
Does Artist Context Help?

Audio-Conditioned Diffusion for Album Cover Generation

Andrea Li

Department of Computing and Mathematical Sciences
California Institute of Technology
ali3@caltech.edu

Fnu Prince

Department of Computing and Mathematical Sciences
California Institute of Technology
fprince@caltech.edu

Abstract

We study whether textual artist context—a short style description, a biographical paragraph, or their combination—improves audio-conditioned album-cover generation built on top of Stable Diffusion 1.5.¹ Our system freezes the pretrained UNet and trains a low-rank adaptation (LoRA) jointly with a small projector that maps a CLAP audio embedding into eight learnable cross-attention tokens. Per-artist text is encoded by Stable Diffusion’s own CLIP text encoder and concatenated with the audio tokens along the sequence dimension, so the UNet attends to both modalities at every block. We compare four trained conditions (AUDIO_ONLY, STYLE, BIOGRAPHY, COMBINED) against a no-training caption baseline that prompts pretrained SD with the artist text alone. Experiments use 60 paired tracks from three artists (Ariana Grande, Drake, LE SSERAFIM); 12 tracks (4/5/3 per artist) are held out for evaluation. Because the dataset is small and the target is subjective visual identity, our evaluation is primarily qualitative. We find that the audio channel, once paired with LoRA fine-tuning, is sufficient to recover each artist’s dominant portrait convention (a single upward-angled face for Drake, a high-ponytail editorial portrait for Ariana Grande, a wordmark-led graphic for LE SSERAFIM), whereas a text-only caption baseline produces abstract graphics that capture none of this identity. Textual context does not override the audio signal; instead it modulates palette and the salience of typographic elements, with the largest effect on the group artist whose covers are wordmark-dominated.

1 Introduction

Album covers communicate genre, mood, persona, and aesthetic before the listener presses play. Text-to-image diffusion models such as Stable Diffusion [Rombach et al., 2022] have made high-quality conditional image generation cheap, but generating images from *music* is harder because the input is not symbolic language: it carries timbre, rhythm, vocal character, harmony, and emotional tone. Bridging audio and visual style requires turning sound into semantic and aesthetic signals an image model can attend to. Recent work in audio-language pretraining [Elizalde et al., 2023, Wu et al., 2023, 2022] provides one bridge: shared audio–text embedding spaces that can be plugged into diffusion conditioning.

¹Code and data are available at <https://github.com/andreali31/AudioImageGen>.

In practice, however, an audio embedding alone may not capture the things that most determine an album cover’s look: the artist’s existing visual brand, narrative themes, color palettes, and pose conventions. We therefore ask:

Does adding lightweight textual artist context—style words, biographical detail, or both—improve audio-conditioned album-cover generation over an audio-only model and over a text-only baseline?

We approach the question with a controlled qualitative comparison rather than a large-scale benchmark. Our contributions are:

1. A simple architecture that combines CLAP audio embeddings with SD’s text-encoder context by sequence-axis concatenation in the UNet’s cross-attention.
2. Four trained conditions (AUDIO_ONLY, STYLE, BIOGRAPHY, COMBINED) plus a caption-only baseline, all reusing the same cached audio embeddings and VAE latents and trained with identical hyperparameters.
3. A held-out qualitative evaluation across three stylistically distinct artists, showing that audio conditioning alone recovers artist-specific portrait conventions, that textual context modulates palette and typographic salience rather than overriding the audio, and that the benefit of context is largest for the artist (LE SSERAFIM) whose covers are dominated by a wordmark rather than a single face.

2 Related Work

Latent diffusion and conditioning. Latent Diffusion Models [Rombach et al., 2022] (LDMs) cast image generation as iterative denoising of a low-resolution latent obtained from a VAE encoder. Stable Diffusion 1.5 is the open-weight LDM we build on. Cross-attention inside the UNet is the standard conditioning channel: any encoder that produces a sequence of $d_{\text{model}} = 768$ tokens can drive generation in place of the original CLIP text encoder.

Adapters and LoRA. Full fine-tuning of an 860M-parameter UNet is impractical at our data scale (60 paired examples). LoRA [Hu et al., 2022] inserts low-rank trainable matrices into attention projections, which lets us steer the model with $\sim 1\%$ of parameters and avoid catastrophic forgetting of the SD prior.

Audio–language representations. CLAP [Elizalde et al., 2023] and its LAION-trained extension [Wu et al., 2023] learn a shared audio–text embedding space, mirroring CLIP [Radford et al., 2021] for the audio modality. Wav2CLIP [Wu et al., 2022] similarly aligns audio to CLIP space. We use LAION-CLAP (laion/clap-htsat-unfused) and obtain a 512-d audio feature per clip.

