Beyond FAD and CLAP

A Modern Perceptual Re-Ranking and a Controllability Audit of Open-Source Instrumental Music Generators

Tony Salomone
Transformer Lab
tony.salomone@lab.cloud

Deep Gandhi
Transformer Lab

Ali Asaria
Transformer Lab

June 25, 2026

Abstract

Vendor reports for text-to-music models rank systems with Fréchet Audio Distance (FAD) and CLAP score computed on a single CLAP encoder, the same family a model may be trained against, raising a circularity concern. We ask two questions about three open-source *instrumental* generators: Stable Audio 3 Medium (SA3), ACE-Step 1.5, and DiffRhythm 2. **(1) Does SA3's reported FAD/CLAP win survive a modern, encoder-diverse perceptual stack?** Re-ranking the same three models on the Song Describer instrumental subset with Kernel Audio Distance (KAD) on two encoders, FAD-∞, Audiobox Aesthetics, and MuQ-MuLan, we find that **SA3 ranks first on all nine metrics**, with non-overlapping bootstrap intervals on every distribution metric. The circularity that motivated the study does not materialize: SA3's lead is as large or larger on the metrics that do *not* use the vendor's encoder, so the win is not a CLAP artifact. **(2) How well do these models obey explicit tempo and key constraints?** On 240 prompts that name a target tempo and key, SA3 obeys (tempo within 4% on 61% of clips, exact key on 64%), DiffRhythm 2 partially, and ACE-Step *near chance*. Explicit controllability is the dimension that most separates the models. SA3 thus leads on *both* perceptual quality and controllability, while the open comparators trade off differently below it. The harness, prompts, per-clip measurements, and reproducible recipe are available from the authors on request.

1 Introduction

Text-to-music systems are typically ranked in their own technical reports using FAD and CLAP score. Both are commonly computed with a LAION-CLAP encoder; when a model is trained with a CLAP objective against that same encoder, the evaluation and the training signal share representation, which can flatter the model on exactly the metric used to crown it [1, 8]. Meanwhile a body of work shows FAD is biased at the few-hundred-sample sizes typical of instrumental references, and that generic FAD/CLAP track human preference only weakly [2, 3, 9].

This paper does not propose a new model. It is a *finding-oriented evaluation*: we take the three open-source instrumental generators that a recent vendor report [1] compares (SA3 Medium, ACE-Step 1.5, DiffRhythm 2) and subject them to two scrutinies. First, a re-ranking on a modern, deliberately encoder-diverse perceptual stack, to test whether the reported ranking is a property of the audio or of the metric. Second, a controllability audit: when a prompt explicitly states a tempo and a key, does the generated audio actually have them? The second question is, to our knowledge, under-reported for open instrumental models and is the paper's main contribution.

2 Background and Related Work

The metric under test. The SA3 report [1] (Table 3, Song Describer 120 s instrumental) ranks SA3-medium > ACE-Step 1.5 > DiffRhythm 2 by FAD and CLAP, both on LAION-CLAP `630k-audioset-best`. We treat this as the claim to be tested, not as ground truth, for the circularity reason above.

Modern distribution metrics. KAD [2] is an unbiased kernel (MMD^2) distance that converges at the few-hundred-sample sizes where FAD is biased and correlates better with human ratings; FAD-∞ [3] extrapolates FAD to the infinite-sample limit. We report both on two encoders to expose encoder sensitivity.

Perceptual and adherence tooling. Audiobox Aesthetics [4] provides per-clip Production Quality (PQ) and Content Enjoyment (CE). MuQ-MuLan [5] is a modern audio–text contrastive model used as a CLAP replacement for prompt adherence. For Finding 2 we use standard MIR conventions: tempo accuracy (Acc1/Acc2) and the `mir_eval` Mirex-weighted key score [6], over the Song Describer dataset's permissively licensed audio [7].

