

AUDIO SLIDERS: MEASURING WHAT A SLIDER DOES TO MUSIC

Taka Khoo

Independent researcher

takakhoo@gmail.com

ABSTRACT

A slider is a small low-rank update to a frozen music generator whose strength is chosen at sampling time, so that one prompt and one seed can be nudged along a single axis: a little happier, a little fuller, a little brighter. Sliders borrowed from image diffusion are usually judged by whether the output drifts toward the right words in a text–audio embedding. We ask what that leaves out. On Stable Audio Open and ACE-Step we test whether a property measured from the waveform follows the slider, whether two unrelated quality predictors still hear music, how much of the original piece survives, and what else moved. Two sliders pass the embedding test and fail the waveform test. A trained slider matches prompt-pair guidance at a third of the cost, and interpolating the prompt breaks the audio. We then let real music choose the axes: independent components of embeddings of 14,985 recordings include arousal and valence, and the generator’s own output spans only 28–83% of the real spread along them. Sliders trained along these axes, from the axis’s own tags or with no text at all, follow them on held-out prompts (rank correlation up to 0.91) and carry a clip across the mean of real music while both quality predictors stay level.

1. INTRODUCTION

Anyone who has worked with a text-to-music model knows the moment. The clip is good, almost right, and you want it a little happier, or a little fuller, or with richer harmony. The only control you have is the prompt, and rewriting it hands you a different piece. What a musician wants here is closer to a knob than a sentence: keep the prompt, keep the seed, and move the result along one axis by an amount you choose.

Image diffusion already has such knobs. Concept Sliders [1] train a low-rank adapter whose scale is set at inference time, and SliderSpace [2] goes further and discovers the axes themselves. Both ideas have crossed into audio, mostly as baselines [3, 4]. In that work a slider counts as working if the output moves toward the attribute text in a joint text–audio embedding and stays close to the original.

For music this leaves a lot unmeasured. An embedding can move while the thing a listener would name does not;

we will show two sliders for which exactly this happens. A slider can also reach its target by wrecking the clip, and similarity to the original says nothing about whether what remains is still music. Nothing in the score says what else moved. And the axes themselves are chosen by whoever writes the prompt pair, with no evidence that these are the directions along which music varies, or that the generator can reach their ends at all.

This paper is about measuring those things, and about what happens when you let real music pick the axes instead. We work with two open generators of different design, a 50-step latent diffusion model and an 8-step rectified-flow model, and with a third as a check. Our contributions are:

1. **A measurement protocol for music sliders.** Every control is scored on a waveform descriptor fixed in advance, in two unrelated embedding models, by two unrelated quality predictors anchored to real recordings, on similarity to the unsteered piece, and on what else it moved. Effects are reported only over the *usable span*, the positions at which quality has not yet dropped.
2. **A comparison of four ways to move the same attribute.** A trained slider reproduces prompt-pair guidance with one forward pass per step instead of three; activation steering is a real competitor; interpolating the text conditioning fails.
3. **Axes taken from real music, and how much of them a generator covers.** Independent components of embeddings of 14,985 recordings include the arousal and valence of the circumplex model [5]. ACE-Step’s output over 648 prompts spans 28–83% of the real spread along them, and Stable Audio Open looks the same.
4. **Sliders along those axes, trained two ways.** From a prompt pair made of the axis’s own tags, and with no text from two sets of the generator’s clips. Both follow the axis. The first moves further; the second keeps more of the piece, moves less of everything else, and a graded form of it reaches ± 2 .

We also report what did not work, because the negative results shape the method: training directly between sets of real recordings, and principal axes of the generator’s own activations, which turned out to describe production rather than music.



© T. Khoo. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** T. Khoo, “Audio Sliders: Measuring What a Slider Does to Music”, manuscript in preparation for ISMIR 2027, not yet peer reviewed.

2. RELATED WORK

2.1 Sliders and steering

Concept Sliders [1] train a LoRA [6] so that its output matches the frozen model’s prediction shifted along the difference between a positive-prompt and a negative-prompt prediction. SliderSpace [2] drops the prompt pair, takes principal components of CLIP embeddings of many images of one concept, and trains one adapter per component. Measured Sliders [7] define image sliders by differentiable measurements and hold the other measurements still; we borrow their selectivity score.

