TADA! TUNING AUDIO DIFFUSION MODELS THROUGH ACTIVATION STEERING

Łukasz Staniszewski[1,2*] **Katarzyna Zaleska**[1] **Mateusz Modrzejewski**[1] **Kamil Deja**[1,2]
[1]Warsaw University of Technology [2]IDEAS Research Institute

Code Audio Examples

ABSTRACT

Audio diffusion models can synthesize high-fidelity music from text, yet their internal mechanisms for representing high-level concepts remain poorly understood. In this work, we use activation patching to demonstrate that distinct semantic musical concepts, such as the presence of specific instruments, vocals, or genre characteristics, are controlled by a small, shared subset of attention layers in state-of-the-art audio diffusion architectures. Next, we demonstrate that applying Contrastive Activation Addition and Sparse Autoencoders in these layers enables more precise control over the generated audio, indicating a direct benefit of the specialization phenomenon. By steering activations of the identified layers, we can alter specific musical elements with high precision, such as modulating tempo or changing a track's mood.

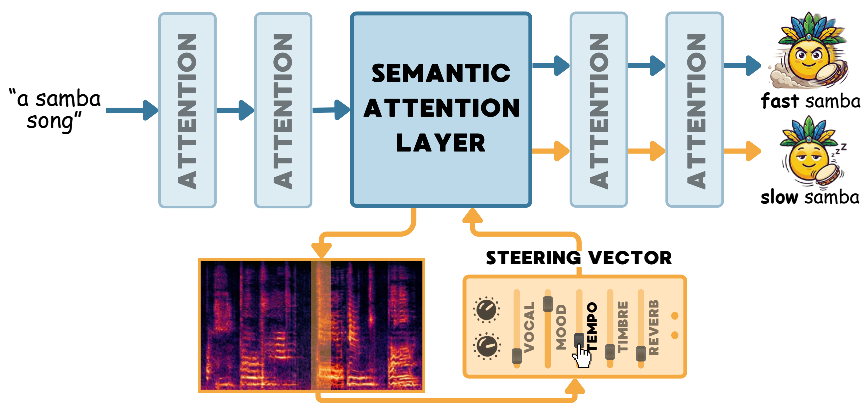

Figure 1: **We study localized steering in Audio Diffusion Models.** By localizing functional layers, we enable precise steering of generations with Contrastive Activation Addition and Sparse Autoencoders.

1 INTRODUCTION

Recent advancements in generative audio have led to Diffusion Models (DMs) capable of synthesizing high-fidelity music from textual descriptions (Gong et al., 2025). While impressive, a significant limitation of those methods is that interaction relies solely on prompting, which acts as a relatively blunt instrument. While a user can ask for "a samba song," prompts lack the precision needed for subtle creative adjustments. For instance, it is nearly impossible to use text to express that a track should have a slightly slower tempo or a marginally lower vocal pitch without triggering the model to regenerate a completely different song. This creates a significant gap for creators who need smooth, precise control over the text-to-music model that goes beyond the limitations of language.

*Corresponding author: luks.staniszewski@gmail.com

Behind this limitation lies an architectural challenge: current audio DMs operate as "black boxes" that entangle various musical skills and semantic attributes within a complex set of millions of parameters. Because these internal mechanisms remain opaque, researchers and creators cannot easily isolate or adjust specific characteristics without affecting the entire global composition.

In this work, we shed light on the inner workings of these audio diffusion models. Drawing inspiration from interpretability methods for language (Meng et al., 2022) and vision (Basu et al., 2024a; Staniszewski et al., 2025), we localize the functional components responsible for generating specific audio concepts across state-of-the-art text-to-music models. Our investigation reveals that, surprisingly, semantic music features such as the presence of a male or female vocalist, instruments, genre, mood, or tempo are governed by remarkably small and specialized subsets of attention layers.

As illustrated in Fig. 1, we build on this observation by employing activation steering techniques, such as Contrastive Activation Addition (CAA (Rimsky et al., 2024)) within localized layers, offering a new tool for controllable generation. We show that restricting interventions to identified functional layers yields significantly higher precision and control than baselines that either apply steering to all layers or to the non-functional set exclusively. Our experiments confirm that this targeted approach effectively modulates attributes such as tempo, mood, voice gender, or instrument presence, while preserving audio fidelity, thereby avoiding the quality degradation observed with standard steering. Furthermore, this localization enables efficient training of Sparse Autoencoders (SAEs (Olshausen & Field, 1997)) within influential regions, revealing highly semantic and interpretable features. This allows fine-grained control over musical attributes, surpassing the limitations of a coarse text prompting. Our contributions can be summarized as follows:

1. We construct a dataset of counterfactual prompt pairs spanning diverse musical concepts to assess the role of layers building modern text-to-music diffusion models.
2. We show that semantic musical attributes, such as tempo, vocals, instruments, mood, and genres, are controlled by small, shared subsets of cross-attention layers across diverse diffusion architectures.
3. We leverage localization to apply Contrastive Activation Addition and Sparse Autoencoders for targeted steering, enabling fine-grained control over musical attributes while preserving overall audio quality.
4. We propose a novel evaluation protocol for audio concept steering that jointly quantifies steering effectiveness, smoothness, and audio quality preservation. We show that under this protocol, localized steering achieves state-of-the-art results across all evaluated concepts.