3 Method

3.1 Models

SA3 Medium (`stabilityai/stable-audio-3-medium`), ACE-Step 1.5 (`ace-step/ACE-Step`), DiffRhythm 2 (`ASLP-lab/DiffRhythm2`). Instrumental mode only; no training or fine-tuning. Per-prompt seeds are deterministic and shared across models.

3.2 Finding 1: perceptual re-ranking

We reproduce the vendor's eval set: from the validated Song Describer captions we select instrumental captions, yielding **395** caption–track pairs. The filter is explicit and provided as code (`select_finding1.py`, available on request): SDD's coherence-validated subset (746 captions), then duration $\geq$ 120 s (712), then drop any caption with a whole-word vocal mention (vocals / singing / rap / speech / choir; 712 → 395). The vendor reports 424 (their keyword list and a manual coherence pass are unpublished, so 424 is not bit-reproducible); an earlier, looser keyword pass of ours gave 416. The three counts agree within 7% and, critically, the ranking is *identical* across them (§4.2). Each model generates a 120 s clip per caption (395 clips/model). The KAD/FAD reference is the real audio of the **329** unique source tracks, trimmed to 120 s and loudness-normalized (EBU R128, −14 LUFS) to suppress loudness-driven metric gaming.

Distribution metrics are computed from per-clip embeddings on two encoders: LAION-CLAP `630k-audioset-best` (the vendor encoder, to expose circularity) and `music_audioset` (a music-trained encoder, as a second view). We report FAD (Fréchet), **KAD** (unbiased MMD^2, RBF kernel, median-heuristic bandwidth, α=100), and FAD-∞, with 95% bootstrap intervals (300 resamples). Per-clip we report Audiobox PQ/CE, MuQ-MuLan cosine adherence, and the vendor CLAP score (cosine of audio/text embeddings on the vendor encoder).

3.3 Finding 2: tempo/key controllability

We construct **240** constraint prompts as a grid: 6 styles × 5 tempos $\{70, 90, 110, 130, 150\}$ BPM × 8 keys (4 major, 4 minor), each phrased in natural language ("An instrumental *lo-fi hip-hop beat* at *90 BPM* in the key of *C major*."). The *request is the ground truth.* Each model generates a 30 s

