The *Name-Free Gap*: Policy-Aware Stylistic Control in Music Generation

Ashwin Nagarajan
University of California, Santa Cruz
asnagara@ucsc.edu

Hao-Wen Dong
University of Michigan
hwdong@umich.edu

Abstract

Text-to-music models capture broad attributes such as instrumentation or mood, but fine-grained stylistic control remains an open challenge. Existing stylization methods typically require retraining or specialized conditioning, which complicates reproducibility and limits policy compliance when artist names are restricted. We study whether lightweight, human-readable modifiers sampled from a large language model can provide a policy-robust alternative for stylistic control. Using MusicGen-small, we evaluate two artists: Billie Eilish (vocal pop) and Ludovico Einaudi (instrumental piano). For each artist, we use fifteen reference excerpts and evaluate matched seeds under three conditions: baseline prompts, artist-name prompts, and five descriptor sets. All prompts are generated using a large language model. Evaluation uses both VGGish and CLAP embeddings with distributional and per-clip similarity measures, including a new min-distance attribution metric. Results show that artist names are the strongest control signal across both artists, while name-free descriptors recover much of this effect. This highlights that existing safeguards such as the restriction of artist names in music generation prompts may not fully prevent style imitation. Cross-artist transfers reduce alignment, showing that descriptors encode targeted stylistic cues. We also present a descriptor table across ten contemporary artists to illustrate the breadth of the tokens. Together these findings define the *name-free gap*, the controllability difference between artist-name prompts and policy-compliant descriptors, shown through a reproducible evaluation protocol for prompt-level controllability.

1 Introduction

Recent progress in text-to-music generation has enabled models such as MusicLM [Agostinelli et al., 2023], MusicGen [Copet et al., 2023], and MusicLDM [Chen et al., 2023] to produce coherent audio from natural language prompts. While these systems capture broad attributes like instrumentation or mood, they offer limited control over the stylistic cues that define an artist's identity. Stylization methods typically require costly retraining or additional adapters, which hinders reproducibility for users. In addition, stylization using artist names often conflicts with deployment policies of text-to-music platforms, motivating the need for policy-compliant alternatives.

We ask whether lightweight, human-readable modifiers sampled from a large language model can provide a policy-compliant alternative for stylistic control. Figure 1 outlines this setup. These descriptors are interpretable, require no model modifications, and can be generated for any artist without using their audio, which avoids copyright issues and supports reproducibility. To test this idea, we evaluate two artists: Billie Eilish, a vocal-driven pop artist, and Ludovico Einaudi, a piano-based instrumental composer. Our contributions include: (1) a reproducible protocol for evaluating prompt-level controllability, (2) baselines and controls including artist-name prompts, min-distance attribution, and cross-artist transfer, and (3) evidence that LLM-sampled descriptors induce measurable, artist-specific shifts while recovering much of the effect of artist names. Together