2 Related work

Interpretability and Steering of Generative Models. Recently, mechanistic interpretability has emerged as a growing research direction, aiming to understand the internal computations of large-scale generative 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), 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 find layers controlling text generated on images with Diffusion Transformers (DiTs). Finally, Zarei et al. (2025) show that important attentions in DiTs can be efficiently traced through the magnitude of their outputs.

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) and Rodriguez et al. (2025b) steer generations by training affine maps representing transformations between distributions of activations. Similarly, difference-based methods have been applied in diffusion models' text encoder space (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, 1997) have (Huben et al., 2024; Bricken et al., 2023) been applied to LLMs to decompose activations into sparse, interpretable features in an unsupervised manner by training an autoencoder for reconstruction 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. Yet another approach (Gandikota et al., 2024) to steering is to train low-rank adapters to represent meaningful directions as weight updates.

Interpretability and Control of Audio Models. Recent research has begun adapting interpretability techniques to the audio domain to understand and control internal computations of models. Initial efforts from 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) developed SMITIN, which uses classifier probes to steer attention heads for specific musical traits, Zhao et al. (2026) adapts Recursive Feature Machines for steering, while Facchiano et al. (2025), closest to our work, 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 recent high performance (Gong et al., 2026) of diffusion-based music generators, interpretability and steering of Audio DMs remains largely underexplored. Recent works predominantly focus on the task of audio editing by manipulating attention maps (Xu et al., 2024; Liu et al., 2025; Yang et al., 2026), or combining 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. 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 contrast to these works, we systematically localize functional components within audio diffusion models and demonstrate that restricting activation steering (via CAA or SAEs) to these specific bottlenecks yields significantly higher controllability, smoothness, and fidelity.

3 BACKGROUND & METHODOLOGY

Audio Diffusion Models. Diffusion models (DMs, Dhariwal & Nichol (2021)) learn to reverse a gradual noising process by predicting, at a given timestep t, the noise $\epsilon \sim \mathcal{N}(0, \mathcal{I})$ added to clean data x_0, minimizing $\mathbb{E}_{t,x_0,\epsilon}\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon,\, t)\|_2^2$. Modern audio DMs, such as Ace-Step (Gong et al., 2025), operate in a compressed latent space: an encoder $\mathcal{E}$ maps waveforms to latent representations $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$, where the diffusion process occurs, and a decoder $\mathcal{D}$ reconstructs the latents back to audio. These models employ U-Net (Ronneberger et al., 2015) or 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 of 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., “female vocal”), we define a set of counterfactual prompt pairs $(\mathcal{P}_c, \mathcal{P}_{\tilde{c}})$ where $\mathcal{P}_c$ contains the concept and $\mathcal{P}_{\tilde{c}}$ does not.

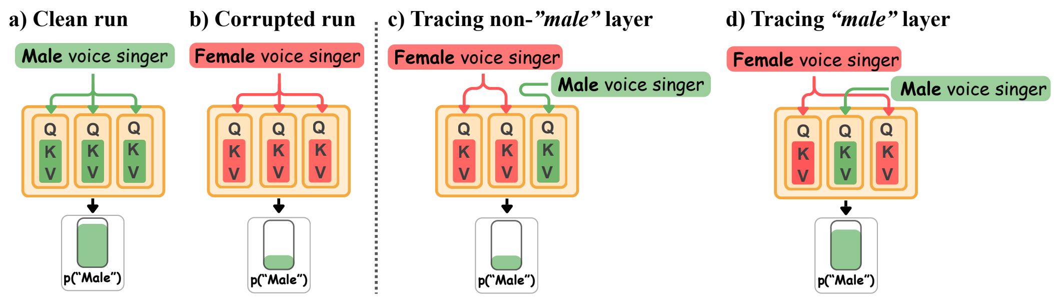

Figure 2: **Layer localization via Activation Patching.** For a given music concept c (e.g., *'male voice'*), we perform (a) a target run with prompt P_c and cache the cross-attention keys and values. In (b) source run, we generate with prompt $P_{\tilde{c}}$, which represents a counterfactual concept (e.g., *'male voice'*) or does not contain c. We patch layer l by substituting cross-attention key ($\mathbf{K}$) and value ($\mathbf{V}$) matrices with those cached from the P_c run. In such a case, other layers receive $P_{\tilde{c}}$. If patching a layer produces audio containing concept c (d), we identify it as a functional layer. Otherwise (c), the layer does not control the concept.