In audio, FreeSliders [3] apply the Concept Sliders target at inference time without training, across image, video, and audio models; most of their audio concepts are sound effects. TADA [4] steers the cross-attention activations of audio diffusion transformers with vectors from prompt pairs, or from sparse-autoencoder features selected with prompt pairs. Facchiano et al. [8] compute difference-in-means vectors for tempo and timbre; AnchorSteer [9] learns text-free concept modules on self-generated clips; Singh et al. [10] find and steer sparse features in autoregressive music models. All of these score a control by text–audio alignment and by perceptual distance to the unsteered output. TADA checks two concepts against a descriptor, and Wang [11] shows with a key estimator and a beat tracker that apparent obedience to an instruction can be the model’s prior. We are not aware of work that scores every control against a waveform property fixed in advance, compares it with plain signal processing, or reports what else moves.

2.2 Where axes come from, and how much of them a model covers

In image diffusion, principal components of a U-Net’s bottleneck activations turn out to be semantic [12, 13]. Guinot et al. [14] learn sparse features of music embeddings to steer retrieval, and independent components of embeddings are known to read more cleanly than principal ones [15]. We found no work that takes axes from embeddings of real music and trains generator controls along them.

Diversity against real music is usually one number per model, such as recall and coverage in an embedding space [16]. Slendebroek and Metaxa [17] compare two commercial systems with human recordings on hand-built signal features and report homogenisation. Our comparison is per axis, on axes learned from real recordings, for open models that a slider can then be trained on. RIME [18] builds paired edit data by applying effects with known parameters; we use such effects the other way round, as a baseline a learned control has to beat.

3. SLIDERS AND HOW WE TRAIN THEM

3.1 Backbones

Two frozen text-to-music generators sit behind one interface. Stable Audio Open 1.0 [19] is a latent diffusion transformer with a v -prediction objective, sampled with

DPM-Solver++(2M) for 50 steps at guidance 7. ACE-Step 1.5 XL turbo [20] is a rectified-flow transformer, $x_t = (1-t)x_0 + t\epsilon$ with the network predicting $\epsilon - x_0$, distilled to 8 Euler steps with no guidance. Everything below is written for a generic prediction $f_\theta(z_t, t, c)$ and applies to both. Stable Audio 3 small appears once, as a check that nothing depends on the first two.

3.2 A slider

A slider is a set of low-rank updates, one per linear layer in the attention and feed-forward blocks. With position $s \in \mathbb{R}$ a layer computes

$$y = Wx + s \frac{\alpha}{r} BAx, \quad (1)$$

with $r = 4$ and $\alpha = 1$. The position is read at every forward pass and can differ per sample, so one batch can hold a clip at several positions. Sliders add, and a slider can be switched on only for $t \leq t_0$, which we call gating.

3.3 Training from a prompt pair

Let c^+ and c^- be a base prompt c with a phrase for each end of the attribute appended, say “happy, joyful” and “sad, melancholic”. Following Concept Sliders, for $s \in \{+1, -1\}$ the slider is trained so that

$$f_{\theta,s}(z_t, t, c) \approx f_\theta(z_t, t, c) + s \eta (f_\theta(z_t, t, c^+) - f_\theta(z_t, t, c^-)), \quad (2)$$

with the right-hand side from the frozen model. The state z_t comes from sampling for a random number of steps with the slider active, so the slider learns on states it will visit. We use $\eta = 4$ for Stable Audio Open and $\eta = 2$ for ACE-Step, 1,000 iterations, and 48 training prompts that share nothing with the evaluation prompts.

3.4 Training from two sets of clips

The second trainer never sees text. Given two sets H and L of latents, we minimise the generator’s own denoising loss with the slider at $+1$ on H and at -1 on L :

$$\mathbb{E}_{x_0 \in H} \|f_{\theta,+1}(z_t, t, c) - u\|^2 + \mathbb{E}_{x_0 \in L} \|f_{\theta,-1}(z_t, t, c) - u\|^2, \quad (3)$$

where u is the regression target for (x_0, ϵ, t) . One update serves both ends with opposite sign, so to first order whatever the two sets share pulls the update in opposite directions and cancels. What remains is the difference between the sets. We build H and L from the generator’s own output: generate a corpus, measure something on every clip, and take the top and bottom 20–30% *within each prompt*, so that no prompt is a shortcut for “high”. The measurement can be a signal descriptor, a quality score, or the projection on a direction in an embedding space, and it can be residualised against other measurements first so the two sets differ in the target and little else. Nothing has to be differentiable through the decoder or the embedding model. A *graded* variant trains every clip at its own position, its standing inside its prompt scaled so that the two tails average ± 1 , drawing clips far from zero more often.