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

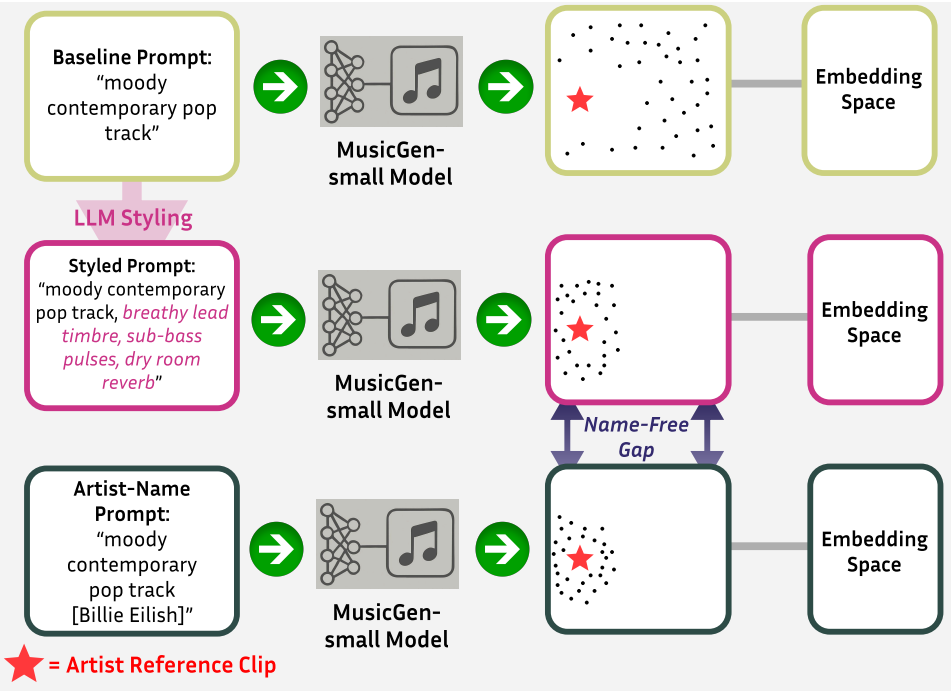

Figure 1: **Prompt-level controllability and the *name-free gap*:** A baseline prompt (top) produces dispersed generations in embedding space relative to the artist reference clip. Adding LLM-sampled descriptors shifts samples closer to the references, while artist-name prompts achieve the strongest alignment. The difference between descriptor and artist-name conditions defines the *name-free gap*.

these results define the *name-free gap*, highlight that our descriptors can achieve similar stylistic effects to artist names, and motivate strategies for controllable music generation that remain accessible under platform policies. For reproducibility, the appendix provides links to GitHub, HuggingFace, and a website with all code, manifests, embeddings, generated audio samples, and demos.[1]

2 Related Work

Text-to-music generation. Recent models such as MusicLM [Agostinelli et al., 2023] generate coherent audio from text, and MusicGen [Copet et al., 2023] shows that a single autoregressive model can support both text and melody conditioning. However, controllability remains limited when relying only on free-form prompts.

Stylization. Prior work on Stylization often relies on finetuned embeddings or learned tokens, including textual inversion for images [Gal et al., 2022], prompt-to-prompt edits [Hertz et al., 2022], and style tokens for speech [Wang et al., 2018]. In music, Rouard et al. [Rouard et al., 2024] introduce discrete bottleneck features that require model adaptation. Our approach instead uses lightweight, human-readable descriptors sampled from a language model, which do not require optimization and provide a policy-compliant alternative to artist names.

Evaluation. Generative audio is typically evaluated in embedding space. VGGish underpins Fréchet Audio Distance (FAD) [Hershey et al., 2017, Kilgour et al., 2019], while CLAP provides broader joint audio text representations [Elizalde et al., 2022, Wu et al., 2023]. We evaluate using FAD and a new min-distance attribution metric in both spaces, along with cross-artist validation as a specificity control, while noting that embedding-based scores remain imperfect proxies for perception [Huang et al., 2025].

[1] https://artisticstyles.github.io/music-style-control-demo/

Artist	Representative LLM-sampled descriptors
Billie Eilish	breathy lead timbre; sub-bass pulses; dry room reverb
Ludovico Einaudi	delicate piano timbre; string ensemble pad; intimate mix
Zayn Malik	airy vocal texture; deep sub bass; minimal trap rhythm
The Weeknd	airy falsetto timbre; analog synth bass; dark reverberant space
Tyler, The Creator	gritty vocal texture; distorted synth bass; experimental offbeat groove
Bruno Mars	smooth falsetto tone; retro synth stabs; groovy disco bass
Taylor Swift	warm lead timbre; acoustic guitar strums; intimate close space
Ariana Grande	silky lead timbre; glossy synth chords; wide reverb space
The Beatles	warm lead timbre; jangly electric guitar; tight backbeat groove
Drake	hazy vocal timbre; 808 bass hits; moody trap groove

Table 1: **Sample LLM descriptors:** The table shows the two evaluated artists and eight additional contemporary artists, each with three LLM-generated descriptors. In the experiments, all three descriptors are appended to the baseline prompt to form styled prompts; see Tables 2 and 3 for the exact prompt strings used.