First, we generate audio with the target prompt $\mathcal{P}_c$, caching the cross-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, while generating audio with the source prompt $\mathcal{P}_{\tilde{c}}$, we patch layer l by substituting its keys and values with the cached ones from $\mathcal{P}_c$ run.

Finally, we compare how the intervention affects similarity between the generated audio and the prompt describing the concept c. If the output audio for the patched run exhibits concept c, we identify l as a *functional layer* for c.

Contrastive Activation Addition (CAA). We compute steering vectors following CASteer (Gaintseva et al., 2026) to enable fine-grained control over musical attributes. Given N contrastive prompt pairs $\{(\mathcal{P}_c^{(i)}, \mathcal{P}_{\tilde{c}}^{(i)})\}_{i=1}^{N}$ for concept c, we collect cross-attention outputs and compute the steering vector $\mathbf{v}_c^{\text{CAA}}$ as:

$$\mathbf{v}_c^{\text{CAA}} = \frac{\mathbf{v}_c}{\|\mathbf{v}_c\|_2}, \text{ where } \quad \mathbf{v}_c = \frac{1}{N}\sum_{i=1}^{N}\left(\bar{\mathbf{h}}_c^{(i)} - \bar{\mathbf{h}}_{\tilde{c}}^{(i)}\right), \tag{2}$$

with $\bar{\mathbf{h}}$ denoting cross-attention outputs averaged across temporal frames. During generation, we steer by modifying cross-attention outputs at functional layers as $\mathbf{h}_l' = \text{ReNorm}(\mathbf{h}_l + \alpha \cdot \mathbf{v}_c^{\text{CAA}}, \mathbf{h}_l)$, where $\alpha \in \mathbb{R}$ controls steering strength, with positive values adding and negative values removing the concept. ReNorm is a re-normalization operation, ensuring that the norm of the output activations $\mathbf{h}_l'$ matches the one before intervention $\mathbf{h}_l$:

$$\text{ReNorm}(\mathbf{h}_l', \mathbf{h}_l) = \frac{\mathbf{h}_l'}{\|\mathbf{h}_l'\|_2} \cdot \|\mathbf{h}_l\|_2. \tag{3}$$

Sparse Autoencoders (SAEs). To further discover interpretable features within cross-attention activations, we train a TopK SAEs (Gao et al., 2025; Bussmann et al., 2024) on the functional layers with the highest response during 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):

$$\mathbf{f} = \text{TopK}\left(\mathbf{W}_{\text{enc}}(\mathbf{h} - \mathbf{b}_{\text{pre}})\right), \tag{4}$$

and further reconstructs them:

$$\hat{\mathbf{h}} = \mathbf{W}_{\text{dec}}\mathbf{f} + \mathbf{b}_{\text{pre}}. \tag{5}$$

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:

$$\text{score}(j, c) = \underbrace{\mu_j(\mathcal{P}_c)}_{\text{TF}} \cdot \underbrace{\log\left(1 + \frac{1}{\mu_j(\mathcal{P}_{\tilde{c}}) + \epsilon}\right)}_{\text{IDF}}, \tag{6}$$

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

$$\mathbf{v}_c^{\text{SAE}} = \sum_{j \in \mathcal{F}_c} \mathbf{W}_{\text{dec}}[:, j], \tag{7}$$

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

$$\mathbf{h}'_l = \mathbf{h}_l + \alpha \cdot \mathbf{v}_c^{\text{SAE}}. \tag{8}$$

4 Layer localization

To measure the importance of individual cross-attention layers in text-to-audio models, we construct a dataset of counterfactual prompt pairs. We consider the following musical concepts: *vocal gender* (female vs. male), *tempo* (slow vs. fast), *mood* (happy vs. sad), and categories such as instruments (*drums*, *flute*, *guitar*, *maracas*, *trumpet*, *violin*), and genres (*jazz*, *techno*, *reggae*). For each contrasting pair $(c, \tilde{c})$, we select captions from the MusicCaps (Agostinelli et al., 2023) dataset that contain keywords associated with concept c (e.g., 'female voice') but do not contain 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 use GPT-4 (Achiam et al., 2023) to generate alternatives $\mathcal{P}_{\tilde{c}}$ by replacing concept-associated terms with their counterparts while preserving all other content. Examples of prompt pairs and concept replacements are provided in the App. A.