Audio-conditioned image generation. Recent work explores generating images directly from audio: SonicDiffusion [Biner et al., 2024], Music2P [Choi et al., 2024], and earlier qualitative pipelines such as Wav2CLIP-driven VQGAN. These systems typically rely on either (i) audio-to-text captioning followed by a standard text-to-image model, (ii) shared audio–image embedding spaces, or (iii) direct conditioning of a diffusion model on audio embeddings. Approach (i) discards the raw acoustic signal entirely, since generation is driven by whatever the captioner verbalizes rather than the sound itself. Approach (ii) can produce visually coherent outputs but has no mechanism for grounding artist-specific visual identity unless that identity is already encoded in the shared embedding space, which for niche or newer artists it often is not. Approach (iii) preserves the acoustic signal but leaves open whether audio alone is sufficient to recover an artist’s visual brand conventions. Our work sits closest to (iii) and directly addresses that open question by introducing lightweight textual artist context as a complementary signal.

3 Method

3.1 Architecture

We freeze the SD 1.5 UNet and add (i) a LoRA adapter [Hu et al., 2022] on the cross-attention projections $\{W_Q, W_K, W_V, W_{\text{out}}\}$ at rank $r = 8$, and (ii) a small audio projector f_θ that maps a 512-d CLAP audio embedding to $K = 8$ tokens of dimension 768. The projector is a 3-layer MLP with LayerNorm:

$$f_\theta(\mathbf{a}) = \text{Reshape}_{(K, 768)}\left(W_2 \text{GELU}(W_1 \text{LN}(\mathbf{a}))\right), \quad \mathbf{a} \in \mathbb{R}^{512}.$$

It additionally carries a learned null embedding $\mathbf{n} \in \mathbb{R}^{K \times 768}$ used for classifier-free guidance.

Textual context is encoded by SD’s frozen CLIP text encoder to a $(77, 768)$ tensor $\mathbf{t}$. The full UNet conditioning is the sequence-axis concatenation

$$\mathbf{c} = [f_\theta(\mathbf{a}); \mathbf{t}] \in \mathbb{R}^{(K+77) \times 768},$$

which is sequence-length agnostic from the UNet’s perspective. Setting $\mathbf{t}$ to the empty-string embedding recovers an audio-only model.

3.2 Context modes

We study four mutually exclusive context conditions, all trained with the same architecture, optimizer, and seed (Table 1). Per-artist context strings are stored once in `artist_profiles.json` and formatted by `artist_context.py`.

Table 1: Context modes. All modes share the same audio pathway; only the text passed to SD’s CLIP text encoder changes.

Mode	Text input	Intuition
AUDIO_ONLY	(empty string)	Audio is the only signal.
STYLE	“polished contemporary pop . . .”	Visual/genre cues.
BIOGRAPHY	“American singer . . .”	Identity/narrative cues.
COMBINED	biography + style	Both.

3.3 Training objective and CFG

We use the standard noise-prediction MSE on SD’s pretrained DDPM scheduler:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{t, \mathbf{z}_0, \epsilon} \left[\|\epsilon - \epsilon_\phi(\mathbf{z}_t, t, \mathbf{c})\|_2^2 \right],$$

where $\mathbf{z}_0 = \text{VAE}(\mathbf{x})$ is the VAE latent of the ground-truth cover and ϕ collects the LoRA parameters. With probability $p_{\text{drop}} = 0.1$, we replace both modalities with their unconditional counterparts ($\mathbf{a} \rightarrow \mathbf{n}$, $\mathbf{t} \rightarrow \mathbf{t}_\emptyset$), so classifier-free guidance at inference time pushes *both* signals stronger jointly.

3.4 Caption-only baseline

The caption baseline (`caption_baseline.py`) loads pretrained SD 1.5 *without* the LoRA adapter or the audio projector and prompts it with the same artist text used in the COMBINED mode. It uses the same per-song seeds as the trained eval so rows in the comparison grid line up. This baseline answers the complement to our main question: *how much of the result comes from the audio?*

4 Experimental Setup

4.1 Dataset

We curated 60 tracks paired with their official album covers across three stylistically distinct artists (Ariana Grande, Drake, LE SSERAFIM; 20 tracks per artist). Covers are square JPG or PNG at

256 × 256 after a bicubic resize and center crop; audio is mono at 48 kHz. From each track we extract a single 10-second center window as the CLAP input, mirroring common practice in audio captioning. The split is fixed: 12 tracks are held out for evaluation (4 Ariana, 5 Drake, 3 LE SSERAFIM); the remaining 48 are used for training. Held-out song stems are listed in Appendix A.

