu_j(\mathcal{P}_c)}_{\text{TF}} \cdot \underbrace{\log\left(1 + \frac{1}{\mu_j(\mathcal{P}_{\tilde{c}}) + \epsilon}\right)}_{\text{IDF}}, \tag{12}$$

where $\mu_j(\mathcal{P}_\rfloor) = \frac{1}{|\mathcal{P}_\rfloor|}\sum_{\mathbf{h} \in \mathcal{P}_\rfloor} f_j(\mathbf{h})$ is the mean activation of feature j in the SAE's latent space on generated audio with prompts $\mathcal{P}_\rfloor$. Features that activate highly for concept c but rarely for other samples receive high scores. We select the top-τ_c scoring features $\mathcal{F}_c$ and, by summing their corresponding decoder columns, we construct the steering vector $\mathbf{v}_c^{\text{SAE}}$ as

$$\mathbf{v}_c^{\text{SAE}} = \sum_{j \in \mathcal{F}_c} \mathbf{W}_{\text{dec}}[:, j], \tag{13}$$

and add it directly to the output of the cross-attention layer:

$$\mathbf{h}_l' = \mathbf{h}_l + \alpha \cdot \mathbf{v}_c^{\text{SAE}}. \tag{14}$$

I.2 Implementation and hyperparameter details

Activation Patching. For Activation Patching (Meng et al., 2022), we adapt audio diffusion models to the NNsight Fiotto-Kaufman et al. (2025) API and perform localization using their library.

Shared setup. Unless stated otherwise, all steering experiments use a single shared inference configuration for Ace-Step: $T = 30$ flow-matching steps with the Euler-discrete scheduler, CFG scale 5.0, audio duration 30s, and the same 100 test prompts (Sec. G). For each method, α is swept over 31 values: 15 negative, $\alpha = 0$, and 15 positive. The $\alpha = 0$ run uses the unmodified pipeline and serves as the per-prompt preservation baseline against which LPAPS distortion is measured. Each α produces 100 steered audio clips, yielding a total of 30×100 samples per concept per method. To choose method settings, we sweep over key hyperparameters using a held-out dataset of 20 prompts with 9 steering strengths.

Concept Sliders. We attach LoRA adapters to all attention projections of every transformer block. Each concept's LoRA is trained for 500 iterations with AdamW at a learning rate of 10^{-4} and a training-time guidance $\eta = 7$. In Table 12, we show that choosing different values for these parameters does not yield a significant difference in steering performance. Additionally, in Table 13, we find, for each concept, on the held-out dataset, the optimal LoRA rank r. Concept Sliders in our main experiments use LoRAs trained with the optimal rank for each concept.

Table 12: **Hyperparameter impact on audio concept modulation with Concept Sliders.** We report preservation-alignment AUC (averaged over piano and vocal concepts) on the held-out dataset. The setting used in the experiments is marked $\star$, the best per axis is in **bold**.

	Learning rate			**Train guidance**			**Training iters**		
Value	1e-4 $\star$	5e-4	5e-3	3	5	7 $\star$	500 $\star$	1000	1500
Avg AUC	**0.102**	0.101	0.083	**0.104**	0.103	0.102	**0.102**	0.101	0.101

Table 13: **Per-concept LoRA-rank sweep for Concept Sliders.** We report preservation–alignment AUC. The per-concept highest scoring parameter (used in the benchmark) is **bolded**.

Rank	Piano	Mood	Tempo	Voc. gen.	Violin	Voc. style	Guitar	Rock	Electr.	Avg
4	0.061	**0.039**	0.038	**0.104**	0.080	0.162	0.069	0.076	0.088	0.080
8	**0.061**	0.036	0.038	0.102	**0.082**	0.155	0.071	0.082	**0.091**	0.080
16	0.061	0.036	**0.043**	0.103	0.080	**0.165**	**0.075**	**0.082**	0.090	**0.082**

AUSteer. We rank activation dimensions across all active layers at each diffusion step, and select the top-s of them to construct the steering vector. In the localized variant, only dimensions in $\{7, 8\}$ are considered in ranking and selection. In Table 14, we sweep budget s separately for both method configurations on the held-out dataset: $s \in \{256, 1024, 2048, 4096, 8192, 16384\}$ for all-layer steering, and $s \in \{256, 1024, 2048, 4096, 5120\}$ for the localized variant (two active layers expose a total of $2d = 5120$ dimensions). In our experiments, we use the best-scoring s for each concept and method's variant. Results in Table 14 suggest that the best s depends on the steered concept and localized steering is more effective than a global approach, with exceptions for *tempo* and *vocal gender* concepts.

Table 14: **Per-concept top-s sweep for AUSteer.** We report preservation–alignment AUC (mean of MuQ-T and CLAP over both steering directions) on the held-out dataset. The best value in each column is **bolded**.

Top-s	Piano	Mood	Tempo	Voc. gen.	Violin	Voc. style	Guitar	Rock	Electr.	Avg
All-layer steering										
256	0.049	0.034	0.035	0.034	**0.038**	0.024	0.052	0.087	0.097	0.050
1024	0.048	**0.046**	0.042	0.037	0.030	0.069	0.068	0.108	0.135	0.065
2048	**0.054**	0.039	0.048	0.039	0.022	0.082	**0.075**	**0.118**	0.143	0.069
4096	0.040	0.030	**0.055**	0.063	0.034	0.106	0.069	0.109	**0.151**	**0.073**
8192	0.024	0.028	0.046	**0.065**	0.029	**0.124**	0.051	0.105	0.115	0.065
16384	0.016	0.033	0.042	0.056	0.035	0.124	0.044	0.104	0.108	0.062
Localized steering (layers $\{7, 8\}$)										
256	0.029	0.030	0.030	0.032	0.029	0.103	0.056	0.082	0.110	0.056
1024	0.060	0.035	0.042	0.029	0.048	0.130	0.068	0.124	0.164	0.078
2048	0.047	0.040	0.042	**0.045**	**0.074**	0.133	0.078	0.127	**0.166**	**0.083**
4096	**0.061**	0.045	**0.046**	0.034	0.066	0.139	0.079	0.126	0.145	0.082
5120	0.049	**0.049**	0.046	0.034	0.059	**0.148**	**0.080**	**0.129**	0.140	0.082

In Table 15, we analyze methods for constructing steering vectors from AUSteer's scoring: scaling activation dimensions by scores (*multiplicative*, Eq. (8)) or concatenating into a single vector and, as in CAA, adding to intermediate activations (*additive*, Eq. (9)). Across all evaluated concepts, the CAA-style approach performs better.

In Table 16, we analyze how AUSteer's ranking-and-selection mechanism, which can be considered an alternative to our localization, distributes activation dimensions across Ace-Step layers. Interestingly, we find that their top-s selection produces surprisingly distinct layer allocations across concepts. In most cases, our functional layers $\{7, 8\}$ contribute less than 50% to the total budget, with the majority of selected dimensions coming from later layers.

Table 15: **Impact of steering vector construction on AUSteer performance.** We report preservation–alignment AUC on the held-out set over all concepts and **highlight** the best method.

Method	Piano	Mood	Tempo	Voc. gen.	Violin	Voc. style	Guitar	Rock	Electr.	Avg
add. (eq. 9)	**0.107**	**0.031**	**0.042**	**0.087**	**0.102**	**0.219**	**0.109**	**0.191**	**0.211**	**0.122**
mult. (eq. 8)	0.028	−0.005	−0.009	0.028	0.003	−0.008	0.009	0.028	−0.013	0.007

Considering AUSteer's layer allocation results (Table 16) and per-concept steering performance (Tables 26 to 34), we find that AUSteer's discriminative localization often degrades steering performance. Precisely, on *piano* (Table 26), *mood* (Table 28), and *violin* (Table 31) concepts, where our layers contribute, respectively, in 2.4%, 12.0%, and 16.3% to the budget, all-layer AUSteer scores, respectively, two (0.049 vs. 0.102), two (0.026 vs. 0.053), and three (0.032 vs. 0.104) times worse than localized approach. Usually, steering across all layers outperforms the localized one when our layers contribute to the budget, i.e., for the *tempo* (47.7%) and *guitar* (63.7%) concepts.

This suggests that AUSteer's discriminative-momentum score selects layers where the concept is most reliably encoded (discriminative) rather than where intervention is most effective (causal), thereby limiting the method's performance on audio concept modulation compared to our localized approach.

Table 16: **Ace-Step layers allocated with AUSteer's rank-and-select localization.** We report the percentage of activation dimensions in attention layers, averaged across diffusion steps, for all nine concepts. Layers with at least 1% allocation to at least one concept are listed, and the functional layers ($\{7, 8\}$) we localize are highlighted.

Layer	Piano	Mood	Tempo	Voc. gen.	Violin	Voc. style	Guitar	Rock	Electr.
5	9.6	0.3	1.3	0.3	17.5	11.1	1.8	0.2	17.7
6	5.3	0.0	0.4	0.0	8.4	2.7	2.2	0.0	0.2
7	0.6	0.0	17.6	17.6	1.4	3.0	1.6	8.8	1.6
8	1.8	12.0	30.1	37.6	14.9	17.3	62.1	53.5	71.8
9	0.0	0.7	1.1	0.0	0.2	15.2	0.2	2.5	0.4
12	0.4	0.0	0.0	5.8	0.6	0.1	0.1	0.0	0.1
13	1.5	0.1	0.0	0.0	0.7	3.3	0.0	0.0	0.0
14	0.1	58.7	3.4	0.0	0.4	3.9	3.2	5.4	0.6
15	0.0	11.6	0.0	0.0	0.1	0.0	0.1	0.0	0.0
16	0.2	0.0	1.1	1.2	0.0	3.4	5.4	0.4	0.1
17	1.6	0.0	0.0	4.4	0.5	0.0	0.1	0.0	0.0
18	11.1	3.1	0.8	1.0	4.1	9.0	0.5	0.4	0.0
19	6.7	11.6	24.1	0.0	8.1	14.2	4.7	0.1	0.0
20	0.2	0.0	6.8	0.4	0.8	0.8	3.4	0.8	2.2
22	59.2	1.8	11.8	7.5	41.3	6.6	10.6	21.2	3.1
23	0.1	0.0	0.1	10.1	0.2	1.9	2.4	0.1	1.9
24	0.2	0.0	1.3	14.1	0.2	6.9	1.2	6.5	0.0

Contrastive Activation Addition (CAA). For steering vectors calculation, we first gather activations from the outputs of the cross-attention blocks across all denoising steps. Our methodology is roughly equivalent to CASteer (Gaintseva et al., 2026), a method designed for text-to-image diffusion models. Unlike vision, where inputs are split spatially into a grid of patches, each a vector of activations, audio models split signals along the time axis into timeframes. The collected layer outputs from positive and negative prompt runs (see Section G) are first averaged along the time axis, then contrasted, and finally averaged over all N prompt pairs. During experiments, we steer only the conditional (*cond*) prediction within classifier-free guidance (CFG). Our analysis in Table 17, conducted on the held-out dataset, shows that steering both conditional and unconditional (*null*) predictions separately does not yield significantly better results, despite being twice as costly.