3 Methods

3.1 Data and Generation

We test whether lightweight textual modifiers can steer generated outputs toward the distribution of a target artist. For each artist, we use a large language model to generate a baseline prompt that captures broad instrumentation and mood. For example, "moody contemporary pop track" for Billie Eilish and "instrumental track with gentle dynamics" for Ludovico Einaudi. Candidate descriptors are proposed by a large language model, with five independent sets sampled per artist. Each styled prompt is formed by appending one of these sets to the baseline. An additional baseline appends only the artist name, providing a control for trivial name-based stylization. Table 1 shows sample descriptors for ten artists, showing that the language model produces concise tokens reflecting stylistic traits across diverse genres. For the prompt used to query the LLM and other specific method details, refer to Appendix D.

Artists are represented by fifteen 15-second reference excerpts, resampled to 32 kHz and used only for embedding-based evaluation. All experiments use the publicly released `musicgen-small` checkpoint without modification. There are seventy total generated audio files per artist (fifty styled clips, ten baseline, and ten artist-name baseline clips).

3.2 Evaluation and Controls

Evaluation is performed in two embedding spaces to control for embedding choice: VGGish [Hershey et al., 2017] and CLAP [Elizalde et al., 2022, Wu et al., 2023]. We focus on three measures: (i) FAD to quantify distributional similarity, (ii) min-distance attribution as a per-clip similarity metric, and (iii) cross-artist validation as a specificity control.

Min-distance attribution. For each generated clip embedding $\mathbf{g}_i$ and reference set $\{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$, we compute

$$d_{\text{min}}(\mathbf{g}_i) = \min_{j=1}^{N} \left(1 - \frac{\mathbf{g}_i \cdot \mathbf{r}_j}{\|\mathbf{g}_i\| \|\mathbf{r}_j\|}\right), \tag{1}$$

the cosine distance to the nearest reference. For each condition c, we report the raw median of these min-distance values: $\text{Median}(d_c)$. Lower values indicate closer alignment of generated clips to the target reference style.

Specificity controls. To test whether descriptors are artist-specific rather than generically helpful, we adopt a cross-artist validation setup. Styled prompts sampled for one artist are evaluated against reference sets for both artists, and per-clip cosine similarities to the reference centroid are averaged. Improvements are computed relative to the baseline as

$$\Delta = \text{mean}(\text{styled centroid similarities}) - \text{mean}(\text{baseline centroid similarities}). \tag{2}$$

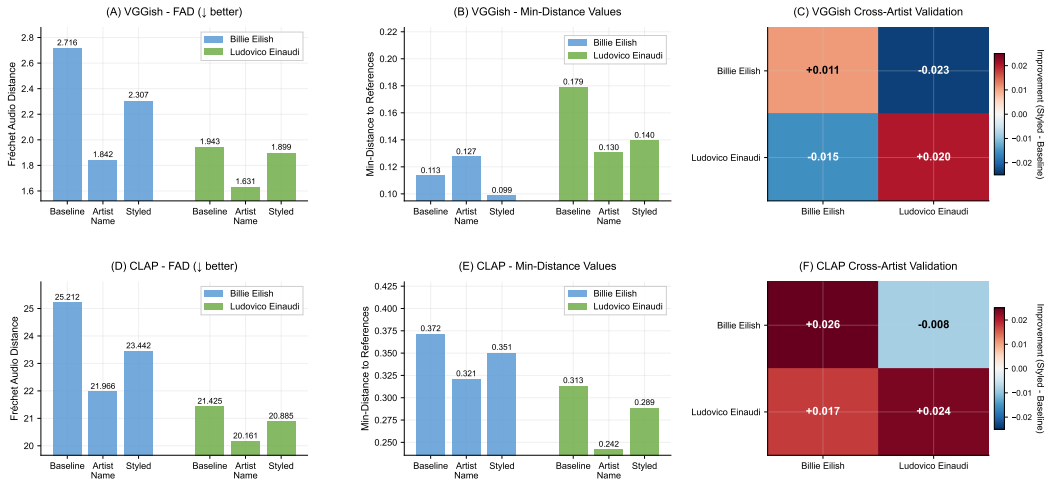

Figure 2: **Dual embedding analysis:** (A,D): Comparison of both CLAP and VGGish FAD for both artists. (B, E): Comparison of min-distance values in CLAP and VGGish for both artists. (C, F): Cross-artist validation in both CLAP and VGGish.