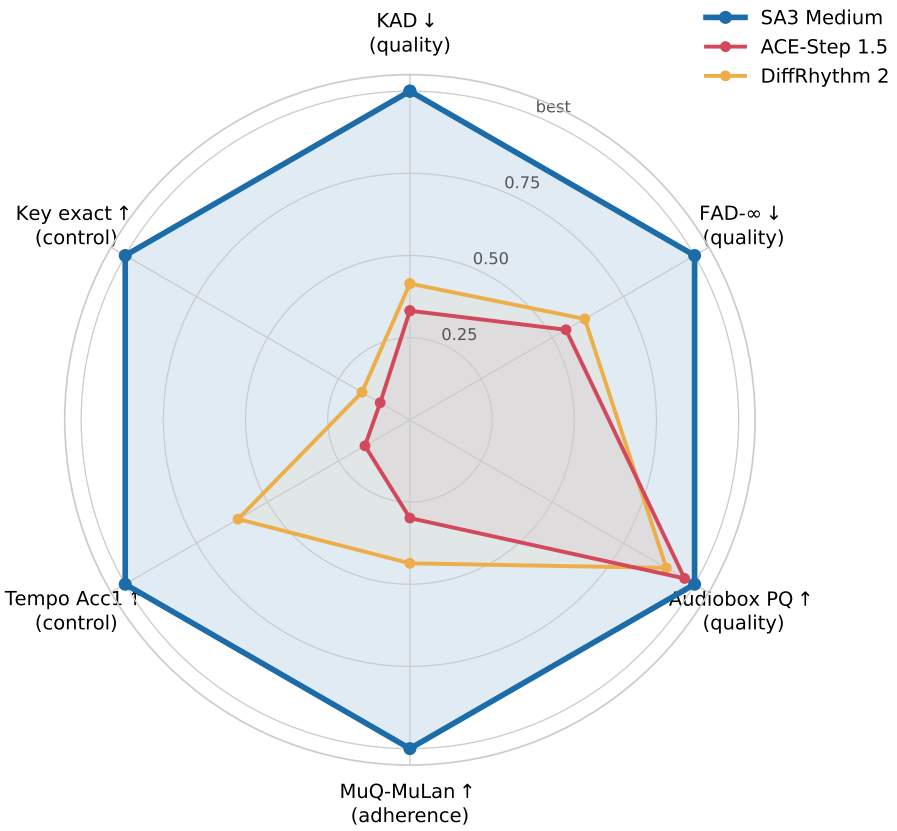

Figure 1: Both findings at a glance. Six representative metrics spanning perceptual quality (KAD, FAD-∞, Audiobox PQ, MuQ-MuLan) and constraint control (tempo, key); each axis is normalized to fraction-of-best (best model = 1, lower-better metrics inverted, so *outward* = *better*). SA3 Medium is the outer envelope on every axis. The field bunches near the edge only on Audiobox PQ (raw audio quality, where the models are close) and collapses inward on the control axes, where ACE-Step is near chance. Numbers in Tables 1–2.

clip per prompt (720 total). From the audio we estimate tempo (librosa) and key (Krumhansl–Schmuckler on CQT chroma) and score: tempo **Acc1** ($|\hat{t} - t|/t \leq 0.04$), **Acc2** (octave-tolerant over $\{\frac{1}{3}, \frac{1}{2}, 1, 2, 3\}$), and absolute BPM error; key by the `mir_eval` Mirex-weighted score (full / fifth / relative / parallel credit) and exact match. Because tempo/key estimators have finite ceilings ($\sim$0.85 tempo Acc1, $\sim$0.72 key accuracy), they cap all models equally and the cross-model comparison remains fair.

4 Results

4.1 Finding 1: the win survives

Table 1 gives the full stack (Figure 1 summarizes both findings visually). **SA3 ranks first on every metric**: the vendor FAD/CLAP baseline *and* the modern metrics that do not share its encoder (KAD on both encoders, Audiobox, MuQ-MuLan). On the distribution metrics SA3's 95% bootstrap intervals do not overlap either comparator (e.g. FAD_{630k} SA3 $[0.180, 0.205]$ vs. ACE $[0.309, 0.331]$ vs. DiffRhythm $[0.272, 0.307]$; KAD_{630k} SA3 $[2.15, 2.94]$ vs. $[6.82, 7.71]$ vs. $[5.25, 6.48]$), so the lead is statistically clear. Our vendor CLAP for SA3 reproduces the report almost exactly

(0.395 vs. 0.390), validating the harness.

Table 1: Finding 1: modern perceptual stack on 395 instrumental clips/model vs. 329 Song Describer references. Best per row in **bold**. ↓ lower better, ↑ higher.

Metric	**SA3 Medium**	ACE-Step 1.5	DiffRhythm 2
FAD · 630k ↓	**0.168**	0.299	0.265
FAD-∞ · 630k ↓	**0.148**	0.270	0.241
KAD · 630k ↓	**2.35**	7.08	5.67
FAD · music ↓	**0.168**	0.365	0.348
KAD · music ↓	**2.28**	11.57	9.40
CLAP · 630k (vendor) ↑	**0.395**	0.290	0.276
Audiobox PQ ↑	**7.65**	7.38	6.89
Audiobox CE ↑	**7.05**	6.78	6.28
MuQ-MuLan ↑	**0.385**	0.115	0.168

Comparison to the vendor report. The vendor (FAD/CLAP) reported 0.107/0.390, 0.193/0.321, 0.293/0.158 for SA3, ACE, DiffRhythm. Our CLAP ranking matches (SA3 > ACE > DiffRhythm); our FAD is uniformly higher (loudness normalization, subset, chunking) but preserves the SA3-$\gg$-rest gap. The re-order vs. the vendor is at 2nd/3rd: we place DiffRhythm 2 ahead of ACE-Step on FAD (0.265 vs. 0.299), KAD, and MuQ, whereas the vendor placed ACE clearly ahead. Thus the comparators reorder below SA3, but SA3's top rank is untouched and is corroborated by three metric families the vendor did not use.