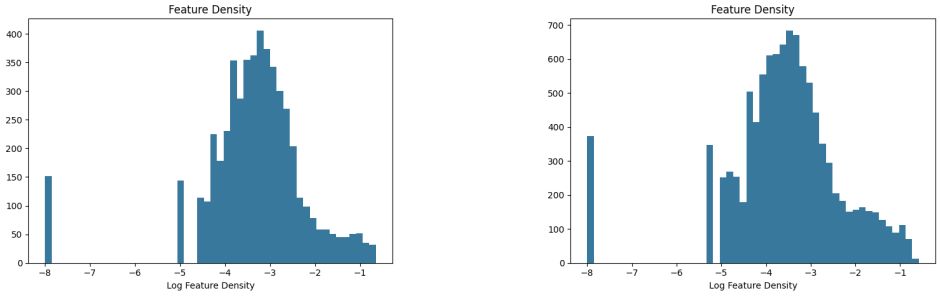
Table 17: **Impact of model's classifier-free guidance (CFG) prediction on concept steering with CAA.** We report preservation-alignment AUC over concepts. Steering separately both conditional and unconditional (*cond* + *null*) predictions does not yield significantly different results than steering only the conditional prediction (*cond*).

Steered CFG passes	Piano	Mood	Tempo	Voc. gen.	Violin	Voc. style	Guitar	Rock	Electr.	Avg
cond	0.104	0.035	**0.059**	0.075	0.105	**0.239**	**0.117**	0.192	**0.191**	0.124
cond + null	**0.110**	**0.040**	0.058	**0.089**	**0.110**	0.230	0.114	**0.203**	0.177	**0.126**

SAEs. We train one BatchTopK SAE (Bussmann et al., 2024; Gao et al., 2025) per localized layer, on the outputs of the attention layers $\{7, 8\}$. Activations are collected at different diffusion steps (every 5 of 30) for MusicCaps (Agostinelli et al., 2023) prompts. We sweep expansion factors $m \in \{2, 4, 8, 16\}$ and sparsity $k \in \{16, 32, 64\}$, and select the configuration that minimizes reconstruction error while keeping the fraction of dead and high-frequency features low: $(m, k) = (2, 32)$ for layer $\{7\}$, trained for 10 epochs, and $(m, k) = (4, 64)$ for layer $\{8\}$, trained for 15 epochs. In Table 18, we report the values: *FVU* (fraction of variance unexplained, measures reconstruction error), $\text{dead}_{\%}$ (the fraction of features that have not fired for more than 10^7 inputs), and $\text{fire}_{\%}$ (the fraction of features that fire at least once every batch), while in Fig. 12, we present feature density in trained SAEs.

Table 18: **Training quality metrics for our BatchTopK SAEs.** Lower FVU indicates better SAE reconstruction, lower $\text{dead}_{\%}$ indicates a smaller fraction of features that never activate; lower $\text{fire}_{\%}$ indicates fewer features that activate every batch (global).

ACE-Step Layer	FVU (↓)	$\text{dead}_{\%}$ (↓)	$\text{fire}_{\%}$ (↓)
7	0.216	0.37%	19.4%
8	0.243	0.25%	20.1%



(a) Layer 7. (b) Layer 8.

Figure 12: **Feature-density histograms for Sparse Autoencoders trained Ace-Step semantic bottleneck.** Features in SAE latent space are considered dead (never used) when their density $< 10^{-7}$, while high-frequency features (activating at least once per batch) are the ones with density $> 10^{-2}$. We consider features in between to be prospective for steering.

Concept features set $\mathcal{F}_c$ is selected per layer by ranking SAE features under the TF–IDF criterion (see eq. 12), scored on model activations collected from the same contrastive prompt pairs that one in CAA and AUSteer. To construct $\mathcal{F}_c$, we use the k highest-scoring SAE features, where the optimal k is chosen separately for each concept on the held-out dataset. In Fig. 13, we sweep $k \in \{5, 10, 20, 50, 100, 500\}$ over both $\{7, 8\}$ functional layers in ACE-Step. Our main experiments with SAE use winning pairs for concepts: piano $(5, 20)$, mood $(500, 10)$, tempo $(20, 10)$, vocal gender $(100, 5)$, violin $(20, 50)$, vocal style $(20, 20)$, guitar type $(5, 50)$, jazz genre $(50, 20)$ and classical genre $(100, 10)$. Interestingly, steering performance for some concepts (e.g., *piano*, *vocal gender*) depends significantly on k, whereas for others (e.g., *mood*, *tempo*) it is uniform across pairs.

The concept's steering vector $\mathbf{v}_c^{\text{SAE}}$ is constructed as a unit-weighted summation of SAE's decoder columns, corresponding to its top-k selected dimensions. Alternatively, during summation, one can weight decoder columns by their corresponding TF–IDF (Eq. (12)) scores or by their average (across prompts) value in latent space (TF). We analyze this approach in Table 19 and show that it is competitive to our uniform feature weighting, when weights are rescaled in a way that they cumulatively equal k. We denote such rescaling as $\Sigma = k$.

We perform negative-direction (concept-suppressing) steering with SAE by applying a positive-direction steering vector with the strength parameter α negated. Constructed from decoder columns, the steering vector lives in the model's activation space. Negated, under the linear representation hypothesis (Park et al., 2024), should correspond to reducing the amount of the concept in the generated audio. An alternative approach would be to identify a negative feature set via TF-IDF scoring over the concept-suppressing prompts dataset and create a steering vector from the corresponding decoder columns. In Table 20 we compare both approaches to negative-direction steering and show that ours (applying on-direction vector with negative steering strength) offers better concept suppression.

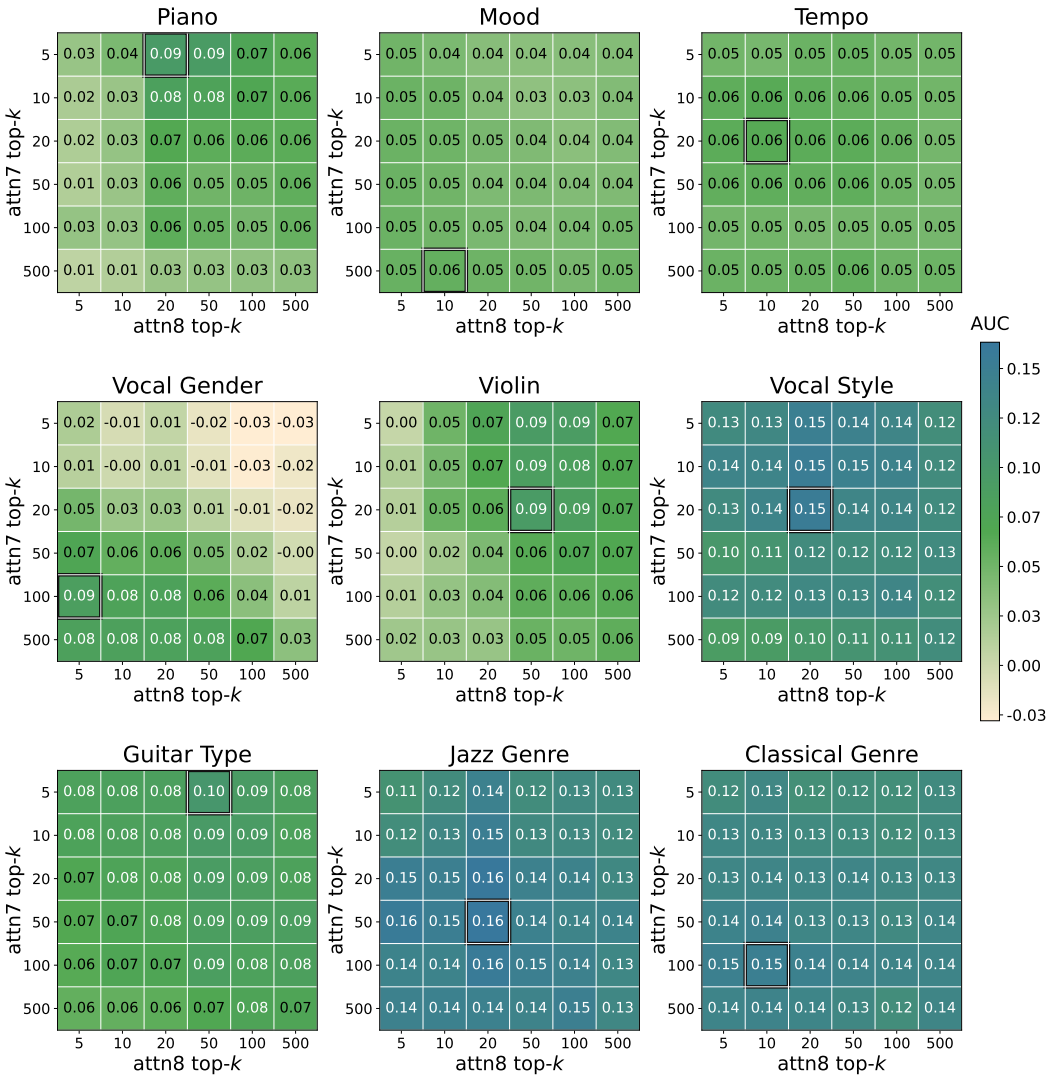


Figure 13: **Impact of SAE feature count (k) on each concept modulation performance.** We find that optimal configuration of k for layers $(attn7, attn8)$ highly depends on given concept. In our experiments, we use the best-scoring pair for each concept, which is highlighted.

Table 19: **Impact of SAE's features weighting on steering performance.** We report AUC on the held-out set. While weighting decoder columns by raw scores yields suboptimal results, rescaling scores ($\Sigma = k$) yields performance competitive with our ($\star$) approach.

Weighting	Piano	Mood	Tempo	Voc. gen.	Violin	Voc. style	Guitar	Rock	Electr.	Avg
unit $\star$	0.124	0.031	**0.065**	0.126	**0.117**	0.233	0.123	**0.236**	**0.174**	**0.137**
TF	0.110	**0.040**	0.034	0.129	0.074	0.239	0.137	0.173	0.143	0.120
TF–IDF	0.095	0.032	0.041	0.112	0.092	0.220	0.143	0.169	0.160	0.118
TF, $\Sigma = k$	**0.128**	**0.040**	0.052	**0.138**	0.090	**0.247**	**0.167**	0.225	0.143	**0.137**
TF–IDF, $\Sigma = k$	0.113	0.033	0.053	0.104	0.108	0.215	0.160	0.225	0.165	0.131

Table 20: **Constructing negative direction (concept-removal) steering vectors with SAE.** We report AUC on the held-out dataset. Our ($\star$) approach of applying a positive (concept-amplifying) vector with negative steering strength performs better than identifying a negative-feature set.

Negative SV Construction	Piano	Mood	Tempo	Voc. gen.	Violin	Voc. style	Guitar	Rock	Electr.	Avg
negated positive dir. vector $\star$	**0.065**	**0.083**	**0.063**	**0.046**	0.070	**0.075**	**0.081**	**0.091**	**0.123**	**0.077**
TF–IDF on negative prompts	0.008	0.033	0.058	0.035	**0.075**	0.045	0.016	0.028	0.039	0.037

J STUDIED AUDIO DIFFUSION ARCHITECTURES

In our experiments, we employ audio generative models with diverse architectures, namely *AudioLDM2* (Liu et al., 2024), *Stable Audio Open* (Evans et al., 2025), and *Ace-Step* (Gong et al., 2025). While AudioLDM2 and Stable Audio Open are trained with diffusion loss (Ho et al., 2020), Ace-Step are trained to learn a velocity field within the flow matching paradigm (Lipman et al., 2023). AudioLDM2 leverages a U-Net (Ronneberger et al., 2015) architecture as a backbone and employs FLAN-T5 (Chung et al., 2024) as text encoder. Conversely, Stable Audio Open is a Diffusion Transformer (DiT, Peebles & Xie (2023)) architecture with a T5 (Raffel et al., 2020) text embedding model. Finally, Ace-Step uses mT5 (Chung et al., 2023) for text encoding, an autoregressive model (Liu et al., 2025b) for lyric encoding, and employs the Linear DiT (Xie et al., 2025) backbone for the denoising process.