3.5 Where the axes come from

We take directions from three places. (i) Principal components of CLAP embeddings of 1,024 clips of one concept, as in SliderSpace. (ii) Principal and independent components [15] of CLAP [21] and MuQ-MuLan [22] embeddings of 14,985 real recordings from FMA [23], after dropping the 40% most vocal clips and removing each genre’s mean, so that an axis describes how recordings of the same genre differ. (iii) Principal components of the generator’s own activations: the mean cross-attention output of every block for 768 clips, with each prompt’s mean removed.

3.6 Other ways to move an attribute

Under identical prompts, seeds, and sampler we also run the right-hand side of Eq. (2) directly at every step, which is the training-free recipe of FreeSliders and costs three forward passes per step; activation steering, which adds the mean difference in cross-attention output between c^+ and c^- to every block [4]; interpolation of the text conditioning toward c^+ or c^- ; and, where one exists, a conventional effect applied to the unsteered clip.

4. WHAT WE MEASURE

For each prompt and seed we render the same initial noise at every slider position and call the result a trajectory. Unless stated, every number uses 24 prompts held out from training, three seeds, and ten-second clips.

Does a measured property follow the slider? Each slider is paired, before any training, with a descriptor computed from the waveform with no learned model in the loop: spectral centroid, onset rate, beat-tracked tempo [24], energy decay time, percussive energy share, low-band and stereo-side energy ratios, key clarity, major-minus-minor key fit, harmonic change rate. We report the mean Spearman correlation between position and descriptor over trajectories (ρ), the share of trajectories whose two ends are in the intended order, and the change in units of the descriptor’s standard deviation across unsteered clips.

Does a second model agree? The projection of each clip on the slider’s text direction is computed in CLAP and again in MuQ-MuLan, which shares no weights or training data with it.

Does it still sound like music? Every clip is scored by Audiobox Aesthetics [25] (content enjoyment, CE, on a 1–10 scale) and by SongEval [26] (musicality, 1–5), a head on MuQ features trained on human ratings of generated songs. Both are anchored by 2,000 excerpts of real recordings, so that a number means something. The *usable span* of a slider is the set of positions, walking outward from zero, over which mean CE stays within 0.5 of the unsteered clips. Effect sizes are reported inside it, which stops a slider from being credited for change it only achieves by destroying the clip.

How much of the piece survives? CLAP similarity between each clip and the same trajectory at position zero.

What else moves? The slope of every descriptor against position, in standard deviations per unit, for every

slider: a leakage matrix. A panel of 80 instrument, genre, mood, and character tags, scored in CLAP, puts the change into words.

Coverage. For a direction d found in real recordings, the coverage of a generator is the standard deviation of its clips’ projections on d divided by that of the recordings. We report it over all prompts, and again after removing each prompt’s mean from the generated clips and each genre’s mean from the recordings, which is what changing only the seed can reach. A slider’s reach along d is reported in the same unit.

5. RESULTS

5.1 Do the named sliders do what their names say?

First the yardstick. Content enjoyment averages 6.12 over 2,000 real recordings, 6.95 for unsteered ACE-Step clips and 6.04 for Stable Audio Open. A slider that drops a clip well below the real-music mean has left the region where the change it made is worth having.

Table 1 and Figure 1 report eight ACE-Step prompt-pair sliders over nine positions. Where a descriptor was assigned in advance it rises with the slider, and MuQ-MuLan agrees with CLAP on the direction for all eight. The tag panel says the same in words: the mood slider raises *happy* and lowers *dark-toned*. Enjoyment stays above the real-music mean across the range for every slider but tension, which collapses past +1, and that is where its usable span ends. SongEval, trained on different data, tells the same story: unsteered clips score 2.77 in musicality against 2.78 for real recordings, the usable span it implies is identical for six of the eight sliders and half a step wider for the other two, and the two predictors agree over clips with Spearman 0.52–0.67. The recipe carries over unchanged to Stable Audio 3 small, where six sliders at 600 iterations all move their descriptor (ρ 0.34–0.84, four with at least 96% of trajectories ordered), with the ungated quality cost of a 50-step model.

5.2 Four ways to move the same attribute

Table 2 runs the alternatives on the same prompts and seeds. The trained slider reproduces prompt-pair guidance, which is after all its own training target, with one forward pass per step instead of three: it scores higher on brightness and harmony and lands inside the confidence interval everywhere else. Activation steering turns out to be a real competitor. It is within 0.05 of the slider on four of six descriptors and keeps more of the piece (0.83–0.87 against 0.74–0.84), while the slider moves its descriptor further on brightness, ensemble, and harmony. Interpolating the text conditioning has the lowest correlation on every attribute and breaks the audio at the ends.