4.2 Subset robustness

The instrumental subset can be drawn three ways: the vendor's 424 (unpublished filter), our as-run 416 (a looser keyword pass that let ∼21 incidental vocal mentions, e.g. "female singers", leak through), and the canonical 395 from the released whole-word filter. Re-computing the entire stack on the clean 395 leaves the ranking *identical on all metrics* and barely moves SA3 (FAD 0.164 → 0.168, CLAP 0.392 → 0.395). Removing the vocal leaks actually *improves* the comparators, most visibly DiffRhythm 2 (FAD 0.309 → 0.265), since the leaked captions were vocal and scored poorly. This is what resolves the ACE/DiffRhythm FAD tie seen on the 416 set into a clean DiffRhythm-second ordering. A 5% change in subset composition (larger than the 8-pair gap to the vendor's 424) is in any case dwarfed by the bootstrap, which resamples the set with replacement (∼37% dropped per draw) and still yields non-overlapping intervals. The ranking is not a subset artifact. Table 1 reports the clean 395 numbers.

4.3 Finding 2: controllability is not uniform

Table 2 shows that explicit tempo/key obedience differs by almost an order of magnitude. **SA3 genuinely controls both**: a median BPM error of **2.3** (it lands almost exactly on the requested tempo for most prompts) and 64% exact key, near the estimator ceiling. The Acc2>Acc1 gap (0.75 vs. 0.61) shows SA3's tempo misses are dominated by octave (half/double-time) errors, not random ones. **ACE-Step is statistically at chance.** Uniform-random exact-key accuracy over the 24 major/minor keys is $1/24 \approx 0.042$; ACE-Step's 0.067 (0.070±0.017 over three seeds, §4.4) is within noise of it. Its tempo Acc1 (0.10) sits just above the ∼0.07 expected from a tempo drawn uniformly over 60–180 BPM (the ±4% window covers ∼7% of that range), i.e. marginally above, but nowhere near control. ACE generates coherent audio (it was competitive in Finding 1) but

does not condition on natural-language tempo/key. DiffRhythm 2 is intermediate on tempo and weak on key. The ranking SA3 $\gg$ DiffRhythm 2 $>$ ACE-Step is unanimous across tempo Acc1, Acc2, and key.

Table 2: Finding 2: tempo/key adherence to explicit constraints (240 prompts/model). Ground truth = the request. Best per column in **bold**.

Model	Tempo Acc1 ↑	Tempo Acc2 ↑	BPM MAE ↓	BPM med. err ↓	Key wt. ↑	Key exact ↑
SA3 Medium	**0.608**	**0.746**	**19.2**	**2.3**	**0.695**	**0.642**
DiffRhythm 2	0.367	0.508	25.5	7.5	0.170	0.108
ACE-Step 1.5	0.096	0.133	31.1	27.0	0.122	0.067

Where control succeeds (SA3, per style). Tempo control is strongest where there is a clear beat, with jazz piano trio at Acc1 0.84 ± 0.08 and lo-fi at 0.76 ± 0.03 (3-seed mean±std, §4.4), and weakest for ambient pad (0.11 ± 0.07); the ambient figure is largely a *measurement* floor (pad textures have no clear pulse), since ambient has SA3's *highest* exact-key rate (0.82, sustained harmony makes key unambiguous). The per-style ordering is stable across seeds, though individual cells (jazz, ambient) carry the largest spread. The single-seed jazz figure of 0.95 was a favorable draw. Control is real and structured, not uniform across content.