K Details on user study

To complement the automatic evaluation, we conducted an anonymous in-browser listening study that compares the seven steering paradigms in their *Standard* (all-layer) and *Localized* configurations on three musical concepts: *piano*, *vocal gender*, and *tempo*.

Stimulus design. The study uses 15 trials, five per concept. Each trial is anchored by a single music prompt drawn from a stylistically diverse pool covering jazz ballads, shoegaze, reggae, Latin music, cinematic soundtracks, trap beats, classic rock, and dreamy meditation. For every trial we pre-render the full α-sweep over seven steering strengths $\alpha \in \{-\alpha_{\max}, -\frac{2}{3}\alpha_{\max}, -\frac{1}{3}\alpha_{\max}, 0, +\frac{1}{3}\alpha_{\max}, +\frac{2}{3}\alpha_{\max}, +\alpha_{\max}\}$ (calibrated to match Table 1 strengths) for all 13 steering methods, yielding $13 \times 7 = 91$ short MP3 clips per trial that are served as static assets. Each clip is rendered to loop seamlessly, so listeners can sweep α without hearing playback restart. The negative direction is meaningful: positive α *adds* the concept (e.g., introduces piano) and negative α *suppresses* it (e.g., removes piano-like content); $\alpha = 0$ is the unsteered baseline and serves as a perceptual anchor.

Per-trial method assignment. Showing all 13 methods on every trial would extend the per-trial duration well beyond the $\sim$1 min target, with predictable drop-off. We instead present *seven* samples per trial, drawn as follows:

- SAE($\{7, 8\}$) is included on every trial as an upper anchor; given the high overall performance and the fact that it has no Standard counterpart, it would otherwise be undersampled relative to the other paradigms.
- Three of the six remaining paradigm families (PCI, Text Emb, Token Emb, FreeSliders, AUSteer, CAA) are drawn uniformly at random per (session, trial). For each chosen family, both the *Standard* and *Localized* variants are included, contributing 6 samples.

The full 7-sample list is then shuffled so position cannot leak method identity. The order of the 15 trials is also independently shuffled per session. Both the trial order and the per-trial sample order are persisted to `localStorage`, so a refresh or a browser tab close does not reroll either. Concept Sliders is not included in this listening study.

Listening interface. The listening interface is presented in Fig. 14. Each trial displays the target concept (e.g., "Tempo") and a textual hint of what positive vs. negative α should accomplish (e.g., for *tempo*: "positive α should make the music faster; negative α should make it slower"). The listener sees seven rows, one per sample, anonymized as "Sample 1" … "Sample 7". Each row exposes a single play button and an α slider with seven stops. Crucially, the α slider does *not* restart playback: when a sample begins playing, all seven of its pre-rendered loops are wired in parallel through a Web Audio gain-crossfade graph (one gain node per α stop, all summing into a master gain that drives the destination). Dragging the slider executes a sample-accurate gain crossfade with a 5 ms time-constant between the active and target α stops, so the concept transition is heard *continuously* rather than as a sequence of discrete switches – which is what makes the *Smoothness* question perceptually meaningful in a listening setting. Only one sample plays at a time: pressing play on another row auto-pauses the currently-active one. To reduce wait time during navigation, the audio buffers for the next trial are prefetched with 6-way parallelism while the current trial is being rated.

Likert questions. Beneath the player are three Likert (1–5) scales:

- **Seamless Edit.** "How successfully did the audio adopt the target concept (as you increased α) without sounding unnatural or losing its original musical identity?" Anchors: 1 = *Failed / ruined the original context*, 5 = *Perfectly seamless*.
- **General Audio Quality.** "Rate the overall acoustic quality and realism of the generated audio." Anchors: 1 = *Poor / highly distorted*, 5 = *Excellent / studio quality*.
- **Smoothness.** "How natural and gradual is the transition of the target concept when changing the steering strength?" Anchors: 1 = *Abrupt / glitchy*, 5 = *Very smooth / natural*.

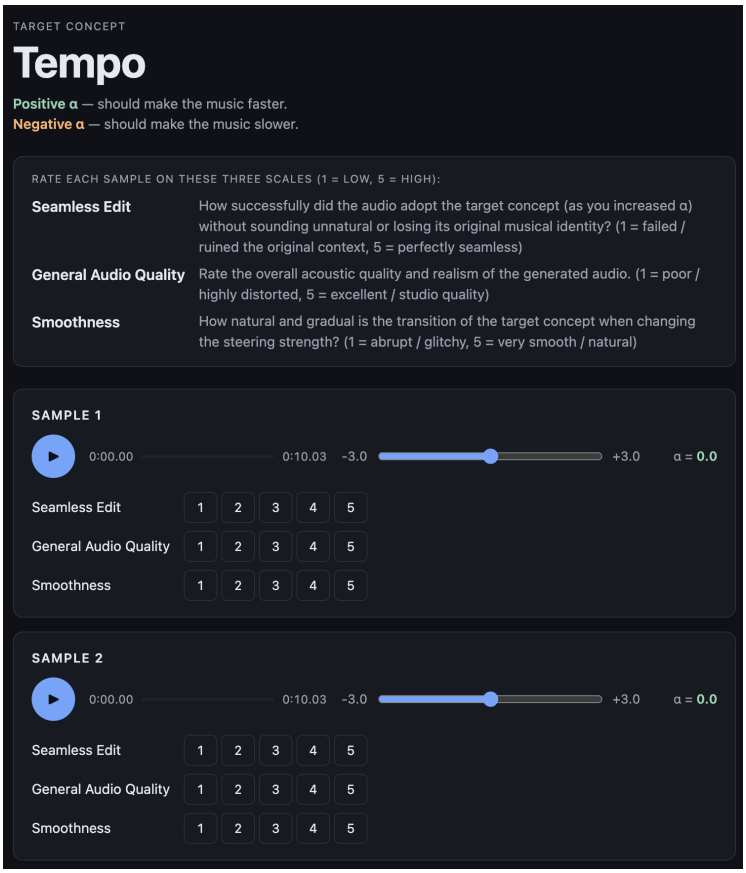


Figure 14: Print screen of the user study interface

A trial cannot be submitted until all three scales have been answered for every one of its seven samples. The landing page additionally asks the listener to self-rate their musical-skill level on a 1–5 scale (1 = *trouble distinguishing instruments in a recording*, through 5 = *professional musician*); this rating is logged but is not used for filtering or for any analysis in the main paper.

Logging, resumption, and analysis filter. For every (sample, trial) pair we log the session id (a UUIDv4 minted on the landing page), the trial's presentation index, the sample index within the trial, the trial concept and prompt, all α values offered (`alphas`), the α stops that were active at any point *while* the sample was playing (`alphas_visited`; values selected with the slider while paused are not logged), whether the play button was pressed at least once (`listened`), and the seven shuffled methods chosen for that trial (`selected_methods`). Submission is trial-by-trial: every *Next* click POSTs the which appends seven scores to the final `results.csv`.

In our analyses, we keep only rows with `listened` = 1, i.e. samples for which the listener actually pressed play. This filter (a) removes a small number of rows where a listener clicked through Likerts without auditioning the sample and (b) ensures that the engagement statistics reported below describe rows that contributed to the published numbers.

Participants and engagement. The study collected 1279 rated samples across 32 sessions, with 7 listeners completing all 15 trials.

Self-rated musical skill skewed towards casual listeners (2 sessions at level 1, 16 at level 2, 9 at level 3, 4 at level 4, 1 at level 5; mean 2.56). Engagement with the α slider was high: in 98.4% of rated samples the listener visited at least three distinct α stops while listening, and the mean number of α stops visited per sample was 6.2 out of 7 – consistent with active sweeping rather than rating the default position.

Question correlations. Although the three Likert scales received broadly similar mean ratings (*Seamless Edit*: 2.92, *Audio Quality*: 3.00, *Smoothness*: 2.92), they are far from redundant. The Pearson correlations between the per-rating scores are r=0.61 between Audio Quality and Smoothness, r=0.52 between Seamless Edit and Smoothness, and r=0.44 between Seamless Edit and Audio Quality. *Seamless Edit*, the dimension that asks specifically whether the steering succeeded, is therefore the most informationally independent of the three, supporting its central role in the per-method and pairwise analyses above.

Skill vs. method interaction. Self-reported musical skill modulates which paradigm a listener prefers (Table 21). Splitting the 32 sessions into *casual listeners* (skill ≤ 2, 497 rated samples) and *trained musicians* (skill ≥ 3, 782 rated samples), the localized activation-space methods CAA and AUSteer receive *higher* mean Seamless-Edit ratings from trained musicians than from casual listeners (CAA: 3.40 vs. 3.23; AUSteer: 3.42 vs. 3.10), while FreeSliders, PCI, Text Emb, and Token Emb move in the opposite direction (e.g. FreeSliders: 2.71 vs. 3.38, a gap of 0.67 points). One natural reading is that musically trained listeners are more sensitive to the artifacts introduced by score- and prompt-level edits, while activation-space edits remain perceptually clean enough to score even higher with critical ears. We caveat that only one session was rated at the highest skill level (n=105 ratings), so we group skill levels 3–5 together rather than reporting a finer breakdown.

Table 21: **Mean Seamless-Edit rating per method (Localized variant), split by self-reported musical skill.** Casual listener = skill ≤ 2; trained musician = skill ≥ 3. Activation-space methods are preferred more strongly by trained listeners; score- and prompt-level methods less so.

Method (loc.)	**Casual listeners** (≤ 2)	**Trained musicians** (≥ 3)	**Δ**
PCI	3.20	2.80	−0.40
Text Embeddings	2.69	2.32	−0.37
Token Embeddins	2.64	2.58	−0.06
FreeSliders	3.38	2.71	−0.67
CAA	3.23	3.40	+0.17
AUSteer	3.10	3.42	+0.32
SAE	3.27	3.17	−0.10

L ADDITIONAL RESULTS FOR LOCALIZED ACTIVATION STEERING

L.1 LOCALIZED ACTIVATION STEERING

In Fig. 15, we present a generalization of our localized activation steering with CAA across 4 musical concepts and 3 audio diffusion models. Notably, for some concepts (mood in AudioLDM2 and Stable Audio Open, and female vocal in ACE-Step), steering without our layers results in concept degradation. In Tables 22 to 24 we present quantitative results supporting this.