4.2 Implementation

SD 1.5 is loaded at fp16 on a single Colab T4 GPU. CLAP audio features are computed once and cached alongside the VAE latents (prepare_data.py); training thus avoids re-running CLAP and the VAE every epoch. We train each of the four modes for E epochs with batch size 4 and learning rate 1×10^{-4} using AdamW, LoRA rank $r = 8$ with $\alpha = 16$, and the audio projector at $K = 8$ tokens. For inference we use the DPM-Solver multistep scheduler at 30 steps and CFG scale 5.0. We set $E = 50$ epochs for every mode, which at batch size 4 over 48 training pairs corresponds to roughly 600 optimizer steps per mode.

4.3 Evaluation protocol

For each held-out song we generate one image per trained mode using the same per-song seed and the same scheduler/CFG settings, then add the caption-only baseline image and the ground-truth cover to form a 6-column sheet. evaluate_heldout.py produces individual images and the per-song sheets; make_paper_figures.py composites the per-artist grids in Section 5. We do not report FID or CLIP-score: at $N = 48$ training images these metrics are dominated by sample variance, and the proposal explicitly targets subjective musical–visual alignment.

5 Results

We present qualitative grids per artist (Figures 1–3). Each row is one held-out song; columns from left to right are: ground-truth cover, caption-only baseline, AUDIO_ONLY, STYLE, BIOGRAPHY, COMBINED.

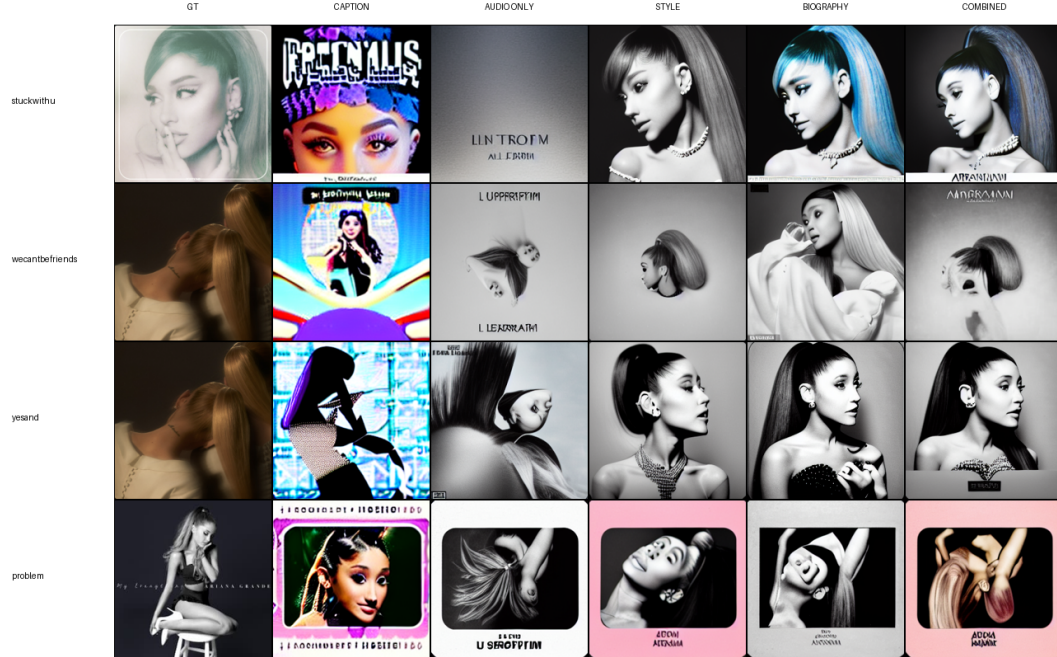


Figure 1: Held-out songs for Ariana Grande. Columns: ground truth, caption-only baseline, AUDIO_ONLY, STYLE, BIOGRAPHY, COMBINED. The caption baseline yields abstract graphics with no consistent subject, while all four trained columns produce the artist’s signature high-ponytail portrait—confirming the audio pathway, not the text, anchors identity. AUDIO_ONLY and STYLE skew monochrome; BIOGRAPHY and COMBINED reintroduce warmer tones and pastel backgrounds.