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 waveforms of 10 seconds (AudioLDM2, Stable Audio Open) and 30 seconds (Ace-Step), using 8 different random seeds per prompt, resulting in 2048 generations per concept. The impact of layer l on concept c is calculated as:

$$\text{I}(l, c) = \frac{\text{sim}(l \leftarrow c, l' \leftarrow \tilde{c}) - \text{sim}(l \leftarrow \tilde{c}, l' \leftarrow \tilde{c})}{\text{sim}(l \leftarrow c, l' \leftarrow c) - \text{sim}(l \leftarrow \tilde{c}, l' \leftarrow \tilde{c})}, \tag{9}$$

where $\text{sim}(l \leftarrow c_1, l' \leftarrow c_2)$ denotes the audio-text similarity between the concept name and a generation where layer l inputs prompt c_1 while all other layers $l' = \mathcal{L} \setminus \{l\}$ receive c_2. The denominator in Eq. (9) serves as a normalization constant representing the upper bound of similarity gain, ensuring $\text{I}(l, c) \in [0, 1]$. We use MuQ (Zhu et al., 2025) for assessing mood, tempo, instruments, and genres, and CLAP (Wu et al., 2022) for distinguishing vocal gender.

Fig. 3 presents the layer-wise impact scores for each model. Across all three architectures, we observe that a small subset of layers concentrates the control over musical concepts. In AudioLDM2, with the U-Net architecture, we localize key components in the decoder, specifically the layers $\{$45,46,51,52$\}$ (4 out of 64 cross-attentions), with a slight contribution from the layers in-between (47–50). In transformer-based architectures, we observe an intense concentration of control in the middle layers (2 out of 24). Namely, in Ace-Step, cross-attentions $\{$7,8$\}$ exhibit strong influence across all concept categories, suggesting their role as a semantic bottleneck. In Stable Audio Open, similar concentration arises in layers $\{$12,13$\}$. These findings suggest that audio diffusion models develop interpretable, functionally specialized layers that are shared across various audio concepts. Additionally, our experiments indicate that such **specialization phenomenon** is not limited to a single model, but is a general property of text-to-music DMs.