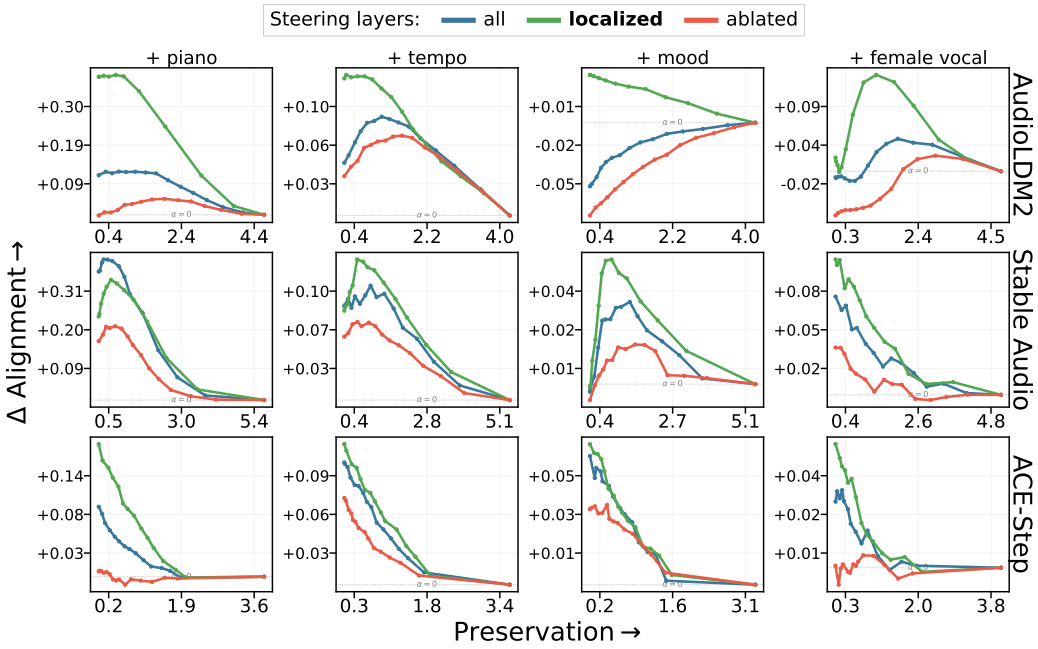


Figure 15: **Layer selection in activation steering is crucial for audio concept modulation.** Steering the activations in localized layers outperforms steering all of the layers, while steering all layers except the localized ones leads to drastic degradation.

Table 22: **Per-concept steering metrics for AudioLDM2 with CAA.** Higher is better for AUC, lower for Smoothness. The best per row is **bolded**.

Concept	Layers	AUC (MuQ) ↑	AUC (CLAP) ↑	Smoothness (MuQ) ↓	Smoothness (CLAP) ↓
Piano	all	0.331	0.175	**0.081**	0.098
	localized	**0.930**	**0.390**	0.111	**0.064**
	ablated	0.099	0.099	0.176	0.171
Tempo	all	0.236	0.008	0.125	0.222
	localized	**0.300**	**0.067**	**0.065**	**0.073**
	ablated	0.201	-0.007	0.135	0.605
Mood	all	-0.066	-0.005	–	0.361
	localized	**0.089**	**0.152**	**0.069**	**0.076**
	ablated	-0.130	-0.073	–	0.827
Female vocal	all	0.095	0.107	0.226	0.233
	localized	**0.304**	**0.337**	**0.193**	**0.128**
	ablated	-0.045	-0.061	0.437	0.472
Average	all	0.149	0.071	0.144	0.229
	localized	**0.406**	**0.236**	**0.109**	**0.086**
	ablated	0.031	-0.011	0.249	0.519

Table 23: **Per-concept steering metrics for Stable Audio Open with CAA.** Higher is better for AUC, lower for Smoothness. The best per row is **bolded**.

Concept	Layers	AUC (MuQ) ↑	AUC (CLAP) ↑	Smoothness (MuQ) ↓	Smoothness (CLAP) ↓
Piano	all	**0.792**	**0.228**	0.093	**0.105**
	localized	0.742	0.220	0.128	0.154
	ablated	0.383	0.094	**0.086**	0.129
Tempo	all	0.268	0.108	0.110	0.100
	localized	**0.323**	**0.216**	0.113	**0.085**
	ablated	0.183	0.031	**0.096**	0.202
Mood	all	0.074	0.241	0.199	0.099
	localized	**0.119**	**0.246**	**0.183**	**0.099**
	ablated	0.034	0.163	0.240	0.099
Female vocal	all	0.106	**0.122**	**0.089**	**0.041**
	localized	**0.153**	0.097	0.110	0.062
	ablated	0.030	0.087	0.148	0.066
Average	all	0.310	0.175	**0.123**	**0.086**
	localized	**0.334**	**0.195**	0.134	0.100
	ablated	0.158	0.094	0.142	0.124

Table 24: **Per-concept steering metrics for ACE-Step with CAA.** Higher is better for AUC, lower for Smoothness. The best per row is **bolded**.

Concept	Layers	AUC (MuQ) ↑	AUC (CLAP) ↑	Smoothness (MuQ) ↓	Smoothness (CLAP) ↓
Piano	all	0.060	0.026	0.038	0.063
	localized	**0.141**	**0.061**	**0.033**	**0.032**
	ablated	-0.008	-0.010	0.488	–
Tempo	all	0.095	0.022	0.032	0.075
	localized	**0.111**	**0.023**	**0.029**	**0.042**
	ablated	0.059	0.018	0.034	0.076
Mood	all	0.043	**0.049**	0.069	0.038
	localized	**0.049**	0.046	**0.042**	**0.035**
	ablated	0.036	0.044	0.087	0.064
Female vocal	all	0.020	0.104	0.138	0.034
	localized	**0.029**	**0.116**	**0.068**	**0.034**
	ablated	-0.002	0.058	0.736	0.082
Average	all	0.055	0.051	0.069	0.053
	localized	**0.083**	**0.061**	**0.043**	**0.036**
	ablated	0.021	0.027	0.336	0.074

L.2 QUANTITATIVE METHODS' CONFIGURATIONS COMPARISON

In Table 25, we provide a full table comparing global (all layers steered) and localized (functional layers steered) configurations for all steering methods.

Table 25: **Effect of layer localization on each steering method.** For every method, we compare its standard variant (steering all diffusion model layers) against our localized variant (applying the intervention only at layers {7,8}). Mean ± standard error pooled across 9 musical concepts and both steering directions (see Tables 26 to 34 in Appendix). The Δ row reports the relative gain from localization, expressed as a % change of the Standard mean (sign chosen so that "+" always denotes improvement): **green** when localization helps and **red** when it hurts, faded when the absolute gain is below the standard error of the difference.

Method	**AUC** (↑)		**Smoothness** (↓)		**Audio Quality** (↑)
	MuQ	CLAP	MuQ	CLAP	
PCI (Görgün et al. (2026))	$0.084_{\pm.019}$	$0.049_{\pm.010}$	$0.185_{\pm.064}$	$0.223_{\pm.066}$	$6.803_{\pm.016}$
+ localized	$0.068_{\pm.016}$	$0.038_{\pm.008}$	$0.277_{\pm.143}$	$0.122_{\pm.017}$	$6.789_{\pm.014}$
Δ	−19%	−22%	−50%	+21%	−0.2%
Text Embeddings (Ezra et al. (2025))	$0.014_{\pm.007}$	$0.011_{\pm.004}$	$0.439_{\pm.224}$	$0.681_{\pm.277}$	$6.690_{\pm.027}$
+ localized	$0.023_{\pm.006}$	$0.014_{\pm.004}$	$0.334_{\pm.123}$	$0.366_{\pm.140}$	$6.780_{\pm.014}$
Δ	+62%	+31%	+25%	+48%	+1.4%
Token Embeddings (Baumann et al. (2025))	$0.035_{\pm.011}$	$0.024_{\pm.006}$	$0.267_{\pm.063}$	$0.211_{\pm.058}$	$6.776_{\pm.023}$
+ localized	$0.032_{\pm.009}$	$0.020_{\pm.006}$	$0.208_{\pm.073}$	$0.282_{\pm.098}$	$6.794_{\pm.013}$
Δ	−8%	−20%	+18%	−33%	+0.3%
FreeSliders (Ezra et al. (2025))	$0.079_{\pm.012}$	$0.049_{\pm.009}$	$0.055_{\pm.007}$	$0.066_{\pm.012}$	$6.790_{\pm.017}$
+ localized	$0.080_{\pm.012}$	$0.049_{\pm.009}$	$0.065_{\pm.014}$	$0.074_{\pm.011}$	$6.790_{\pm.017}$
Δ	+1%	−1%	−18%	−13%	−0.0%
Concept Sliders (Gandikota et al. (2024))	$0.086_{\pm.010}$	$0.067_{\pm.011}$	$0.056_{\pm.006}$	$0.045_{\pm.006}$	$6.762_{\pm.019}$
+ localized	$0.068_{\pm.010}$	$0.050_{\pm.009}$	$0.099_{\pm.037}$	$0.060_{\pm.009}$	$6.776_{\pm.019}$
Δ	−21%	−25%	−77%	−34%	+0.2%
AUSteer (Feng et al. (2026))	$0.081_{\pm.020}$	$0.058_{\pm.013}$	$0.212_{\pm.069}$	$0.065_{\pm.011}$	$6.758_{\pm.014}$
+ localized	$0.096_{\pm.016}$	$0.063_{\pm.012}$	$0.054_{\pm.006}$	$0.050_{\pm.005}$	$6.775_{\pm.018}$
Δ	+19%	+8%	+74%	+22%	+0.3%
CAA (Rimsky et al. (2024))	$0.075_{\pm.016}$	$0.046_{\pm.010}$	$0.092_{\pm.034}$	$0.071_{\pm.015}$	$6.787_{\pm.016}$
+ localized	$0.104_{\pm.019}$	$0.065_{\pm.011}$	$0.063_{\pm.018}$	$0.052_{\pm.010}$	$6.792_{\pm.017}$
Δ	+39%	+40%	+32%	+26%	+0.1%

M AUDIO CONCEPT STEERING BENCHMARK

M.1 METHODS COMPARISON ACROSS CONCEPTS

We compare all steering methods across 9 musical features, in both positive and negative directions. Tables 26 to 34 present evaluation of methods across musical concepts.

Table 26: **Evaluating steering *piano* concept with Ace-Step.** We describe metrics in Section 5.2. Symbols + and - denote the quality of steering towards positive (*add piano*) and negative (*remove piano*) directions, *avg* = (pos + neg) / 2. **Best** and second results are highlighted.

Method (*piano*)	**AUC MuQ** (↑)			**AUC CLAP** (↑)			**Smoothness MuQ** (↓)			**Smoothness CLAP** (↓)			**Avg Quality** (↑)
	+	-	avg	+	-	avg	+	-	avg	+	-	avg	
PCI	0.104	0.004	0.054	0.044	0.001	0.022	0.066	0.418	0.242	0.077	0.501	0.289	6.814
PCI **(loc.)**	0.081	0.007	0.044	0.027	0.010	0.018	0.080	0.547	0.313	0.094	0.270	0.182	6.809
Text Embeddings	0.056	0.005	0.030	0.027	0.022	0.024	0.065	0.415	0.240	0.073	0.051	0.062	6.803
Text Embeddings **(loc.)**	0.060	0.015	0.038	0.020	0.019	0.019	0.069	0.385	0.227	0.075	0.151	0.113	6.796
Token Embeddings	0.127	0.040	0.084	0.049	0.021	0.035	0.084	0.085	0.084	0.078	0.082	0.080	6.793
Token Embeddings **(loc.)**	0.102	0.040	0.071	0.031	0.026	0.028	0.081	0.084	0.083	0.089	0.099	0.094	6.780
FreeSliders	0.085	0.049	0.067	0.033	0.026	0.030	0.049	0.038	0.044	0.052	0.069	0.061	6.787
FreeSliders **(loc.)**	0.093	0.056	0.075	0.036	0.028	0.032	0.038	0.031	**0.035**	0.038	0.066	0.052	6.786
Concept Sliders	0.098	0.059	0.078	0.056	0.032	0.044	0.041	0.029	0.035	0.031	0.057	0.044	6.793
Concept Sliders **(loc.)**	0.071	0.050	0.060	0.039	0.026	0.033	0.056	0.041	0.048	0.050	0.070	0.060	6.795
AUSteer	0.097	0.000	0.049	0.012	0.029	0.020	0.039	0.653	0.346	0.133	0.097	0.115	6.762
AUSteer **(loc.)**	0.153	0.052	0.102	0.040	0.034	0.037	0.022	0.053	0.037	0.030	0.049	0.040	6.784
CAA	0.060	0.044	0.052	0.026	0.024	0.025	0.038	0.066	0.052	0.063	0.095	0.079	6.785
CAA **(loc.)**	0.141	0.073	0.107	0.061	0.044	0.053	0.033	0.051	0.042	0.032	0.044	**0.038**	6.788
SAE **(loc.)**	0.196	0.084	**0.140**	0.080	0.034	**0.057**	0.025	0.055	0.040	0.028	0.049	0.038	6.791

Table 27: **Evaluating steering *tempo* concept with Ace-Step.** We describe metrics in Section 5.2. Symbols + and - denote the quality of steering towards positive (*fast song*) and negative (*slow song*) directions, *avg* = (pos + neg) / 2. **Best** and second results are highlighted.