5.3 The embedding says yes, the waveform says no

On Stable Audio Open, sliders trained from the prompt pairs for note density and for percussion move the output along their CLAP direction at every setting. Neither moves

Slider	Descriptor	ρ	Ordered	MuQ ρ	Span	Moved	Kept	CE ₀	CE _{ends}
brightness	centroid_oct	0.97	1.00	0.80	-1.0 to +2.0	4.41	0.74	6.95	6.04
density	onset_rate	0.78	0.92	0.86	-0.5 to +2.0	1.25	0.84	6.95	6.09
mood	majorness	0.39	0.75	0.88	-1.0 to +2.0	0.87	0.79	6.95	6.65
harmony	harmonic_change	0.41	0.76	0.73	-1.5 to +2.0	0.84	0.83	6.95	6.64
groove	—	—	—	0.56	-1.5 to +2.0	—	0.79	6.95	6.88
ensemble	pc	0.60	0.90	0.77	-1.5 to +2.0	0.92	0.80	6.95	6.56
tension	—	—	—	0.86	-1.5 to +1.0	—	0.83	6.95	5.64
melody	key_clarity	0.48	0.78	0.89	-1.0 to +2.0	0.97	0.80	6.95	6.47

Table 1. Prompt-pair sliders on ACE-Step 1.5 XL turbo. ρ : mean rank correlation between position and descriptor. Ordered: share of trajectories with correctly ordered ends. MuQ ρ : the same correlation for the MuQ-MuLan direction score. Span: usable span. Moved: descriptor change over the span in standard deviations of unsteered clips. Kept: CLAP similarity to the unsteered clip at the ends of the span. CE: content enjoyment unsteered and at the extreme positions.

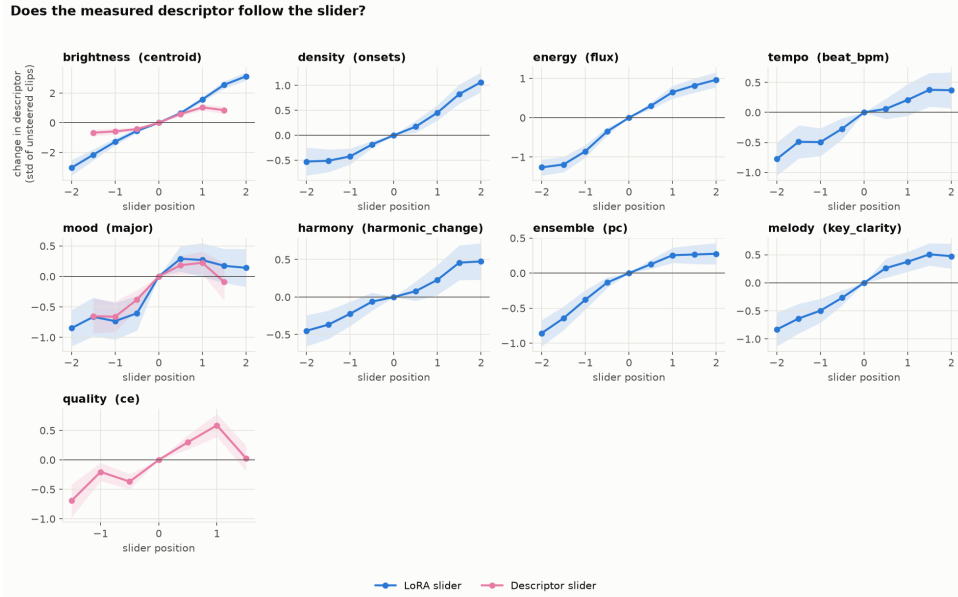


Figure 1. ACE-Step: change in each slider’s descriptor against position, in standard deviations of unsteered clips, for prompt-pair sliders (blue) and sliders trained between two sets of clips (pink). Shaded bands are 95% intervals over trajectories.

its descriptor: the rank correlation with onset rate or percussive share never exceeds 0.37. Scored by embedding alone, both would have been reported as working. The descriptor has to be sound as well. With a tempo estimate prone to octave errors the tempo slider scored $\rho = -0.01$; with a beat tracker it scores 0.57 and moves the mean tempo from 90 to 167 BPM.