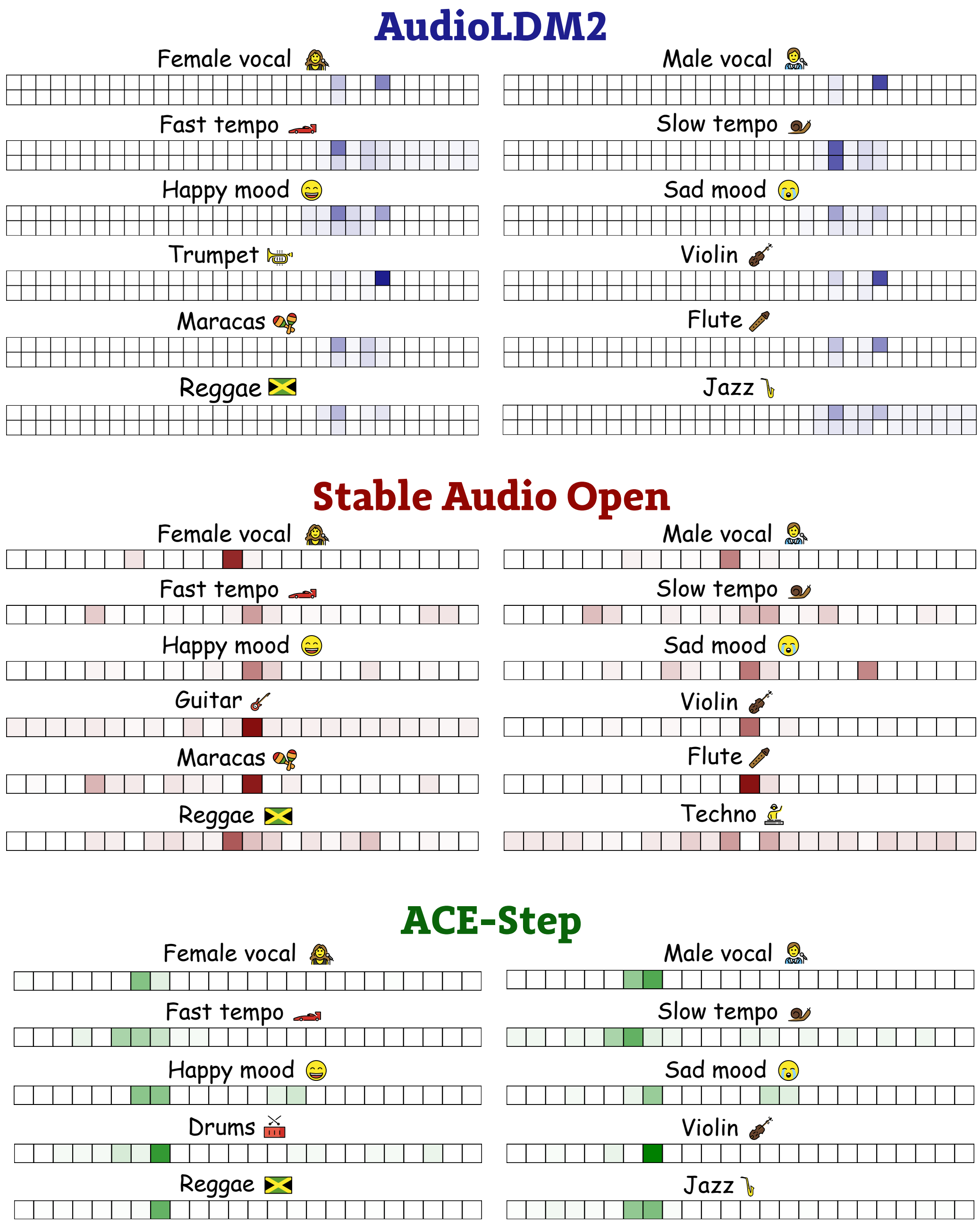

Figure 3: **Functional cross-attention layers in AudioLDM2 (Liu et al., 2024), Stable Audio Open (Evans et al., 2025), and ACE-Step (Gong et al., 2025) models.** We demonstrate that singular layers control different musical concepts, including vocal gender, tempo, mood, instruments, and genres across diverse audio diffusion architectures.

5 STEERING AUDIO DIFFUSION MODELS

Given a strong specialization of cross-attention layers, we further evaluate their usefulness on a downstream task of concept steering with audio DMs. The goal is to modulate the generated audio to increase the likelihood of including the concept c according to the steering strength α, while preserving other audio characteristics untouched.

5.1 BENCHMARK

Evaluation Metrics. We evaluate steering methods over a range of steering strengths $\alpha \in \{\alpha_{\text{min}}, \dots, \alpha_{\text{max}}\}$, with positive and negative directions evaluated independently. There exists an inherent trade-off between preserving the characteristics of the original audio and achieving alignment with the steered concept: stronger steering generally produces greater concept alignment at the cost of reduced preservation of the unsteered output. A preferable method in this case is one that, for the same level of audio preservation, achieves better alignment towards the target concept. Inspired by the perception-distortion tradeoff framework (Blau & Michaeli, 2018), we propose quantifying the **steering effectiveness** by computing the area under the alignment-preservation curve (**AUC (↑)**).

We use LPAPS (Iashin & Rahtu, 2021) to measure perceptual distance from the unsteered baseline ($\alpha = 0$), and define the preservation score as $p(\alpha) = P_{\text{max}} - \text{LPAPS}(\alpha)$, where higher values indicate better preservation of the original audio. We set P_{max} to the LPAPS obtained when the diffusion model is prompted with a different text prompt describing the target concept, representing the natural level of audio change expected from a prompt-level intervention, and providing an interpretable reference for the preservation axis. For the alignment axis we use the sign-corrected delta $\Delta a(\alpha) = \text{sign}(\alpha) \cdot (a(\alpha) - a(0))$, where $a(\cdot)$ is a similarity between audio generated with given steering level and the textual prompt describing this direction, measured with either CLAP (Wu et al., 2022) or MuQ (Zhu et al., 2025) model. Sign correction ensures that both steering directions are on a common scale where higher is better. For fair comparison across methods, all curves are truncated to the same preservation range $[0, P_{\text{max}}]$. The AUC is then approximated via the trapezoid rule on raw α samples: $\text{AUC} = \int_0^{P_{\text{max}}} \Delta a(p)\, dp$.

Following Ezra et al. (2025), we evaluate **Smoothness (↓)**, assessing how uniformly a method distributes concept changes across its steering scale. We normalize the alignment delta to $[0, 1]$ by dividing by the maximum correct-direction delta. Next, we compute consecutive gaps $g_i = \Delta a_{i+1} - \Delta a_i$ ordered by increasing $|\alpha|$, and define $\text{S} = \text{std}(\{g_i\})$.

Finally, we assess **Audio Quality (↑)** using Audiobox Aesthetics (Tjandra et al., 2025), averaging Content Enjoyment (CE), Content Usefulness (CU), Production Complexity (PC), and Production Quality (PQ) scores at 15 uniformly spaced preservation values for both directions.

Baselines. We compare localized steering for audio modulation against various baselines. *PCI* (Görgün et al., 2026) injects a concept-describing prompt starting from a specific diffusion timestep. As the maximum distortion PCI can produce is bounded by injecting the concept prompt at all diffusion steps, we use this maximum achievable LPAPS as P_{max} during evaluation. For prompt-level interventions, we also compare to *Text Embeddings* (Ezra et al., 2025), where steering direction is obtained through the difference of positive and negative prompt embeddings and applied to the neutral embedding, and *Token Embeddings* (Baumann et al., 2025), where steering vector is calculated and applied only to embeddings of tokens indicating modulated subjects.