Higher positive diagonal values (same-artist) and negative or lower off-diagonal values (cross-artist) indicate that descriptors encode targeted stylistic cues rather than generic quality gains.

4 Results and Discussion

Main Findings. Across both artists and embedding spaces, artist-name prompts provide the strongest improvements. Descriptor-augmented prompts consistently improve over the baseline and close a substantial portion of the *name-free gap*. Panels A and D in Figure 2 show that artist names achieve the lowest FAD, with descriptors narrowing the difference. Panels B and E show reduced median min-distance under both artist-name and descriptor conditions relative to the baseline, though for Billie Eilish in Panel B the artist-name prompts actually perform slightly worse than the baseline, while the descriptor condition achieves the best result.

Specificity and Controls. Panels C and F test specificity: diagonal cells (same artist) are higher positive values, while off-diagonal cells (cross artist) are negative or lower positive values, indicating that descriptors capture targeted stylistic cues rather than generic quality gains.

Discussion. Our findings show that prompt-level modifiers provide a lightweight means of steering generative music without retraining. Artist names act as strong style carriers, and name-free descriptors recover a substantial portion of this effect, which suggests that restricting artist names alone may not fully prevent style imitation. Limitations include evaluation on two artists, short 15 second reference clips, and reliance on embedding-based proxies that do not replace listening tests; results are also sensitive to embedding choice, with VGGish and CLAP not always agreeing.

5 Contributions and Future Works

This work contributes a reproducible framework for evaluating prompt-level controllability in generative audio, combining matched seeds, fixed references, and evaluation in two embedding spaces, with controls such as artist-name prompts and cross-artist transfer, and metrics including FAD and min-distance attribution. Using this framework, we quantify the *name-free gap* between artist-name prompts and policy-compliant descriptors, and outline how it can organize future research questions. While demonstrated with two artists, the approach is model-agnostic and applies broadly across music and related generative audio tasks. Future directions include expanding the artist and genre set, testing longer clips and musical structure, incorporating human listening studies, exploring alternative embeddings and robustness analyses, and developing systematic descriptor design and benchmarks for policy-compliant controllability.

References

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.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.05284.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. *arXiv preprint arXiv:2308.01546*, 2023. URL https://arxiv.org/abs/2308.01546.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018.

Simon Rouard, Yossi Adi, Jade Copet, Axel Roebel, and Alexandre Défossez. Audio conditioning for music generation via discrete bottleneck features. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024. URL https://arxiv.org/abs/2407.12563.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, Malcolm Slaney, Ron J Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

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

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. *arXiv preprint arXiv:2206.04769*, 2022.

Yusong Wu, Hao Cheng, Zhifeng Chen, Nan Duan, Shujie Wang, Ming Chen, Michael Zeng, and Xiaodong Liu. CLAP: Contrastive language-audio pretraining. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. URL https://arxiv.org/abs/2211.06687.

Yichen Huang, Zachary Novack, Koichi Saito, Jiatong Shi, Shinji Watanabe, Yuki Mitsufuji, John Thickstun, and Chris Donahue. Aligning text-to-music evaluation with human preferences. *arXiv preprint arXiv:2503.16669*, 2025.

OpenAI. Gpt-5: Large language model. https://openai.com/, 2025. Used to generate descriptor candidates via the ChatGPT web interface.

Modan Tailleur, Junwon Lee, Mathieu Lagrange, Keunwoo Choi, Laurie M. Heller, Keisuke Imoto, and Yuki Okamoto. Correlation of fréchet audio distance with human perception of environmental audio is embedding dependant. *arXiv preprint arXiv:2403.17508*, 2024.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting fréchet audio distance for generative music evaluation. In *ICASSP*, 2024. arXiv:2311.01616.

A Extended Overview

We include one vocal pop artist (Billie Eilish) and one instrumental piano artist (Ludovico Einaudi) to cover common use cases, and this pairing helps us cover both vocal and instrumental production regimes. The study compares baseline prompts, prompts augmented with artist names, and prompts augmented with name-free descriptors to quantify prompt-level controllability.

Each artist is represented by fifteen 15-second reference excerpts. Outputs are generated under three conditions: baseline prompts, prompts augmented with the artist name, and prompts augmented with five independent three-token descriptor sets sampled from a large language model; we used GPT-5 in our experiments. Matched seeds are used across all conditions to control for stochastic variability.