Something similar shows up when the training recipe changes. Retraining the mood slider with rank 16, with a doubled guidance multiplier, on cross-attention only, or for three times as long leaves the CLAP direction score at 0.92–0.93 in every case, while the descriptor correlation ranges from 0.29 to 0.52. The embedding cannot tell these sliders apart; the waveform can.

5.4 Leaving the first steps alone

On the 50-step model a slider at ± 1 takes the piece with it. Switched on only from step 21, the brightness slider keeps 1.41 of its 2.37 standard deviations of range, similarity to the unsteered clip rises from 0.67 to 0.85, and the loss in

enjoyment falls from 1.0 to 0.3 points (Table 3). The earliest steps decide the layout of the piece; a slider that leaves them alone keeps it.

5.5 What else moves, and cancelling it

Sliders are not clean. The mood and melody sliders both raise the spectral centroid by 0.8 standard deviations per unit, and several raise loudness. A measured leak can be cancelled with a second slider: adding the brightness slider at -0.53 and -0.55 times their position brings the leak to -0.02 and $+0.09$ and raises similarity to the unsteered clip at ± 1 from 0.85 to 0.90 and from 0.87 to 0.92. The CLAP direction correlation drops too, from 0.92 to 0.69 and from 0.93 to 0.72, which suggests that part of what the embedding scores as happier or more melodic was simply brightness.

5.6 Axes that real music chose

Fitted on two random halves of the recordings and matched one to one, the leading eight axes agree with mean co-

Slider	Slider		Guidance		Act. steer.		Interp.	
	ρ	MuQ	ρ	MuQ	ρ	MuQ	ρ	MuQ
brightness	0.97	0.80	0.89	0.81	0.93	0.81	0.49	0.35
density	0.78	0.86	0.74	0.81	0.71	0.86	0.10	0.38
mood	0.39	0.88	0.45	0.87	0.41	0.88	0.25	0.63
harmony	0.41	0.73	0.08	0.68	0.24	0.72	-0.01	0.23
groove	–	0.56	–	0.71	–	0.58	–	0.12
ensemble	0.60	0.77	0.55	0.75	0.58	0.67	0.25	0.25
tension	–	0.86	–	0.88	–	0.91	–	0.69
melody	0.48	0.89	0.51	0.83	0.53	0.83	0.31	0.38
CE at ends	6.37		6.54		6.52		4.95	
Passes per step	1		3		1		1	

Table 2. Four ways to move the same attribute on ACE-Step, same prompts and seeds. ρ : rank correlation of position with the waveform descriptor (– where none was assigned); MuQ: with the MuQ-MuLan direction. CE at ends: mean content enjoyment at the two extreme positions, averaged over sliders.

On after	Moved	CLAP	Chroma	CE
0	2.37	0.67	0.68	5.15
10	2.11	0.71	0.71	5.34
14	1.98	0.74	0.74	5.45
21	1.41	0.85	0.84	5.87
29	0.42	0.97	0.97	6.08

Table 3. Timestep gating for the brightness slider on Stable Audio Open at positions ± 1 : descriptor moved (std), CLAP and chroma similarity to the unsteered clip, content enjoyment. Unsteered CE is 6.16.

sine 0.98 (MuQ-MuLan, principal), 0.94 (CLAP, principal), 0.90 (MuQ-MuLan, independent), and 0.83 (CLAP, independent). Of 1,024 sparse-autoencoder features on CLAP, only 9 are found by two seeds. A harder test refits the axes on 47,942 further recordings from a different slice of the archive: the leading subspace carries over (the top 32 directions overlap 97%) while single axes rotate inside it, and the matched cosine of the MuQ-MuLan independent components falls to 0.61.

The independent components of MuQ-MuLan are the ones that read like music. One runs from quiet, minor-key, and dreamy to aggressive, dry, and rhythmic; another from dark and distorted to playful and simple. These are recognisably the arousal and valence of the circumplex model [5], and nobody asked for them. Others contrast jazz with electronic music, plucked and bowed strings with synthesizer and choir, and piano with organ and rock. Arousal, valence, and the jazz–electronic axis reappear in the second corpus with the same tags (cosine 0.65–0.76); the strings–synthesizer axis does not.