Additionally, we compare to *FreeSliders* (Ezra et al., 2025), a training-free method where modulation is performed in diffusion prediction space, and *AUSteer* (Feng et al., 2026), an activation steering method with discriminative scalar selection, identifying and scaling only the most concept-relevant dimensions of the steering vector rather than applying a uniform perturbation.

Details. Methods are evaluated with Ace-Step, generating 30-second audio clips with 30 diffusion steps. For Contrastive Activation Addition (CAA), we compute steering vectors leveraging contrastive prompt pairs (see App. B for examples). Evaluation is performed on 100 diverse prompts spanning a wide variety of music styles and concepts, allowing us to assess the steering effect on a wide range of generations.

For SAEs, we train them on the outputs of attention layers $\{7, 8\}$, collecting activations with MusicCaps (Agostinelli et al., 2023) prompts. We conduct a hyperparameter search over expansion factors ($m \in \{2, 4, 8, 16\}$) and sparsity parameters ($k \in \{16, 32, 64\}$). We select the best configuration based on the lowest reconstruction error and fraction of both dead and high-frequency features, which is SAE ($m = 2, k = 32$) for layer $\{7\}$ and ($m = 4, k = 64$) for layer $\{8\}$. SAEs are trained for 10 and 15 epochs, respectively. Concept-specific features are selected based on generations from contrastive prompt pairs (same as in the case of CAA), according to the scoring methodology presented in Table 1.

Concept	Attention #7		Attention #8	
	Selection	Top-τ	Selection	Top-τ
Piano	TF-IDF (Eq. (6))	5	TF-IDF (Eq. (6))	20
Mood	TF-IDF (Eq. (6))	20	TF-IDF (Eq. (6))	20
Tempo	TF-IDF (Eq. (6))	20	TF-IDF (Eq. (6))	20
Vocal	TF-IDF (Eq. (6))	20	TF-IDF (Eq. (6))	20

Table 1: **Feature selection methodology for Sparse Autoencoders and the number of the top-scoring features selected per concept.**

5.2 AUDIO STEERING RESULTS.

First, we provide a rationale for localizing functional layers in Audio Diffusion Models for activation steering. Table 2 presents the quantitative results of steering the Ace-Step model with Contrastive Activation Addition (CAA), comparing our targeted intervention strategy against global baselines. We compare steering exclusively the identified functional layers ($\{7, 8\}$) against steering all layers ($\mathcal{L}$) and, crucially, an ablation setting where we steer all layers except the functional ones ($\mathcal{L} \setminus \{7, 8\}$).

Concept	Layers	(#)	Preservation (↓) LPAPS	Δ Alignment (↑) CLAP	MuQ
Piano	$\mathcal{L}$	24	3.374	0.130	0.211
	$\mathcal{L} \setminus \{7,8\}$	22	3.216	0.016	0.016
	$\{7,8\}$	2	2.799	0.131	0.238
Vocal Gender	$\mathcal{L}$	24	4.026	0.259	0.107
	$\mathcal{L} \setminus \{7,8\}$	22	3.635	0.134	0.056
	$\{7,8\}$	2	3.182	0.186	0.076
Tempo	$\mathcal{L}$	24	3.769	0.204	0.235
	$\mathcal{L} \setminus \{7,8\}$	22	3.227	0.040	0.065
	$\{7,8\}$	2	3.402	0.190	0.212
Mood	$\mathcal{L}$	24	3.678	0.220	0.075
	$\mathcal{L} \setminus \{7,8\}$	22	3.296	0.106	0.021
	$\{7,8\}$	2	3.122	0.162	0.093

Table 2: **Steering Ace-Step with CAA.** $\mathcal{L}$ denotes steering all cross-attention layers, $\mathcal{L} \setminus \{7, 8\}$ excludes layers 7 and 8, while $\{7, 8\}$ employs localized layers for steering.

The results provide striking validation of our localization hypothesis. We observe a "semantic bottleneck", consisting of controlling layers, within which the model's steering capacity for high-level concepts is almost entirely concentrated. As shown in Table 2, steering these two layers alone ($\{7, 8\}$) yields high alignment scores across all concepts. Conversely, when we apply steering to the remaining 22 layers while leaving the functional layers untouched ($\mathcal{L} \setminus \{7, 8\}$), the ability to control the generation collapses.

Beyond successful alignment, targeted steering demonstrates superior preservation of the original audio's characteristics. By limiting the intervention to the semantic bottleneck, we minimize collateral damage to unrelated acoustic features. This is evidenced by the Preservation metrics (LPAPS), where our steering consistently maintains lower distances to the original audio compared to using the vast majority of non-functional layers. For reference, the LPAPS computed between two sets of audio clips generated from 100 paired prompts, but with different random seeds, equals 5.7127.