Method (*tempo*)	**AUC MuQ** (↑)			**AUC CLAP** (↑)			**Smoothness MuQ** (↓)			**Smoothness CLAP** (↓)			**Avg Quality** (↑)
	+	-	avg	+	-	avg	+	-	avg	+	-	avg	
PCI	0.052	0.075	0.063	0.008	0.027	0.017	0.081	0.055	0.068	0.190	0.076	0.133	6.746
PCI **(loc.)**	0.047	0.054	0.051	0.007	0.019	0.013	0.081	0.059	0.070	0.204	0.070	0.137	6.742
Text Embeddings	0.015	0.011	0.013	−0.016	0.021	0.002	0.234	0.115	0.175	2.731	0.072	1.402	6.678
Text Embeddings **(loc.)**	0.016	0.015	0.016	−0.001	−0.001	−0.001	0.208	0.119	0.163	0.404	0.684	0.544	6.782
Token Embeddings	0.012	0.025	0.018	0.011	0.020	0.015	0.166	0.150	0.158	0.082	0.088	0.085	6.741
Token Embeddings **(loc.)**	0.025	0.012	0.019	0.013	0.016	0.014	0.097	0.118	0.107	0.077	0.094	0.086	6.778
FreeSliders	0.071	0.076	0.073	0.014	0.031	0.023	0.029	0.033	**0.031**	0.078	0.046	0.062	6.759
FreeSliders **(loc.)**	0.080	0.075	0.078	0.015	0.029	0.022	0.034	0.028	0.031	0.080	0.073	0.077	6.762
Concept Sliders	0.065	0.059	0.062	0.014	0.019	0.017	0.050	0.030	0.040	0.119	0.054	0.087	6.760
Concept Sliders **(loc.)**	0.063	0.061	0.062	0.016	0.017	0.017	0.049	0.029	0.039	0.132	0.060	0.096	6.760
AUSteer	0.095	0.090	0.092	0.020	0.038	0.029	0.038	0.042	0.040	0.070	0.046	0.058	6.769
AUSteer **(loc.)**	0.077	0.082	0.080	0.019	0.036	0.028	0.060	0.038	0.049	0.064	0.045	0.054	6.769
CAA	0.095	0.073	0.084	0.022	0.036	**0.029**	0.032	0.039	0.036	0.075	0.032	0.054	6.781
CAA **(loc.)**	0.111	0.099	0.105	0.023	0.028	0.025	0.029	0.037	0.033	0.042	0.040	**0.041**	6.786
SAE **(loc.)**	0.121	0.100	**0.110**	0.021	0.014	0.018	0.048	0.032	0.040	0.054	0.072	0.063	6.761

Table 28: **Evaluating steering *mood* concept with Ace-Step.** We describe metrics in Section 5.2. Symbols + and - denote the quality of steering towards positive (*happy song*) and negative (*sad song*) directions, *avg* = (pos + neg) / 2. **Best** and second results are highlighted.

Method (*mood*)	**AUC MuQ** (↑)			**AUC CLAP** (↑)			**Smoothness MuQ** (↓)			**Smoothness CLAP** (↓)			**Avg Quality** (↑)
	+	-	avg	+	-	avg	+	-	avg	+	-	avg	
PCI	0.028	0.027	0.027	0.020	0.049	0.034	0.102	0.097	0.099	0.103	0.104	0.103	6.766
PCI **(loc.)**	0.008	0.031	0.020	0.008	0.044	0.026	0.178	0.077	0.127	0.103	0.069	0.086	6.748
Text Embeddings	0.010	−0.001	0.004	−0.002	0.019	0.008	0.281	0.181	0.231	0.573	0.062	0.318	6.693
Text Embeddings **(loc.)**	0.007	0.009	0.008	0.000	0.027	0.014	0.178	0.184	0.181	0.407	0.084	0.246	6.764
Token Embeddings	0.002	0.001	0.001	−0.005	0.026	0.010	0.462	0.265	0.364	0.938	0.116	0.527	6.710
Token Embeddings **(loc.)**	−0.005	0.001	−0.002	−0.009	0.017	0.004	–	0.291	–	1.839	0.102	0.971	6.755
FreeSliders	0.034	0.026	0.030	0.024	0.043	0.033	0.052	0.058	0.055	0.050	0.028	0.039	6.769
FreeSliders **(loc.)**	0.028	0.025	0.027	0.015	0.043	0.029	0.054	0.064	0.059	0.129	0.051	0.090	6.766
Concept Sliders	0.036	0.065	0.051	0.040	0.099	0.070	0.066	0.051	0.058	0.044	0.024	0.034	6.716
Concept Sliders **(loc.)**	0.024	0.043	0.033	0.032	0.075	0.054	0.103	0.046	0.074	0.052	0.025	0.039	6.745
AUSteer	0.007	0.044	0.026	0.031	0.097	0.064	0.459	0.058	0.259	0.093	0.038	0.066	6.723
AUSteer **(loc.)**	0.037	0.069	**0.053**	0.045	0.105	**0.075**	0.091	0.040	0.066	0.053	0.025	0.039	6.732
CAA	0.043	0.048	0.046	0.049	0.088	0.069	0.069	0.059	0.064	0.038	0.034	0.036	6.739
CAA **(loc.)**	0.049	0.048	0.049	0.046	0.095	0.071	0.042	0.058	**0.050**	0.035	0.019	**0.027**	6.746
SAE **(loc.)**	0.023	0.070	0.046	0.036	0.091	0.063	0.090	0.067	0.078	0.059	0.075	0.067	6.735

Table 29: **Evaluating steering *vocal gender* concept with Ace-Step.** We describe metrics in Section 5.2. Symbols + and - denote the quality of steering towards positive (*female vocal*) and negative (*male vocal*) directions, *avg* = (pos + neg) / 2. **Best** and second results are highlighted.

Method (*vocal gender*)	**AUC MuQ** (↑)			**AUC CLAP** (↑)			**Smoothness MuQ** (↓)			**Smoothness CLAP** (↓)			**Avg Quality** (↑)
	+	-	avg	+	-	avg	+	-	avg	+	-	avg	
PCI	0.045	0.010	0.027	0.132	0.058	0.095	0.088	0.249	0.168	0.053	0.109	0.081	6.882
PCI **(loc.)**	0.021	0.015	0.018	0.101	0.023	0.062	0.123	0.316	0.219	0.046	0.105	0.076	6.845
Text Embeddings	−0.075	0.041	−0.017	−0.017	0.028	0.006	–	0.096	–	4.537	0.074	2.306	6.693
Text Embeddings **(loc.)**	−0.008	0.001	−0.004	0.025	0.007	0.016	0.187	0.563	0.375	0.088	0.119	0.103	6.821
Token Embeddings	−0.005	−0.008	−0.007	0.004	0.010	0.007	0.597	0.454	0.525	0.160	0.241	0.201	6.882
Token Embeddings **(loc.)**	−0.008	0.005	−0.002	0.025	0.001	0.013	1.021	0.374	0.698	0.143	0.474	0.308	6.850
FreeSliders	0.044	0.025	0.035	0.106	0.068	0.087	0.064	0.136	0.100	0.043	0.075	0.059	6.864
FreeSliders **(loc.)**	0.047	0.027	0.037	0.106	0.072	0.089	0.084	0.217	0.151	0.043	0.087	0.065	6.862
Concept Sliders	0.072	0.081	**0.076**	0.112	0.060	0.086	0.051	0.027	**0.039**	0.024	0.038	**0.031**	6.842
Concept Sliders **(loc.)**	0.045	0.043	0.044	0.084	0.034	0.059	0.051	0.100	0.075	0.030	0.068	0.049	6.865
AUSteer	−0.002	0.040	0.019	0.135	0.101	**0.118**	0.853	0.065	0.459	0.049	0.079	0.064	6.794
AUSteer **(loc.)**	0.018	0.044	0.031	0.108	0.078	0.093	0.141	0.044	0.092	0.042	0.079	0.060	6.845
CAA	0.020	0.005	0.012	0.104	0.070	0.087	0.138	0.574	0.356	0.034	0.082	0.058	6.848
CAA **(loc.)**	0.029	0.007	0.018	0.116	0.085	0.101	0.068	0.334	0.201	0.034	0.081	0.057	6.864
SAE **(loc.)**	0.044	0.018	0.031	0.122	0.068	0.095	0.066	0.094	0.080	0.038	0.072	0.055	6.852

Table 30: **Evaluating steering *vocal style* concept with Ace-Step.** We describe metrics in Section 5.2. Symbols + and - denote the quality of steering towards positive (*rap vocal*) and negative (*sing vocal*) directions, *avg* = (pos + neg) / 2. **Best** and second results are highlighted.

Method (*vocal style*)	**AUC MuQ** (↑)			**AUC CLAP** (↑)			**Smoothness MuQ** (↓)			**Smoothness CLAP** (↓)			**Avg Quality** (↑)
	+	-	avg	+	-	avg	+	-	avg	+	-	avg	
PCI	0.155	−0.010	0.073	0.056	−0.001	0.028	0.068	1.277	0.672	0.074	1.250	0.662	6.823
PCI **(loc.)**	0.136	−0.010	0.063	0.053	−0.011	0.021	0.069	2.739	1.404	0.091	–	–	6.800
Text Embeddings	−0.005	−0.014	−0.010	−0.026	0.006	−0.010	1.029	2.983	2.006	0.630	0.122	0.376	6.674
Text Embeddings **(loc.)**	0.036	−0.002	0.017	0.024	−0.010	0.007	0.077	2.508	1.293	0.246	2.549	1.398	6.811
Token Embeddings	0.045	0.007	0.026	0.032	0.004	0.018	0.089	0.190	0.139	0.081	0.197	0.139	6.877
Token Embeddings **(loc.)**	0.043	0.005	0.024	0.033	−0.005	0.014	0.103	0.365	0.234	0.085	0.324	0.204	6.842
FreeSliders	0.118	0.019	0.068	0.051	0.011	0.031	0.033	0.093	0.063	0.049	0.254	0.152	6.842
FreeSliders **(loc.)**	0.121	0.019	0.070	0.052	0.011	0.031	0.031	0.087	0.059	0.034	0.216	0.125	6.843
Concept Sliders	0.148	0.031	0.089	0.099	0.019	**0.059**	0.047	0.055	0.051	0.046	0.043	0.044	6.822
Concept Sliders **(loc.)**	0.107	0.016	0.061	0.049	0.005	0.027	0.043	0.126	0.084	0.057	0.156	0.106	6.824
AUSteer	0.094	0.044	0.069	0.058	0.042	0.050	0.045	0.033	0.039	0.046	0.035	**0.041**	6.813
AUSteer **(loc.)**	0.106	0.037	0.072	0.030	0.023	0.027	0.026	0.069	0.048	0.070	0.046	0.058	6.815
CAA	0.122	0.023	0.073	0.046	0.020	0.033	0.045	0.073	0.059	0.050	0.073	0.061	6.822
CAA **(loc.)**	0.163	0.020	0.091	0.055	0.012	0.034	0.034	0.081	0.057	0.053	0.103	0.078	6.831
SAE **(loc.)**	0.179	0.042	**0.110**	0.053	0.021	0.037	0.021	0.037	**0.029**	0.045	0.085	0.065	6.832

Table 31: **Evaluating steering *violin* concept with Ace-Step.** We describe metrics in Section 5.2. Symbols + and - denote the quality of steering towards positive (*add violin*) and negative (*remove violin*) directions, *avg* = (pos + neg) / 2. **Best** and second results are highlighted.