5.7 How much of real music does a generator cover?

Along every one of these axes the generator is narrower than real music (Figure 2). Over 648 prompts crossing genre, instrument, and mood, ACE-Step spans 28–83% of the real spread; seeds of a single prompt span 38–69% of what recordings of one genre span. The average generated clip sits 1.35 standard deviations toward the quiet end of

Axis	From	ρ	Ord.	Real	Seeds	Kept	CE	Mus.
acoustic/electr.	sets	0.55	0.93	0.64	1.48	0.86	7.11	2.68
arousal	pair	0.91	1.00	1.85	3.35	0.80	6.84	2.71
arousal	sets	0.82	0.99	1.19	2.14	0.84	7.13	2.70
classical/funk	sets	0.64	0.96	0.77	1.65	0.85	7.12	2.69
jazz/electronic	pair	0.86	0.97	1.86	3.78	0.82	7.06	2.72
jazz/electronic	sets	0.61	0.94	0.61	1.23	0.87	7.09	2.70
piano	pair	0.78	0.97	1.56	2.93	0.80	6.78	2.74
piano	sets	0.63	0.96	0.75	1.42	0.87	7.08	2.69
strings/synth	pair	0.80	0.97	1.78	3.28	0.73	6.70	2.64
strings/synth	sets	0.40	0.82	0.44	0.81	0.87	7.09	2.69
valence	pair	0.42	0.75	0.86	1.71	0.85	6.75	2.67
valence	sets	0.50	0.93	0.62	1.23	0.87	7.07	2.73

Table 4. Sliders along axes of real music on ACE-Step, positions -1 to $+1$. From: a prompt pair made of the axis’s tags, or two sets of the model’s clips. ρ , Ord.: rank correlation between position and the output’s projection on the axis, and share of trajectories with -1 and $+1$ in the right order. Real, Seeds: movement along the axis between -1 and $+1$ in standard deviations of real recordings and of one prompt’s seeds. Kept: CLAP similarity to the unsteered clip. CE, Mus.: enjoyment and SongEval musicality at ± 1 (unsteered: 6.95 and 2.77).

the arousal axis and 1.36 toward the playful end of the valence axis. On CLAP’s first three principal axes ACE-Step covers 35–39% and Stable Audio Open 48–52%, from different training data and a different architecture. The top eight principal directions of the generated corpus contain only 31% of the top eight of the real one: the directions along which generated clips differ most are mostly not the ones along which recordings do.

5.8 Sliders along real-music axes

We turned the axes into sliders in two ways (Table 4). The set trainer takes the top and bottom 20% of 24 seeds per prompt along the axis. The second route is simpler than it sounds: read the tags at the two ends of an axis and use them as the prompt pair in Eq. (2). For arousal these are “quiet, dreamy, melancholic, minor key” and “aggressive, energetic, rhythmic, dry”, words that an independent component of real-music embeddings supplied. Either way the output is scored by its projection on the axis from real recordings, in standard deviations of those recordings.

All twelve sliders follow their axis. The prompt pair is the stronger route on four of five axes, with rank correlations of 0.78–0.91 and 1.6–1.9 standard deviations of real music between -1 and $+1$. Unsteered clips sit 1.21 standard deviations toward the quiet end of the arousal axis; the slider takes the same prompts and seeds across the real-music mean to -0.26 at position -2 , and to $+2.16$ at $+1$, where its usable span ends. For scale, the standard deviation of the generator’s whole output over 648 prompts on this axis is 0.83. The set-trained sliders move less (0.44–1.19), keep more of the piece (0.84–0.87 against 0.73–0.85), and hold both quality scores level at both ends; on valence they are also the more reliable (93% of trajectories ordered against 75%).

The model explores less of each musical axis than real music does

Axes: independent components of MuQ-MuLan embeddings of 14,985 FMA recordings. Model: ACE-Step 1.5 XL turbo, 15,552 clips.

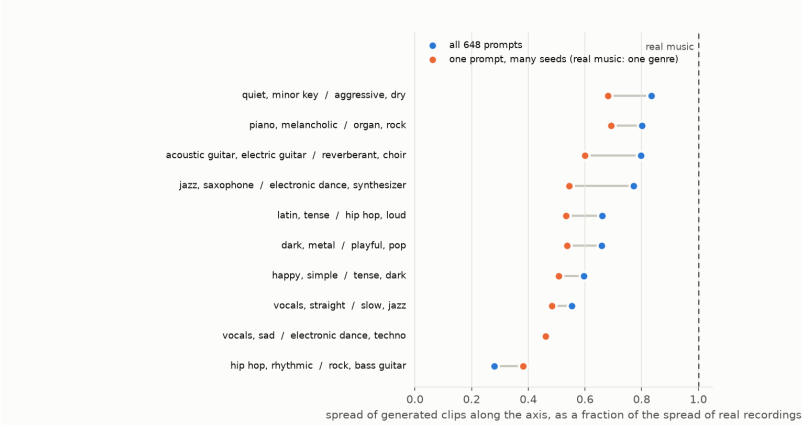


Figure 2. How much of each real-music axis 15,552 ACE-Step clips span, as a fraction of the spread of real recordings, over all 648 prompts (blue) and for one prompt with many seeds (orange). Each axis is labelled by the tags at its two ends.