Finally, in Table 3, we present a comprehensive comparison of all steering methods across four concepts. The results using Sparse Autoencoders (SAE($\{7, 8\}$)) further corroborate these findings. By steering along specific feature directions within functional layers, we achieve alignment scores that surpass those from raw activation steering across combined layers, offering the highest degree of interpretability. This is a noteworthy case, given the recent studies by Kantamneni et al. (2025) questioning the practical utility of SAEs compared to standard baselines in LLMs.

Looking across all concepts, SAE($\{7, 8\}$) and CAA($\{7, 8\}$) consistently rank among the top methods in terms of AUC, demonstrating that localizing interventions to the semantic bottleneck layers is a broadly effective strategy regardless of the steering mechanism. Beyond effectiveness, the Smoothness scores reveal a complementary advantage of localized steering, scoring the best for localized interventions.

Table 3: **Evaluating Audio Concept Steering with Ace-Step.** We describe metrics in Section 5.1. Symbols +/- denote positive/negative steering directions, *avg* = (pos + neg) / 2. On almost all the concepts, localized activation steering outperforms other baselines, including prompt-level, score-level, and activation-level interventions.

Concept	**Method**	**AUC LPAPS-MuQ (↑)**			**AUC LPAPS-CLAP (↑)**			**Smoothness (↓)**		**Average Quality (↑)**
		+	-	*avg*	+	-	*avg*	**MuQ**	**CLAP**	
Piano	PCI	0.075	0.004	0.039	0.031	0.001	0.016	0.242	0.289	6.813
	Text Embeddings	0.042	0.006	0.024	0.024	0.022	0.023	0.201	0.078	6.803
	Token Embeddings	0.109	0.035	0.072	0.044	0.021	0.032	0.076	0.087	6.792
	FreeSliders	0.063	0.046	0.055	0.025	0.024	0.025	0.050	0.078	6.784
	AUSteer	0.082	-0.000	0.041	0.011	0.028	0.019	0.587	0.124	6.769
	CAA	0.038	0.037	0.038	0.021	0.023	0.022	0.048	0.059	6.781
	CAA $\{\mathbf{7,8}\}$	0.113	0.073	0.093	0.048	0.046	0.047	**0.033**	**0.036**	6.785
	SAE $\{\mathbf{7,8}\}$	0.173	0.075	**0.124**	0.059	0.044	**0.052**	0.037	0.044	6.787
Vocal Gender	PCI	0.028	0.003	0.015	0.093	0.067	0.080	0.231	0.068	6.881
	Text Embeddings	-0.091	0.043	-0.024	0.047	-0.010	0.019	-	0.301	6.705
	Token Embeddings	0.016	0.001	0.009	-0.005	0.020	0.008	0.351	0.279	6.882
	FreeSliders	0.024	0.018	0.021	0.097	0.082	0.089	0.133	**0.057**	6.862
	AUSteer	0.004	-0.001	0.002	0.139	0.090	**0.114**	0.176	0.101	6.824
	CAA	0.012	0.021	0.016	0.084	0.067	0.076	0.134	0.067	6.854
	CAA $\{\mathbf{7,8}\}$	0.021	0.020	0.020	0.087	0.068	0.077	0.135	0.062	6.860
	SAE $\{\mathbf{7,8}\}$	0.016	0.053	**0.034**	0.080	0.032	0.056	**0.126**	0.108	6.820
Tempo	PCI	0.039	0.067	0.053	0.006	0.023	0.014	0.068	0.133	6.746
	Text Embeddings	0.013	0.015	0.014	-0.014	0.023	0.004	0.207	1.155	6.677
	Token Embeddings	0.034	0.068	0.051	-0.005	0.017	0.006	0.067	0.563	6.713
	FreeSliders	0.068	0.068	0.068	0.016	0.028	0.022	0.036	0.074	6.762
	AUSteer	0.082	0.077	0.079	0.018	0.031	**0.024**	0.038	0.058	6.775
	CAA	0.078	0.065	0.071	0.006	0.012	0.009	0.038	0.122	6.779
	CAA $\{\mathbf{7,8}\}$	0.101	0.076	0.089	0.024	0.025	**0.024**	0.048	**0.053**	6.778
	SAE $\{\mathbf{7,8}\}$	0.110	0.110	**0.110**	0.018	0.027	0.022	**0.032**	0.059	6.766
Mood	PCI	0.024	0.025	0.025	0.016	0.048	0.032	0.099	0.103	6.766
	Text Embeddings	0.005	0.002	0.003	-0.003	0.019	0.008	0.284	0.352	6.693
	Token Embeddings	0.024	0.025	0.025	0.018	0.030	0.024	0.082	0.092	6.775
	FreeSliders	0.035	0.026	0.031	0.021	0.042	0.031	0.083	0.052	6.771
	AUSteer	0.021	0.037	0.029	0.045	0.092	0.068	0.090	0.075	6.742
	CAA	0.021	0.005	0.013	0.036	0.061	0.048	0.124	**0.040**	6.772
	CAA $\{\mathbf{7,8}\}$	0.035	0.033	0.034	0.021	0.066	0.044	0.068	0.053	6.750
	SAE $\{\mathbf{7,8}\}$	0.039	0.064	**0.051**	0.036	0.104	**0.070**	**0.052**	0.050	6.726

Audio quality remains largely stable across all methods, with average scores in the range $[6.7, 6.9]$, confirming that no method causes severe perceptual degradation. The slight quality drop observed for Text Embeddings across all concepts is consistent with its higher distortion levels, as the method operates at the prompt level and tends to make broader changes to the generated audio.