Evaluation is conducted in two embedding spaces: VGGish, following the canonical definition of Fréchet Audio Distance (FAD) [Hershey et al., 2017, Kilgour et al., 2019], and CLAP [Elizalde et al., 2022, Wu et al., 2023], a joint audio text model. We report FAD and min-distance attribution in both spaces, and we perform cross-artist validation as a specificity control. Together, these measures capture distributional alignment, per-clip proximity to references, and specificity.

B Reproducibility Checklist

Reference excerpts are drawn from publicly available recordings and evaluated only in 15-second segments. A manifest of excerpt IDs and timestamps is released, while the audio itself is not redistributed. All experiments use the publicly available `musicgen-small` checkpoint without modification. Candidate descriptors are generated automatically by GPT-5 [OpenAI, 2025], and the full sampled sets are provided in a manifest file.

Generation and evaluation scripts are released, with control over prompts and random seeds, as well as embedding extraction. Matched seeds are passed directly into the generation script by setting `torch.manual_seed(s)` before each run. This ensures that, for each seed, one baseline clip, one artist-name clip, and five styled clips (one per descriptor set) are generated under identical random initialization. As a result, differences between conditions arise only from the prompt modifiers rather than random noise, enabling paired evaluation.

Evaluation is conducted in both VGGish and CLAP embedding spaces. In VGGish, we report Fréchet Audio Distance (FAD_{vgg}) following prior definitions [Hershey et al., 2017, Kilgour et al., 2019]. For CLAP, we use an open-source implementation [Elizalde et al., 2022], computing analogous FAD_{clap}. In both spaces we additionally report min-distance attribution as a per-clip metric, using raw median values. Results are aggregated across descriptor sets; FAD is reported as mean $\pm$ standard deviation, while min-distance is reported as raw medians.

Released artifacts include descriptors, reference manifests, embeddings, metrics, and figure scripts, enabling exact reproduction of reported results without requiring access to copyrighted recordings.

All experiments were lightweight. Each generation condition (10 seeds $\times$ 15 s clips at 32 kHz) ran on a consumer laptop (Apple M2, 24 GB RAM). Total compute across all conditions was under 8 CPU hours. This modest requirement enables reproduction without specialized hardware or access to GPUs.

C Extended Related Work

Style representations in speech. Global Style Tokens (GST) learn style embeddings in TTS models via a reference encoder, enabling prosody control without explicit supervision [Wang et al., 2018]. We treat GST as a conceptual precedent, while our approach uses training-free, prompt-level interventions for music generation.

Embedding spaces for evaluation. Generative audio evaluation relies heavily on pretrained embeddings. VGGish underpins the canonical definition of Fr'echet Audio Distance (FAD) [Hershey et al., 2017, Kilgour et al., 2019], while CLAP provides joint audio text representations used for retrieval and cross-modal tasks [Elizalde et al., 2022]. Recent work shows that metric outcomes depend on embedding choice [Tailleur et al., 2024], and adaptations of FAD for music evaluation

Billie Eilish	Prompt text
Baseline	a moody contemporary pop track with subtle electronic textures, minimal percussion, and an atmospheric groove
Styled prompt 1	a moody contemporary pop track with subtle electronic textures, minimal percussion, and an atmospheric groove, breathy lead timbre, sub-bass pulses, dry room reverb
Styled prompt 2	a moody contemporary pop track with subtle electronic textures, minimal percussion, and an atmospheric groove, close-mic lead timbre, sparse percussion, intimate mix space
Styled prompt 3	a moody contemporary pop track with subtle electronic textures, minimal percussion, and an atmospheric groove, airy lead timbre, 808 bass tones, slow sparse groove
Styled prompt 4	a moody contemporary pop track with subtle electronic textures, minimal percussion, and an atmospheric groove, soft lead texture, detuned analog pad, minor-key low tempo
Styled prompt 5	a moody contemporary pop track with subtle electronic textures, minimal percussion, and an atmospheric groove, delicate lead timbre, distorted bass texture, syncopated glitch rhythm

Table 2: **Billie Eilish prompts.** Baseline and five styled prompts formed by appending the three-token sets.