5.9 Each trainer moves what it was trained on

Sorting clips by a waveform descriptor, residualised against loudness, brightness, and enjoyment, gives sliders that are selective. Over positions -1 to $+1$ the set-trained harmony slider reaches $\rho = 0.72$ on harmonic change rate with selectivity 6.6, against 0.27 and 0.8 for the prompt-pair slider; density and ensemble show 4.7 and 2.4 against 1.5 and 1.5. The embedding tells the opposite story. The MuQ-MuLan text direction for harmony follows the prompt-pair slider ($\rho = 0.63$) and ignores the set-trained one (-0.07). Neither kind of score can stand in for the other, which is the point of measuring both.

Set-trained sliders stop at ± 1 . Beyond it, similarity falls below 0.72 and musicality drops by 0.3–0.5 in both directions while CE stays flat, a failure only the second predictor sees. The graded variant removes the cliff: six sliders retrained this way are usable from -2 to $+2$ under both predictors, with energy moving spectral flux 1.75 standard deviations across that range and arousal 1.46 standard deviations of real music, at the cost of slightly less movement inside ± 1 . Tempo did not become a slider either way ($\rho = 0.06$).

5.10 What did not work

Axes inside the generator. We hoped the generator’s own activations would give musical axes without any embedding model. Eight principal axes carry 90% of the within-prompt variance of ACE-Step’s cross-attention outputs and reproduce across disjoint prompt sets (0.82). Steering along the first two raises level and lowers spectral flatness ($\rho = 0.77$ and -0.64) and moves tags from lo-fi toward complex and live: these are production axes. The other six leave at least 95% similarity to the unsteered clip at twelve natural standard deviations and act differently from prompt to prompt (consistency 0.10–0.21). Unlike the image case [13], the leading directions here describe polish; arousal, valence, and instrument contrasts only appeared in embeddings of real music.

Training between sets of real recordings. With H and L drawn from the recordings themselves, the slider damaged the clip equally in both directions. A penalty tying the mean of its predictions at ± 1 to the frozen model removed the damage and left almost no steering: tempo did not move and major-minor fit went from 0.09 to 0.15. Real recordings sit off the generator’s distribution and the mismatch swamps the loss, so real music served to find the axes and the generator’s own clips to train along them.

6. LIMITATIONS AND OUTLOOK

No listening test was run. Every claim about quality rests on two learned predictors and every claim about meaning on signal descriptors and tags, and these are proxies: energy decay time is confounded by sustained instruments, and SongEval was trained on songs with vocals. Most clips are ten seconds long. The usable-span tolerance of 0.5 CE is a choice. Coverage depends on our prompt list and on FMA as the sample of real music, and both embedding models were trained on real recordings, so generated audio may be slightly out of their distribution. Axis labels come from an 80-word vocabulary, and a different vocabulary would name the same direction differently. The internal-axis result is for token-averaged cross-attention outputs of an eight-step distilled model; other sites in the network may hold what this one does not.

What we take from the study is a way of working rather than a set of sliders. Decide what a control should change before training it, measure that thing from the waveform, let a quality model anchored to real recordings say how far the control may be pushed, and let real music, rather than a prompt writer, suggest which axes are worth a control at all.

7. ETHICS STATEMENT

The real recordings used to find axes come from the Free Music Archive, which is released under Creative Commons licenses, and only embeddings and direction vectors

derived from them are redistributed. The generators are open-weight models used under their licenses. Sliders of this kind lower the effort needed to produce many variations of a generated piece; they do not change what the underlying models can produce, and they inherit those models' biases, including the narrowness along musical axes that this paper measures. We release code, trained sliders, and the evaluation clips so that the measurements can be checked and extended.