Overall, these results validate our core hypothesis, showing that targeting the semantic bottleneck for steering audio diffusion models yields a favorable balance of effectiveness, smoothness, and quality preservation compared to both global activation steering and prompt-level interventions.

6 CONCLUSIONS

In this work, we demonstrated that distinct semantic musical concepts, such as instrument presence, vocals, and genre, are controlled by a small, shared subset of attention layers in audio diffusion architectures. By identifying these functional regions through activation patching, we showed that targeted interventions using Contrastive Activation Addition and Sparse Autoencoders enable precise manipulation of attributes such as tempo, mood, vocal gender, and piano presence without degrading audio quality. Our results confirm that this layer-specific steering outperforms global baselines in both precision and fidelity, offering a robust method for fine-grained musical control that overcomes the limitations of text prompting.

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/.

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.

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.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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`.

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`.

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.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. *arXiv preprint arXiv:2502.16681*, 2025.

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.

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.

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, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755377. URL https://doi.org/10.1145/3746027.3755377.

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, et al. Steermusic: Enhanced musical consistency for zero-shot text-guided and personalized music editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2000–2010, 2026.

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, 1997.

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.

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/.

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.

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.

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. AI, Arts Creativity.

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

CONTENTS

A TRACING DATASET

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.

Concept	Original Prompt $\mathcal{P}_c$	Counterfactual Prompt $\mathcal{P}_{\tilde{c}}$
Vocal Gender (female → male)		
	"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 (slow → fast)		
	"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 (happy → sad)		
	"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 (violin → trumpet)		
	"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 (reggae → metal)		
	"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 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**.

Category	Concept	Filter Keywords	Example Replacements
Vocal	**female**	female, woman, girlish	female → male woman → man girlish → boyish
	male	male, man	man → woman his → her himself → herself
Tempo	**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
Mood	**happy**	happy, joyful, upbeat, uplifting, cheerful	joyful → sorrowful cheerful → gloomy warm → cold
	sad	sad, depressing, dejected	sad → happy lonely → connected emotional → lighthearted
Instrument	**drums**	drums, drum, percussion	drums → saxophone percussionist → saxophonist
	flute	flute, flutes	flute → maracas flutist → maracast
	maracas	maracas, maraca	maracas → flute maraca → flute
	trumpet	trumpet, trumpets	trumpet → violin trumpeter → violinist
	violin	violin, fiddle	violin → trumpet string → $\emptyset$
Genre	**jazz**	jazz, blues	jazz → opera band → choir
	reggae	reggae, jamaican, tropical, beach	reggae → metal chill → heavy beach → concert

Table 5: **Keywords for dataset construction.** For each concept, we list the keywords used to filter captions from MusicCaps (selecting prompts containing these terms) and the replacement keywords used to generate counterfactual prompts.

B Steering Experiment Details

This section provides details on the steering experiments.

B.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 $\mathcal{P}_{\bar{c}}$ with the contrasting attribute. Table 6 shows the prompt templates used for each concept.

The base prompts span 50 diverse musical styles and genres, including: "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.

Concept	**Positive Prompt** $\mathcal{P}_c$	**Negative Prompt** $\mathcal{P}_{\bar{c}}$
Piano	"{base} with piano"	"{base}"
Mood	"a happy {base}"	"a sad {base}"
Tempo	"fast {base}"	"slow {base}"
Vocal Gender	"{base} with female vocals"	"{base} with male vocals"

Table 6: **Contrastive prompt templates for steering.** The {base} placeholder is filled with diverse musical descriptions (e.g., "a song", "a jazz piece", "electronic music").

B.2 Evaluation Prompts

For steering evaluation, we use 100 diverse prompts that cover a wide range of musical styles, for example:

- **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'

B.3 Audio-Text Alignment Prompts

To measure concept alignment (Alignment metric in Table 2), 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.

Concept	**Text Query**	**Alignment Model**
Piano	"a piano song"	MuQ
Mood (sad → happy)	"a cheerful track"	MuQ
Tempo (slow → fast)	"a fast track"	MuQ
Vocal Gender (male → female)	"This is a music of a female vocal singing"	CLAP

Table 7: **Text queries for audio-text alignment measurement.** These prompts are used to compute similarity scores between generated audio and target concepts.