Method (*violin*)	**AUC MuQ** (↑)			**AUC CLAP** (↑)			**Smoothness MuQ** (↓)			**Smoothness CLAP** (↓)			**Avg Quality** (↑)
	+	-	avg	+	-	avg	+	-	avg	+	-	avg	
PCI	0.148	0.018	0.083	0.065	0.011	0.038	0.076	0.239	0.158	0.075	0.186	0.130	6.828
PCI **(loc.)**	0.127	0.019	0.073	0.051	0.014	0.033	0.064	0.146	0.105	0.076	0.165	0.120	6.833
Text Embeddings	0.088	−0.006	0.041	0.045	0.007	0.026	0.055	0.472	0.263	0.045	0.105	0.075	6.824
Text Embeddings **(loc.)**	0.089	0.004	0.047	0.033	0.016	0.025	0.051	0.473	0.262	0.057	0.126	0.092	6.822
Token Embeddings	0.146	0.027	0.087	0.062	0.018	0.040	0.037	0.175	0.106	0.041	0.104	0.072	6.817
Token Embeddings **(loc.)**	0.113	0.024	0.069	0.049	0.023	0.036	0.044	0.165	0.105	0.048	0.137	0.092	6.814
FreeSliders	0.100	0.053	0.077	0.050	0.027	0.038	0.040	0.074	0.057	0.036	0.051	0.043	6.807
FreeSliders **(loc.)**	0.114	0.058	0.086	0.054	0.027	0.040	0.025	0.056	0.041	0.033	0.069	0.051	6.811
Concept Sliders	0.160	0.052	0.106	0.060	0.046	**0.053**	0.026	0.084	0.055	0.039	0.045	0.042	6.791
Concept Sliders **(loc.)**	0.137	0.048	0.092	0.052	0.033	0.043	0.030	0.062	0.046	0.062	0.062	0.062	6.810
AUSteer	0.070	−0.006	0.032	0.006	0.020	0.013	0.049	1.097	0.573	0.150	0.085	0.118	6.785
AUSteer **(loc.)**	0.151	0.056	0.104	0.053	0.041	0.047	0.018	0.081	0.050	0.034	0.056	0.045	6.804
CAA	0.071	0.025	0.048	−0.007	0.013	0.003	0.043	0.062	0.053	0.204	0.107	0.155	6.798
CAA **(loc.)**	0.147	0.068	0.108	0.054	0.050	0.052	0.020	0.046	**0.033**	0.028	0.043	**0.036**	6.811
SAE **(loc.)**	0.181	0.078	**0.130**	0.049	0.037	0.043	0.024	0.042	0.033	0.052	0.033	0.042	6.778

Table 32: **Evaluating steering *guitar type* concept with Ace-Step.** We describe metrics in Section 5.2. Symbols + and - denote the quality of steering towards positive (*acoustic guitar*) and negative (*electric guitar*) directions, *avg* = (pos + neg) / 2. **Best** and second results are highlighted.

Method (*guitar type*)	**AUC MuQ** (↑)			**AUC CLAP** (↑)			**Smoothness MuQ** (↓)			**Smoothness CLAP** (↓)			**Avg Quality** (↑)
	+	-	avg	+	-	avg	+	-	avg	+	-	avg	
PCI	0.146	0.012	0.079	0.073	0.001	0.037	0.059	0.139	0.099	0.059	0.762	0.410	6.730
PCI **(loc.)**	0.117	0.013	0.065	0.049	0.015	0.032	0.056	0.139	0.098	0.067	0.329	0.198	6.723
Text Embeddings	0.010	0.007	0.009	−0.029	0.026	−0.002	0.139	0.305	0.222	–	0.068	–	6.569
Text Embeddings **(loc.)**	0.005	0.005	0.005	−0.020	0.019	−0.001	0.193	0.300	0.247	0.680	0.235	0.457	6.683
Token Embeddings	0.084	0.004	0.044	0.038	−0.007	0.015	0.062	0.646	0.354	0.058	0.679	0.369	6.730
Token Embeddings **(loc.)**	0.061	0.011	0.036	0.011	0.006	0.009	0.059	0.242	0.150	0.126	0.375	0.250	6.728
FreeSliders	0.156	0.019	0.087	0.065	0.018	0.042	0.026	0.114	0.070	0.033	0.148	0.090	6.694
FreeSliders **(loc.)**	0.152	0.018	0.085	0.064	0.017	0.040	0.023	0.218	0.121	0.046	0.193	0.120	6.692
Concept Sliders	0.148	0.030	0.089	0.074	0.042	0.058	0.030	0.137	0.084	0.028	0.063	0.046	6.653
Concept Sliders **(loc.)**	0.139	0.028	0.084	0.060	0.034	0.047	0.036	0.152	0.094	0.037	0.084	0.060	6.660
AUSteer	0.158	0.040	0.099	0.045	0.026	0.036	0.027	0.151	0.089	0.032	0.075	0.053	6.670
AUSteer **(loc.)**	0.110	0.044	0.077	0.039	0.020	0.029	0.025	0.090	0.057	0.036	0.126	0.081	6.657
CAA	0.107	0.014	0.060	0.027	0.010	0.019	0.031	0.216	0.124	0.046	0.210	0.128	6.690
CAA **(loc.)**	0.153	0.027	0.090	0.055	0.016	0.036	0.025	0.119	0.072	0.050	0.196	0.123	6.686
SAE **(loc.)**	0.205	0.069	**0.137**	0.077	0.046	**0.061**	0.016	0.099	**0.057**	0.020	0.052	**0.036**	6.649

Table 33: **Evaluating steering *jazz genre* concept with Ace-Step.** We describe metrics in Section 5.2. Symbols + and - denote the quality of steering towards positive (*jazz song*) and negative (*rock song*) directions, *avg* = (pos + neg) / 2. **Best** and second results are highlighted.

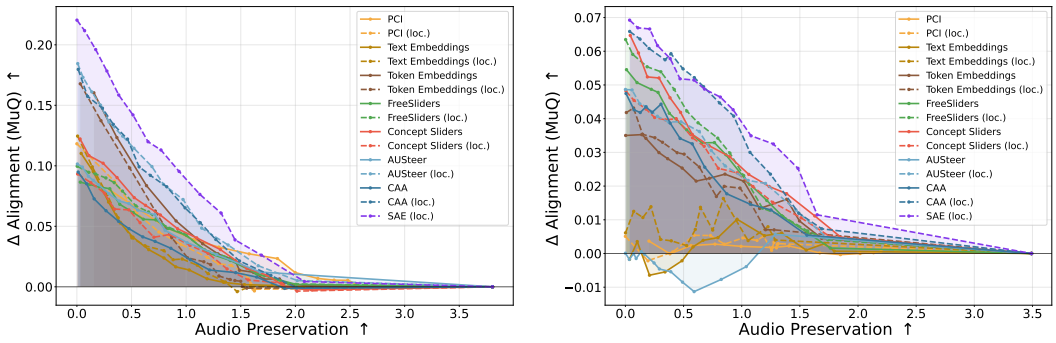
Method (*jazz genre*)	**AUC MuQ** (↑)			**AUC CLAP** (↑)			**Smoothness MuQ** (↓)			**Smoothness CLAP** (↓)			**Avg Quality** (↑)
	+	-	avg	+	-	avg	+	-	avg	+	-	avg	
PCI	0.363	0.049	0.206	0.176	0.010	0.093	0.032	0.140	0.086	0.042	0.231	0.137	6.822
PCI **(loc.)**	0.297	0.038	0.168	0.146	0.012	0.079	0.034	0.114	0.074	0.040	0.160	0.100	6.801
Text Embeddings	0.014	0.028	0.021	−0.025	0.067	0.021	0.352	0.092	0.222	1.184	0.065	0.624	6.633
Text Embeddings **(loc.)**	0.058	0.031	0.045	0.005	0.042	0.024	0.107	0.077	0.092	0.245	0.058	0.151	6.760
Token Embeddings	−0.011	0.031	0.010	−0.006	0.043	0.019	0.965	0.170	0.568	0.697	0.076	0.386	6.706
Token Embeddings **(loc.)**	0.006	0.038	0.022	−0.006	0.015	0.004	0.318	0.091	0.205	0.830	0.135	0.482	6.794
FreeSliders	0.145	0.115	0.130	0.086	0.068	0.077	0.035	0.046	**0.041**	0.035	0.064	0.049	6.783
FreeSliders **(loc.)**	0.138	0.108	0.123	0.078	0.067	0.072	0.033	0.055	0.044	0.035	0.073	0.054	6.782
Concept Sliders	0.189	0.114	0.152	0.157	0.106	0.132	0.041	0.074	0.057	0.021	0.052	0.037	6.727
Concept Sliders **(loc.)**	0.170	0.098	0.134	0.119	0.081	0.100	0.026	0.055	0.041	0.020	0.041	**0.030**	6.753
AUSteer	0.240	0.134	0.187	0.112	0.079	0.095	0.032	0.073	0.053	0.027	0.039	0.033	6.761
AUSteer **(loc.)**	0.250	0.120	0.185	0.132	0.078	0.105	0.034	0.077	0.056	0.025	0.043	0.034	6.786
CAA	0.216	0.120	0.168	0.091	0.056	0.073	0.039	0.063	0.051	0.030	0.040	0.035	6.810
CAA **(loc.)**	0.270	0.142	0.206	0.146	0.083	0.115	0.026	0.060	0.043	0.019	0.051	0.035	6.811
SAE **(loc.)**	0.318	0.126	**0.222**	0.170	0.103	**0.137**	0.051	0.120	0.085	0.041	0.078	0.060	6.776

Table 34: **Evaluating steering *classical genre* concept with Ace-Step.** We describe metrics in Section 5.2. Symbols + and - denote the quality of steering towards positive (*classical song*) and negative (*electronic song*) directions, *avg* = (pos + neg) / 2. **Best** and second results are highlighted.