Ludovico Einaudi	Prompt text
Baseline	contemporary instrumental track with gentle dynamics, melodic progression, and subtle harmonic textures
Styled prompt 1	contemporary instrumental track with gentle dynamics, melodic progression, and subtle harmonic textures, solo piano texture, repetitive arpeggios, classical reverb
Styled prompt 2	contemporary instrumental track with gentle dynamics, melodic progression, and subtle harmonic textures, delicate piano timbre, string ensemble pad, intimate mix
Styled prompt 3	contemporary instrumental track with gentle dynamics, melodic progression, and subtle harmonic textures, flowing piano lead, simple chord texture, warm acoustic space
Styled prompt 4	contemporary instrumental track with gentle dynamics, melodic progression, and subtle harmonic textures, lyrical piano melody, soft dynamics, contemplative tempo
Styled prompt 5	contemporary instrumental track with gentle dynamics, melodic progression, and subtle harmonic textures, expressive piano tone, minimalist patterns, gentle sustain

Table 3: **Ludovico Einaudi prompts.** Baseline and five styled prompts formed by appending the three-token sets.

have been proposed [Gui et al., 2024]. Our study reports results in both VGGish and CLAP to control for embedding choice and provide complementary perspectives.

D Detailed Methods

D.1 Prompt Construction

Candidate descriptors are generated automatically by GPT-5 [OpenAI, 2025] using a fixed instruction template. For each artist, the model is asked to provide a neutral baseline sentence and five independent sets of three short style tokens, subject to constraints on length, format, and coverage of timbre, instrumentation, and mix or rhythm. This yields fifteen descriptors per artist, which are concatenated with the baseline to form styled prompts. For clarity, Tables 2 and 3 list the exact baseline and the five styled prompts per artist.

In addition, we define an artist-name baseline by appending the artist identifier directly to the baseline description:

$$\texttt{artist_baseline} = f\text{"}\{baseline\}[\{artist_name\}]\text{"}$$

This provides a control condition where stylization is achieved through the artist label alone. The complete instruction text given to GPT-5 is provided in Appendix D.2, and the sampled outputs are released to ensure reproducibility.

D.2 Prompt Template

The full instruction used to query GPT-5 is shown below. This template was applied identically for each artist, with the artist name inserted in the placeholder.

```
You are a veteran music producer and mix engineer. Your job is to
    produce concise production descriptors that steer the open-source
    MusicGen model.

TASK
For the artist "{{ARTIST_NAME}}", return a single JSON object
    containing:
1) "baseline": one neutral, model-friendly sentence for a contemporary
     pop track suitable for 12-15 s generations (no artist mentions);
2) "sets": exactly 5 diverse "style token" sets, each with 3 short
    tokens.

CONSTRAINTS
- Tokens are 2-4 words, lowercase ASCII, no proper nouns, no artist
    names, no lyrics, no brands.
- Avoid tokens that require realistic singing; if needed, describe the
     lead timbre (e.g., "breathy lead timbre") instead of vocals.
- Each 3-token set must cover: (i) lead timbre/texture, (ii)
    instrumentation/timbre, (iii) mix/space OR rhythm/harmony.
- The 5 sets must be meaningfully different (no near-duplicates).

OUTPUT (JSON only, no commentary, no markdown):
{
  "artist_name": "{{ARTIST_NAME}}",
  "baseline": "<one-sentence baseline prompt>",
  "sets": [
    ["t1","t2","t3"],
    ["t1","t2","t3"],
    ["t1","t2","t3"],
    ["t1","t2","t3"],
    ["t1","t2","t3"]
  ]
}
```

D.3 Reference Material

Reference excerpts are drawn from publicly available recordings, restricted to fifteen-second segments. Each excerpt is resampled to 32 kHz to match the generator output. The list of excerpt IDs and timestamps is included in the released manifest, but no copyrighted audio is redistributed.

D.4 Generation Protocol

All experiments use the publicly released `musicgen-small` checkpoint. For each artist, ten random seeds are fixed and rendered with the baseline prompt. The same seeds are then used with the artist-name baseline, yielding ten additional outputs. Each of the five descriptor sets is also applied to the same seeds, producing fifty styled outputs. In total, each artist has seventy outputs (10 baseline, 10 artist-name, 50 styled). This matched-seed design controls for randomness across conditions and allows paired comparisons.