8. REFERENCES

- [1] R. Gandikota, J. Materzyńska, T. Zhou, A. Torralba, and D. Bau, "Concept sliders: LoRA adaptors for precise control in diffusion models," in *Proc. European Conference on Computer Vision (ECCV)*, 2024, arXiv:2311.12092.
- [2] R. Gandikota, Z. Wu, R. Zhang, D. Bau, E. Shechtman, and N. Kolkin, "SliderSpace: Decomposing the visual capabilities of diffusion models," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, arXiv:2502.01639.
- [3] R. Ezra, H. Zisling, N. Berman, I. Naiman, A. Gorkor, L. Nochumsohn, E. Nachmani, and O. Azencot, "FreeSliders: Training-free, modality-agnostic concept sliders for fine-grained diffusion control in images, audio, and video," *arXiv:2511.00103*, 2025.
- [4] Ł. Staniszewski, K. Zaleska, M. Modrzejewski, and K. Deja, "TADA! tuning audio diffusion models through activation steering," *arXiv:2602.11910*, 2026.
- [5] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. International Conference on Learning Representations (ICLR)*, 2022, arXiv:2106.09685.
- [7] Y. Chen, B. Wei, and X. Yin, "Measured sliders: Learning continuous controls from differentiable image measurements," *arXiv:2609.05234*, 2026.
- [8] S. Facchiano, G. Strano, D. Crisostomi, I. Tallini, T. Mencattini, F. Galasso, and E. Rodolà, "Activation patching for interpretable steering in music generation," *arXiv:2504.04479*, 2025.
- [9] C.-H. Chang, K.-S. Ho, C.-Y. Tsai, K.-L. Chen, Y.-H. Yang, and J.-J. Ding, "AnchorSteer: Self-discovered concept injection for structure-preserving music editing," in *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026, arXiv:2605.31053.
- [10] N. Singh, M. Cherep, and P. Maes, "Discovering and steering interpretable concepts in large generative music models," in *Proc. International Conference on Learning Representations (ICLR)*, 2026, arXiv:2505.18186.
- [11] Y. Wang, "Do text-to-music models really follow instructions? a counterfactual evaluation of key and beat grouping," *arXiv:2608.11899*, 2026.
- [12] M. Kwon, J. Jeong, and Y. Uh, "Diffusion models already have a semantic latent space," in *Proc. International Conference on Learning Representations (ICLR)*, 2023, arXiv:2210.10960.
- [13] R. Haas, I. Huberman-Spiegelglas, R. Mulayoff, S. Graßhof, S. S. Brandt, and T. Michaeli, "Discovering interpretable directions in the semantic latent space of diffusion models," in *Proc. 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2024, arXiv:2303.11073.
- [14] J. Guinot, A. Riou, E. Quinton, and G. Fazekas, "Steering dense music retrieval with open-vocabulary concept discovery," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2026, arXiv:2608.08757.
- [15] H. Yamagiwa, M. Oyama, and H. Shimodaira, "Discovering universal geometry in embeddings with ICA," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, arXiv:2305.13175.
- [16] Z. Novack, G. Zhu, J. Casebeer, J. McAuley, T. Berg-Kirkpatrick, and N. J. Bryan, "Presto! distilling steps and layers for accelerating music generation," *arXiv:2410.05167*, 2024.
- [17] Z. Slendebroek and D. Metaxa, "The algorithmic flattening of sound: Computational evidence and justice implications of AI music homogenization," *arXiv:2608.06106*, 2026.
- [18] N. Schaffer and N. Singh, "RIME: Enabling large-scale agentic music post-production," *arXiv:2607.19605*, 2026.
- [19] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable audio open," *arXiv:2407.14358*, 2024.
- [20] J. Gong, Y. Song, W. Zhao, S. Wang, S. Xu, J. Guo, and X. Yang, "ACE-Step 1.5: Pushing the boundaries of open-source music generation," *arXiv:2602.00744*, 2026.
- [21] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, arXiv:2211.06687.

- [22] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, “MuQ: Self-supervised music representation learning with mel residual vector quantization,” *arXiv:2501.01108*, 2025.
- [23] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A dataset for music analysis,” in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2017, arXiv:1612.01840.
- [24] F. Foscarin, J. Schlüter, and G. Widmer, “Beat this! accurate beat tracking without DBN postprocessing,” in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2024, arXiv:2407.21658.
- [25] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W.-N. Hsu, “Meta audibox aesthetics: Unified automatic quality assessment for speech, music, and sound,” *arXiv:2502.05139*, 2025.
- [26] J. Yao, G. Ma, H. Xue, H. Chen, C. Hao, Y. Jiang, H. Liu, R. Yuan, J. Xu, W. Xue, H. Liu, and L. Xie, “SongEval: A benchmark dataset for song aesthetics evaluation,” *arXiv:2505.10793*, 2025.