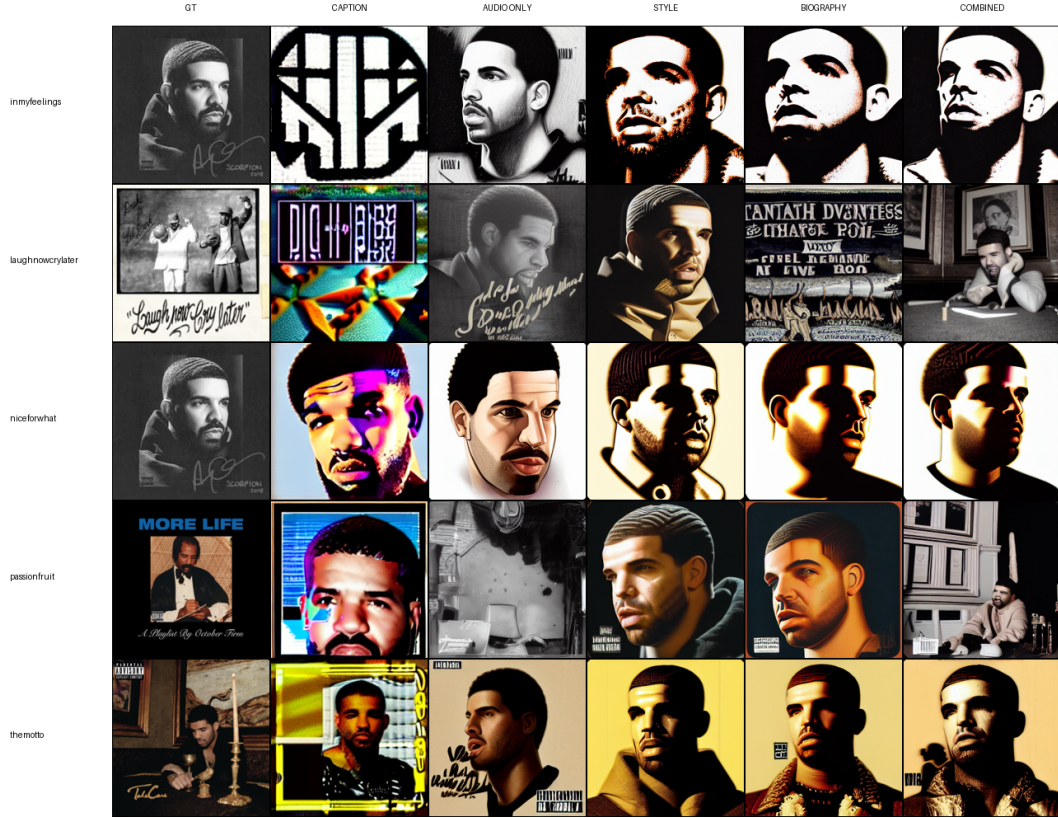


Figure 2: Held-out songs for Drake. The caption baseline produces logo-like graphics unrelated to the artist, while every trained column converges on a close-cropped male face at the artist’s characteristic upward angle. Per-song variation down each trained column indicates the audio embedding injects track-level signal. AUDIO_ONLY tends monochrome; STYLE and COMBINED shift to warmer sepia/gold; BIOGRAPHY adds dense but garbled typography.

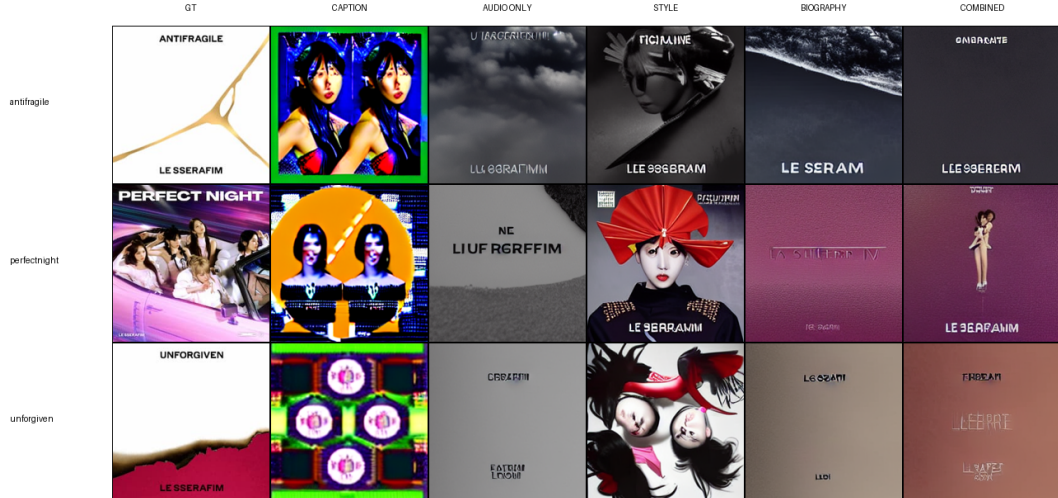


Figure 3: Held-out songs for LE SSERAFIM. Unlike the other two artists, real covers are wordmark-dominated, and the trained columns follow suit—most foreground large (garbled) “LE SSERAFIM”-like lettering. No condition reliably renders the five-member group, a limitation we attribute to data scarcity (Section 6). BIOGRAPHY gives the most saturated palettes (deep reds), while COMBINED returns to brighter, minimal layouts.