4.4 Seed robustness

To separate model capability from seed luck, we regenerated all 240 constraint prompts under three independent draws per model (SA3 and ACE-Step with salted deterministic seeds; DiffRhythm 2 is inherently non-deterministic, since its inference script never seeds `random`, so re-runs are independent samples). Table 3 reports the mean±std. **Per-seed standard deviations (∼0.01–0.03) are far smaller than the between-model gaps (0.19–0.50)**: SA3's tempo lead over DiffRhythm 2 is ≈4.6 standard deviations, so the ranking is stable to seed choice. The aggregate headline numbers are likewise robust: SA3's single-run values (Acc1 0.608, key exact 0.642) lie within one standard deviation of the 3-seed means. The seed sensitivity that does exist is concentrated in *per-style* cells with weak rhythmic content (ambient, ±0.07) or small favorable draws (jazz, single-seed 0.95 vs. 0.84 mean), not in the model ranking or the aggregates.

Table 3: Seed robustness: adherence as mean±std over 3 independent generations of all 240 prompts. The ranking is unchanged on every metric; model gaps dwarf seed variance.

Model	Tempo Acc1	Tempo Acc2	Key weighted	Key exact
SA3 Medium	**0.583±0.024**	**0.726±0.016**	**0.684±0.012**	**0.622±0.019**
DiffRhythm 2	0.397±0.032	0.523±0.025	0.140±0.023	0.078±0.024
ACE-Step 1.5	0.082±0.010	0.119±0.010	0.122±0.006	0.070±0.017

5 Discussion

What the two findings say together. Read jointly, the results are less about crowning SA3 than about what standard benchmarking can and cannot see. Finding 1 lives on the perceptual axis that FAD/CLAP reports already target; there SA3's win is real and survives an independent stack, but it is the axis on which these models are converging. Finding 2 measures an axis those reports omit entirely, and there the same three models separate by nearly an order of magnitude (tempo Acc1

0.61 vs. 0.10, key 0.64 vs. 0.07). The practically decisive difference between open instrumental generators is therefore not how good they sound, on which they are comparatively close, but whether they do what they are told, and that is precisely the dimension current leaderboards do not measure. Controllability is invisible to the benchmarks that rank these systems.

Did the circularity materialize? The study is motivated by the risk that the vendor's CLAP encoder, the family SA3 trains against, flatters SA3. If it did, SA3's lead would *shrink* on metrics not built on that encoder. It does not. SA3's separation is as large or larger on the independent metrics: its distance on the music-encoder KAD is 4.1× smaller than the runner-up's (vs. 2.4× on the 630k KAD it ostensibly benefits from), and on MuQ-MuLan, a different contrastive model entirely, it leads by 2.3×; the only narrow margin (Audiobox PQ, +4%) is itself an independent metric, not the circular one. So the circularity we flagged is real in principle but did *not* materialize for these three models. If anything, the vendor encoder slightly understated SA3's lead. We state this as the negative result it is, rather than leaving circularity an unresolved motivation: the contribution of Finding 1 is not that circularity flips the ranking, but that an independent, less-biased stack *confirms* it.

Do these metrics predict human preference? We ran no listening test, so we cannot answer this directly for our models, the study's deepest limitation (below). Two points nonetheless set our stack apart from the FAD/CLAP ranking it scrutinizes. First, the metrics were chosen for *published* human-correlation evidence: KAD was introduced precisely because it tracks human ratings better than FAD (Spearman ≈ -0.93 vs. -0.80) [2], and Audiobox Aesthetics is a quality predictor *trained on human annotations* [4]. Neither is a generic distribution distance. Second, the three agreeing families (distributional KAD/FAD, learned-aesthetic Audiobox, contrastive MuQ-MuLan) are methodologically independent, so their unanimous ranking is harder to ascribe to a shared blind spot than three variants of FAD would be. These are arguments from the literature and from metric diversity, not human data on *these* models, which is the first limitation below.