D.5 Evaluation Metrics

Fréchet Audio Distance (FAD). We compute FAD in both VGGish and CLAP embedding spaces, following the canonical definition of Kilgour et al. [2019]. FAD_{vgg} provides comparability with prior work, while FAD_{clap} incorporates a joint audio text representation.

Min-Distance Attribution. For each generated embedding, we compute its cosine distance to the closest reference embedding. This provides a per-clip measure of nearest-neighbor alignment

and complements distributional metrics like FAD. We report raw median min-distance values per condition.

Embedding Choice. Embedding choice is known to influence evaluation outcomes [Tailleur et al., 2024]. We therefore report results in both VGGish and CLAP [Elizalde et al., 2022, Wu et al., 2023]: VGGish supports canonical FAD and comparability with prior work, while CLAP is widely adopted for cross-modal evaluation. Adaptations of FAD for music have also been proposed [Gui et al., 2024].

D.6 Statistical Testing

For each artist and condition, metrics are computed over ten matched seeds per descriptor set. We aggregate across the five descriptor sets per artist: FAD is reported as mean ± standard deviation; min-distance is reported as raw medians. All comparisons are paired by seed, ensuring that differences between baseline, artist-name, and styled prompts arise only from the prompt modifiers rather than random variability.

D.7 Control Condition

For the cross-artist control, all five descriptor sets sampled for one artist are applied to the other artist's baseline prompt using the same ten seeds. Metrics are computed against the corresponding reference set, and results are averaged across descriptor sets. Degraded alignment in this condition indicates that descriptors encode artist-specific rather than generic stylistic cues.

E Ethics, Broader Impacts, and Extended Reproducibility

Positive societal impacts. This work contributes a reproducible and interpretable framework for studying stylistic controllability in text-to-music systems. By releasing descriptors, manifests, embeddings, and evaluation scripts, we enable other researchers to conduct controlled experiments on prompt-level interventions. The use of human-readable descriptors supports transparency and may inform creative tools for lightweight stylistic steering without retraining or specialized expertise.

Risks and safeguards. As with any generative system, there is potential misuse for unauthorized style imitation. We mitigate this risk by restricting evaluation to short excerpts (15-second), avoiding redistribution of copyrighted audio, and framing alignment quantitatively in embedding space rather than via perceptual mimicry. All released artifacts exclude original recordings and are anonymized where appropriate.

Limitations and ethical considerations. The study is constrained to two artists, short (15-second) clips, and embedding-based metrics, which do not capture long-term musical structure or fully align with perception. Descriptors are sampled from a language model and may reflect biases in phrasing or coverage. Future work should extend to more artists and genres, incorporate listening studies, and test robustness across embedding spaces. This work is an initial step toward policy-compliant controllability in generative music.

Reproducibility. All reference audio is restricted to 15-second segments and is not redistributed. The demo page hosts only generated clips with metadata removed. Released artifacts include code, manifests, embeddings, and generated audio.

Artifacts

To support reproducibility, we release anonymized artifacts associated with this work:

- Code and evaluation scripts: `https://github.com/artisticstyles/artisticstyles-neurips`
- Website with representative generations: `https://artisticstyles.github.io/music-style-control-demo/`
- Full generations and reference manifests: `https://huggingface.co/datasets/ArtisticStyling/music-style-control-data`

License for released assets. All new artifacts introduced in this work (descriptor sets, manifests, embeddings, evaluation outputs, scripts, and demo website) are released under the MIT license. This permissive license ensures that other researchers can freely reproduce, adapt, and build upon our results.

Licenses for external assets. MusicGen-small is released under the CC-BY-NC 4.0 license, VGGish under the Apache-2.0 license, Microsoft CLAP under the MIT license, and the reference implementation of Fréchet Audio Distance (FAD) under the MIT license. GPT-5 was accessed via the ChatGPT web interface in accordance with OpenAI's Terms of Use. Reference audio excerpts are not redistributed; only 15-second segments were used for evaluation under fair-use principles.