Method (*classical genre*)	**AUC MuQ** (↑)			**AUC CLAP** (↑)			**Smoothness MuQ** (↓)			**Smoothness CLAP** (↓)			**Avg Quality** (↑)
	+	-	avg	+	-	avg	+	-	avg	+	-	avg	
PCI	0.239	0.049	0.144	0.117	0.028	0.072	0.036	0.108	0.072	0.038	0.088	0.063	6.813
PCI **(loc.)**	0.206	0.019	0.113	0.102	0.013	0.057	0.034	0.133	0.084	0.043	0.104	0.073	6.798
Text Embeddings	0.075	−0.006	0.034	−0.009	0.055	0.023	0.068	0.242	0.155	0.515	0.051	0.283	6.642
Text Embeddings **(loc.)**	0.067	−0.000	0.033	0.010	0.044	0.027	0.114	0.215	0.164	0.319	0.068	0.194	6.783
Token Embeddings	0.079	0.022	0.050	0.055	0.064	0.060	0.043	0.163	0.103	0.035	0.052	0.044	6.732
Token Embeddings **(loc.)**	0.074	0.025	0.050	0.056	0.051	0.053	0.046	0.118	0.082	0.037	0.053	0.045	6.805
FreeSliders	0.176	0.106	0.141	0.088	0.081	0.084	0.029	0.041	0.035	0.039	0.033	0.036	6.810
FreeSliders **(loc.)**	0.171	0.100	0.136	0.088	0.080	0.084	0.043	0.041	0.042	0.039	0.031	0.035	6.809
Concept Sliders	0.111	0.035	0.073	0.085	0.092	0.089	0.053	0.118	0.086	0.048	0.027	0.037	6.756
Concept Sliders **(loc.)**	0.088	−0.000	0.044	0.075	0.075	0.075	0.047	0.739	0.393	0.051	0.020	0.035	6.776
AUSteer	0.205	0.106	0.156	0.095	0.098	0.097	0.033	0.061	0.047	0.041	0.027	0.034	6.748
AUSteer **(loc.)**	0.204	0.115	0.159	0.133	0.112	**0.122**	0.026	0.042	0.034	0.047	0.034	0.041	6.788
CAA	0.177	0.091	0.134	0.075	0.085	0.080	0.022	0.050	0.036	0.025	0.032	**0.029**	6.811
CAA **(loc.)**	0.211	0.119	**0.165**	0.109	0.091	0.100	0.030	0.037	**0.033**	0.034	0.038	0.036	6.806
SAE **(loc.)**	0.179	0.095	0.137	0.112	0.125	0.118	0.043	0.100	0.072	0.073	0.064	0.068	6.752

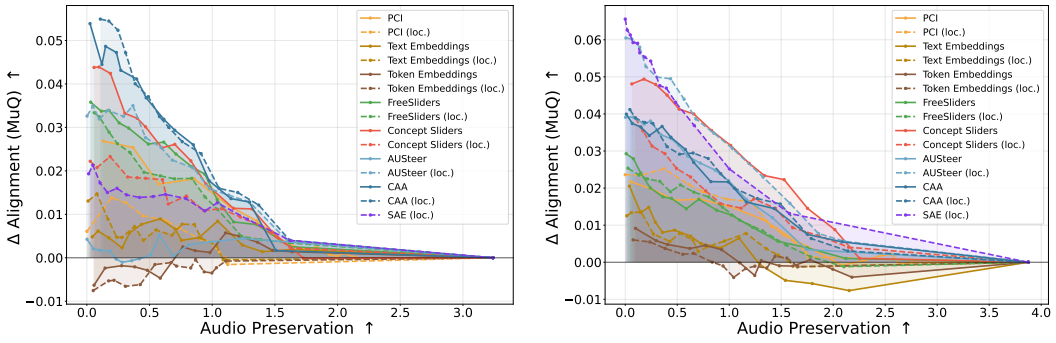
M.2 ALIGNMENT–PRESERVATION CURVES

Figures 16 to 24 visualize the alignment–preservation curves whose areas yield the AUC scores in Tables 26 to 34. For all nine benchmark concepts and both steering directions, each panel sweeps the steering strength α for every evaluated method, plotting the sign-corrected alignment delta $\Delta a(\alpha)$ against the preservation score $p(\alpha)$ (Section 5.2). The AUC metric value reported in the tables is the area under the corresponding curve, integrated up to the shared cutoff $P_{\max}$. Curves closer to the upper-right corner represent better gain in alignment at matched preservation.



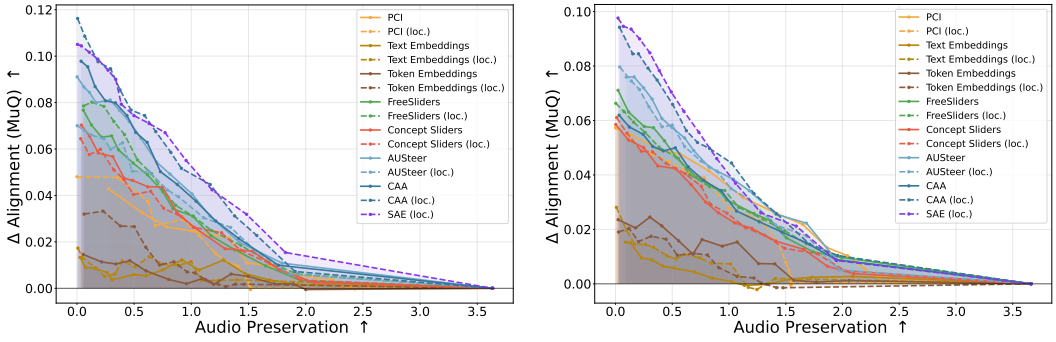
(a) add piano (+) (b) remove piano (–)

Figure 16: **Alignment–preservation curves for *piano* steering.** Used for computing AUC values in Tab. 26.



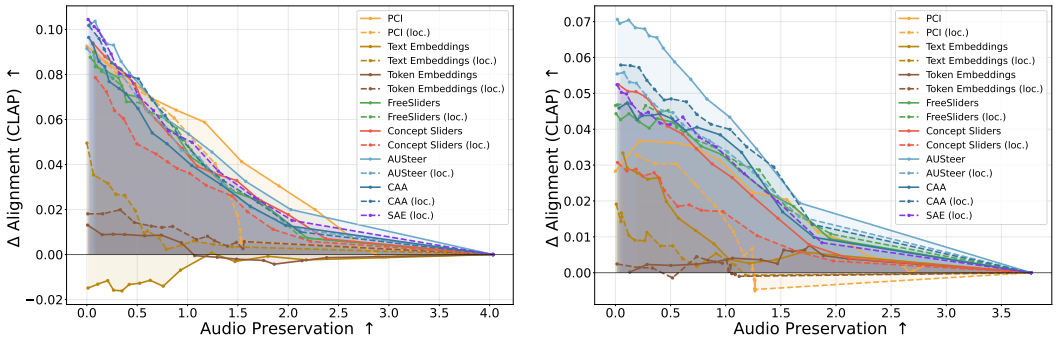
(a) brighter mood (+) (b) darker mood (–)

Figure 17: **Alignment–preservation curves for *mood* steering.** Used for computing AUC values in Tab. 28.



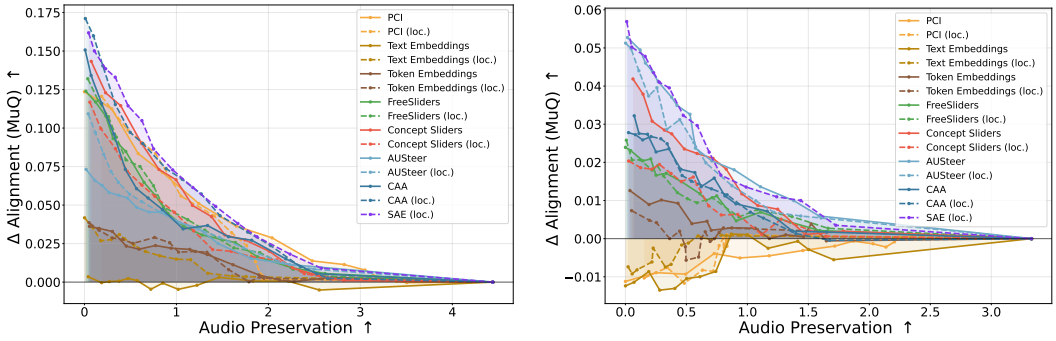
(a) faster tempo (+) (b) slower tempo (–)

Figure 18: **Alignment–preservation curves for *tempo* steering.** Used for computing AUC values in Tab. 27.



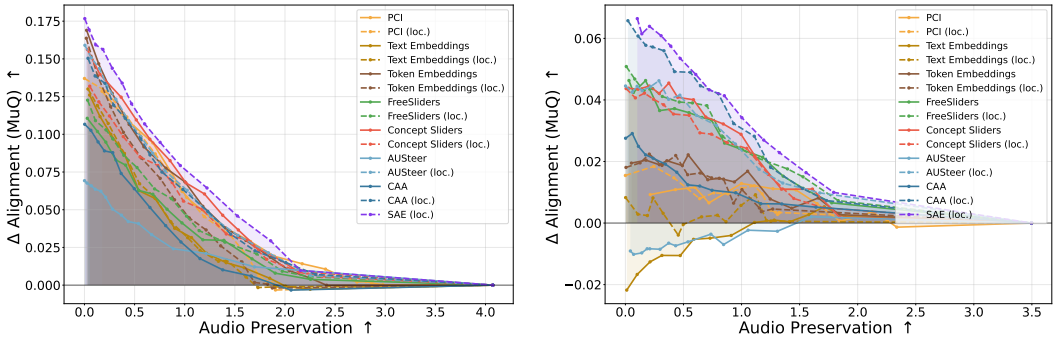
(a) female vocals (+) (b) male vocals (–)

Figure 19: **Alignment–preservation curves for *vocal gender* steering.** Used for computing AUC values in Tab. 29.



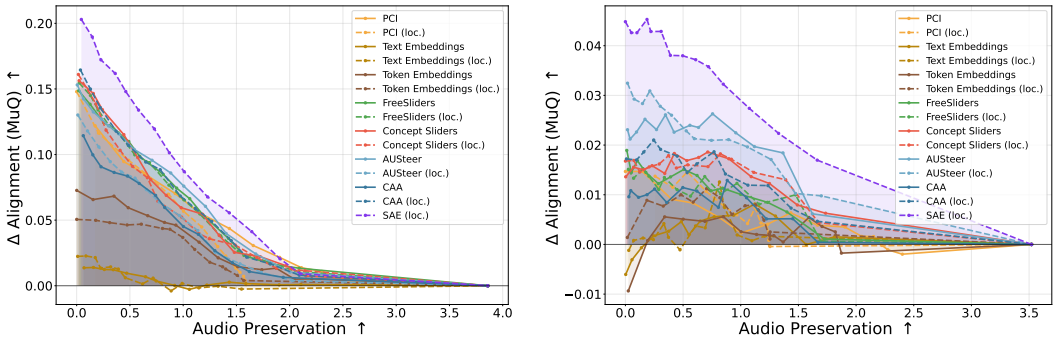
(a) rap vocal style (+) (b) singing vocal style (–)

Figure 20: **Alignment–preservation curves for *vocal style* steering.** Used for computing AUC values in Tab. 30.