Limitations. (i) *No human evaluation: the deepest limitation.* Every result is an automated proxy for human judgment; although our metrics carry the published human-correlation evidence noted above, we collected no human preference on these three models, so Finding 1's perceptual claim rests on metric consensus rather than direct listening. A pairwise listening test is the primary planned extension as we widen the model set. It would additionally let us measure each metric's correlation with human preference on our own data, the very question raised above. (ii) The vendor's exact 424-pair subset is not reproducible (their keyword list and manual pass are unpublished); our filter is provided as code and yields 395, with the 416/424 variants agreeing within 7% and the ranking invariant (§4.2). (iii) FAD's bootstrap intervals sit above the point estimate, reflecting FAD's known upward bias under resampling; we therefore lead with KAD and report FAD for comparability. (iv) Finding 2 uses librosa + Krumhansl–Schmuckler estimators rather than the heavier madmom/essentia stack, for robustness; estimator ceilings cap all models equally but absolute adherence rates are lower bounds. (v) ACE-Step exposes structured tempo/key conditioning fields that we deliberately did not use. Finding 2 measures *natural-language* obedience specifically, the realistic prompting interface; ACE may fare better with structured control inputs. (vi) Finding 2 results are quantified over three seeds (§4.4); the model ranking is stable but individual per-style cells carry larger spread, and a madmom/essentia key-estimator sensitivity pass remains future work.

Takeaway for practitioners. If you need a backing track in a specific key for a vocalist to sing over, or at a specific tempo to cut to picture, SA3 Medium is currently the only open instrumental generator that reliably delivers it from a text prompt: it lands the requested tempo ∼60% of the time and the exact key ∼64%. DiffRhythm 2 hits tempo about a third of the time and key rarely; ACE-Step effectively ignores the request. When the constraint *is* the deliverable, this

gap dwarfs the audio-quality differences that leaderboards rank on, and it is exactly the axis those leaderboards omit. Until a model closes it, treat explicit tempo/key as a model-selection criterion, not a prompt detail.

6 Conclusion

On a modern, encoder-diverse perceptual stack, SA3 Medium's reported win over ACE-Step 1.5 and DiffRhythm 2 survives and is corroborated by metrics outside the vendor's encoder family. SA3 is also the only one of the three that meaningfully obeys explicit tempo/key constraints. Controllability, not raw audio quality, is where these open models differ most, and it is invisible to the FAD/CLAP rankings in current reports. The findings rest on automated metrics chosen for their human-correlation pedigree; anchoring them with a listening test, and correlating each metric against human preference on these models, is the natural next step as we broaden the comparison.

References

[1] Stability AI. *Stable Audio 3 Technical Report.* arXiv:2605.17991, 2026.

[2] *KAD: Kernel Audio Distance for evaluating generative audio.* arXiv:2502.15602, 2025.

[3] *Adapting Fréchet Audio Distance for Generative Music Evaluation (FAD-∞).* 2024.

[4] *Audiobox Aesthetics: Unified Automatic Quality Assessment for Audio.* arXiv:2502.05139, 2025.

[5] *MuQ / MuQ-MuLan: Self-Supervised Music Representation and Audio–Text Alignment.* arXiv:2501.01108, 2025.

[6] C. Raffel et al. *mir_eval: A Transparent Implementation of Common MIR Metrics.* ISMIR, 2014.

[7] I. Manco et al. *The Song Describer Dataset.* arXiv:2311.10057 / Zenodo 10072001, 2023.

[8] *DRAGON: Distributional Rewards for Generative Audio (instrumental eval harness).* arXiv:2504.15217, 2025.

[9] *A Human Study of Music-Generation Preference and Metric Alignment.* arXiv:2506.19085, 2025.