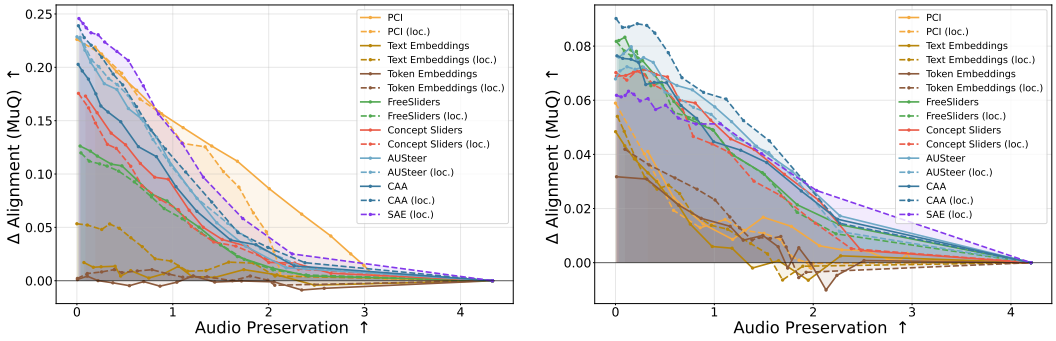
(a) add violin (+) (b) remove violin (–)

Figure 21: **Alignment–preservation curves for *violin* steering.** Used for computing AUC values in Tab. 31.



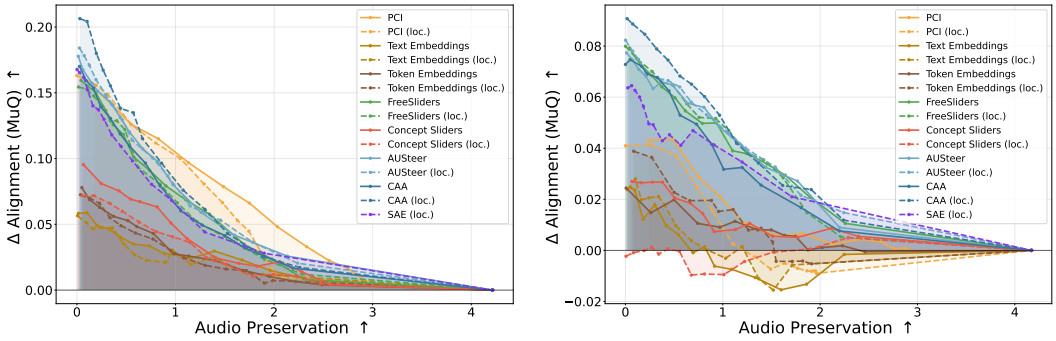
(a) acoustic guitar (+) (b) electric guitar (–)

Figure 22: **Alignment–preservation curves for *guitar type* steering.** Used for computing AUC values in Tab. 32.



(a) jazz genre (+) (b) rock genre (–)

Figure 23: **Alignment–preservation curves for *jazz genre* steering.** Used for computing AUC values in Tab. 33.



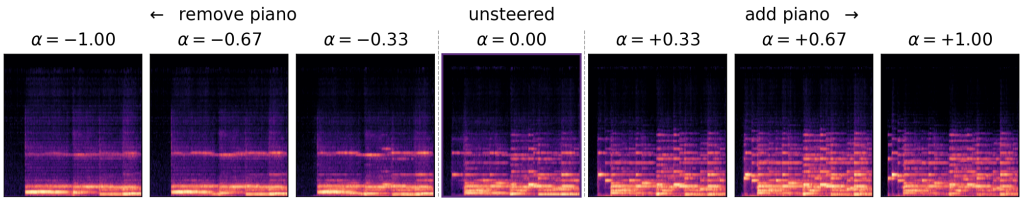
(a) classical genre (+) (b) electronic genre (–)

Figure 24: **Alignment–preservation curves for *classical genre* steering.** Used for computing AUC values in Tab. 34.

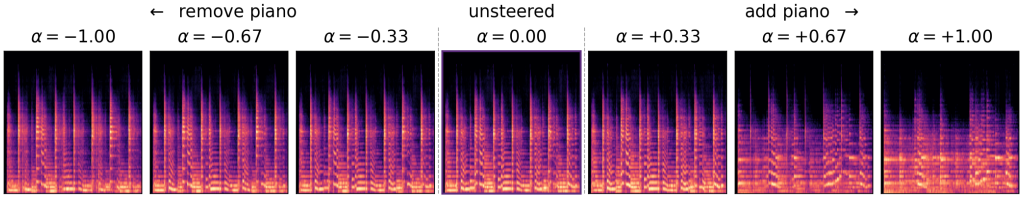
N SPECTROGRAMS

We complement our results with qualitative examples in Figure 25, showing mel-spectrograms of generated audio modulated via SAE steering. For each prompt, we sweep the steering coefficient α over seven evenly spaced values, with $\alpha = 0$ corresponding to the unsteered track.

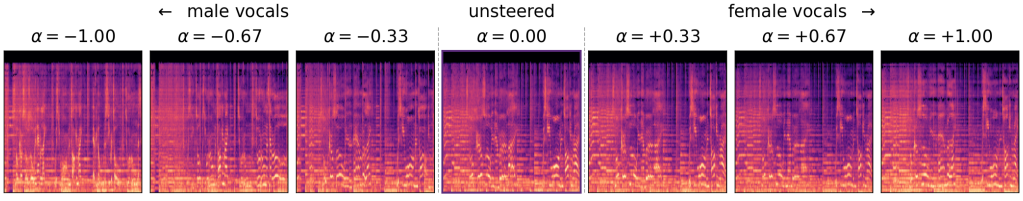
The spectrograms make it visually apparent how SAE interventions controllably edit the time–frequency content of the generation as α is modulated.



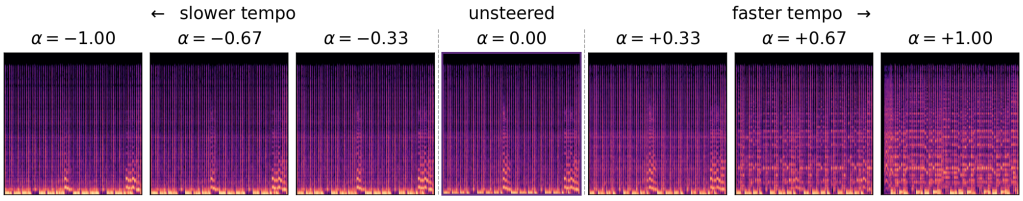
(a) Piano: “calm coldplay style song with slight piano”



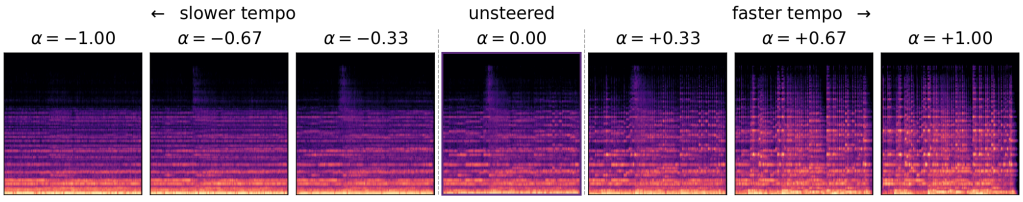
(b) Piano: “Latin Spanish Music with subtle guitar”



(c) Vocal: “jazz trio with brushed drums, walking upright bass, and smoky piano”



(d) Tempo: “Latin music with percussion and guitar”



(e) Tempo: “dramatic opera, lush orchestral strings, and swelling brass”

Figure 25: **Mel spectrograms for audio concept modulation.** Each row presents the result of the steering audio generation process with SAEs. The central column ($\alpha = 0$) is the unsteered baseline.

O DETAILS ON MULTI-CONCEPT STEERING

We evaluate activation steering methods on multi-concept steering with nine combinations, including

1. five pairs: *piano with violin* (**P+V**), *piano with female vocals* (**P+FV**), *acoustic guitar with female vocals* (**A+FV**), *piano with male vocals* (**P+$\overline{\text{FV}}$**), and *jazz with slower tempo* (**J+$\overline{\text{T}}$**);
2. four triples: *piano with violin and jazz* (**P+V+J**), *acoustic guitar with female vocals and brighter mood* (**A+FV+M**), *piano with violin and slower tempo* (**P+V+$\overline{\text{T}}$**), and *acoustic guitar with male vocals and darker mood* (**A+$\overline{\text{FV}}$+$\overline{\text{M}}$**).

For CAA, AUSteer, and SAE, we construct the multi-concept steering vector as the unit-weight summation of the single-concept vectors. Negative direction vectors (e.g., male vocal or darker mood) are constructed by a sign flip of the corresponding positive direction vectors.

All generations share the main benchmark settings (Ace-Step model with $T = 30$ inference steps and CFG: $\omega = 5$). The configuration consists of 100 tracks, each 30 seconds long. We sweep 16 alphas linearly in one direction. For each concept combination evaluation, we calculate the alignment separately for each constituent concept with CLAP or MUQ-T, and report the mean across concepts. Steering strengths are calibrated to match the average preservation cutoff for the main benchmark.

Results in Table 35 compare activation steering methods in non-localized and localized variants across diverse concept combinations for both (MUQ and CLAP) text-audio alignment models. Localized steering variants nearly always offer better steering performance: CAA_{loc} improves on CAA_{all} on all nine combinations in both metrics, and $\text{AUSteer}_{\text{loc}}$ improves on $\text{AUSteer}_{\text{all}}$ on eight of nine under MUQ-T. In 15 out of 18 cases, localized steering offers the best performance.

Table 35: **Multi-concept steering performance measured with average Area Under LPAPS-MUQ and LPAPS-CLAP Curves.** Best result per combination of concepts is **bolded**. Single column denotes steering on combinations of concepts, where **P** = Piano, **V** = Violin, **FV** = Female Vocal, **A** = Acoustic Guitar, **J** = Jazz Music, **M** = Brighter Mood, $\overline{\textbf{FV}}$ = Male Vocal, $\overline{\textbf{T}}$ = Slower Tempo, and $\overline{\textbf{M}}$ = Darker Mood.

Method	Steering Two Concepts					Steering Three Concepts			
	P + V	**P + FV**	**A + FV**	**P + $\overline{\text{FV}}$**	**J + $\overline{\text{T}}$**	**P + V + J**	**A + FV + M**	**P + V + $\overline{\text{T}}$**	**A + $\overline{\text{FV}}$ + $\overline{\text{M}}$**
AUC MuQ									
$\text{AUSteer}_{\text{all}}$	+0.057	+0.001	+0.022	+0.013	+0.090	+0.062	+0.005	+0.048	+0.045
CAA_{all}	+0.048	+0.011	+0.028	+0.008	+0.068	+0.047	+0.026	+0.029	+0.032
$\text{AUSteer}_{\text{loc}}$	+0.099	+0.018	+0.029	**+0.039**	+0.086	+0.088	+0.021	+0.069	+0.051
CAA_{loc}	+0.094	**+0.034**	**+0.052**	+0.027	+0.103	+0.081	**+0.035**	+0.061	+0.039
SAE_{loc}	**+0.114**	+0.027	+0.041	+0.034	**+0.110**	**+0.100**	+0.018	**+0.081**	**+0.063**
AUC CLAP									
$\text{AUSteer}_{\text{all}}$	-0.002	+0.047	**+0.042**	**+0.051**	+0.035	+0.012	+0.028	+0.001	**+0.054**
CAA_{all}	-0.003	+0.037	+0.019	+0.036	+0.036	-0.002	+0.036	-0.003	+0.044
$\text{AUSteer}_{\text{loc}}$	+0.022	+0.046	+0.026	+0.037	+0.050	+0.021	+0.036	+0.010	+0.050
CAA_{loc}	**+0.027**	+0.048	+0.032	+0.049	+0.050	**+0.029**	**+0.047**	+0.013	+0.047
SAE_{loc}	+0.018	**+0.049**	+0.041	+0.045	**+0.056**	+0.029	+0.027	**+0.014**	+0.050