5.1 What audio adds: vs. the caption baseline

The starkest contrast in every figure is between the caption-only column and the four trained columns. Prompting pretrained SD with the artist text alone produces images in an “abstract album-art” mode: geometric motifs, logo-like marks, and saturated color fields that share a generic graphic-design flavor but bear no resemblance to the artists’ actual covers and contain no coherent human subject. This is unsurprising—the textual profiles describe genre and biography but never say “a photographic portrait of this specific person,” and pretrained SD has no grounding linking the artist’s name to their visual identity.

Once the LoRA and audio projector are trained, the picture changes qualitatively. All four trained columns recover the dominant visual template of each artist: a single male face for Drake, a high-ponytail female portrait for Ariana Grande, and a wordmark-led graphic for LE SSERAFIM. Because the only training signal that distinguishes one song from another within an artist is the audio embedding (the text is per-artist and therefore constant across that artist’s rows), the visible row-to-row variation *within* a trained column must be attributable to the audio. We read this as evidence that the audio channel is carrying track-level information into the generation, while the fine-tuning is what supplies artist identity that the caption baseline lacks.

5.2 What context adds: vs. audio-only

Comparing AUDIO_ONLY to the three text-augmented modes isolates the contribution of context. The effect is real but *modulatory* rather than structural: adding text rarely changes *what* is depicted (the portrait template is already set by the audio pathway) but consistently changes *how* it is rendered. Two axes recur across artists. First, palette: AUDIO_ONLY and STYLE tend toward high-contrast monochrome, whereas BIOGRAPHY and COMBINED reintroduce warmer, more saturated color—most visibly the sepia/gold tones for Drake and the deep reds for LE SSERAFIM. Second, typographic salience: the text-bearing modes, especially BIOGRAPHY, are far more likely to render lettering, since the biographical prompts are long and word-dense. For LE SSERAFIM, whose real covers are wordmark-centric, this is arguably an improvement; for Drake and Ariana, where covers are portrait-centric, the lettering appears as a garbled overlay artifact rather than a faithful design element.

The most common failure modes are consistent with a small-data SD 1.5 system: unreadable text (the base model cannot spell), occasional duplicated or distorted facial features, and saturation artifacts under strong guidance. None of these are specific to a single context mode.

5.3 Per-artist behaviour

The three artists stress the system differently. Drake and Ariana Grande both have a portrait-centric visual identity built around a single recognizable person, and for both the model captures that template robustly: even AUDIO_ONLY produces a plausible portrait, and context mainly tunes the palette. LE SSERAFIM is the hardest case. Their visual identity is a five-member group plus a prominent wordmark, and neither property survives cleanly: the wordmark appears (garbled) but the group never renders as five distinct people, even under BIOGRAPHY, whose prompt explicitly states “five-member ... girl group.” This suggests the bottleneck is not the *presence* of the relevant fact in the text but the model’s inability to *ground* a compositional, multi-subject concept from only sixteen training covers. In other words, context helps most where the target look is a global attribute (palette, typography) and least where it is a structured composition (a specific number of people).

6 Discussion

Audio is a per-song signal; text is a per-artist signal. By construction, the textual context in this study is per-artist, not per-song. Variation across rows (songs) within a column is therefore a clean attribution to the audio embedding. Variation across columns (modes) within a row is attribution to the text channel. The figures support this decomposition: within any single trained column we observe clear row-to-row differences in pose, framing, and tone (the audio effect), while moving across columns within a fixed row changes palette and typographic salience but largely preserves the subject (the text effect). The two channels thus occupy complementary roles rather than competing for the same degrees of freedom.

Complementarity vs. overriding. A natural failure mode is that strong artist-level text simply overrides the audio signal, collapsing all songs of an artist to one look. The COMBINED-vs.-AUDIO_ONLY comparison is the cleanest test of that. We do not observe collapse: COMBINED retains the per-song variation present in AUDIO_ONLY while layering on the palette and typographic shifts contributed by the text. The relationship is therefore better described as complementary—text re-weights surface attributes of an image whose structure is still driven by the audio—than as one channel dominating the other. The one partial exception is LE SSERAFIM, where the strong wordmark prior in the text occasionally crowds out the portrait entirely, the closest thing to overriding that we see.

Effect of dataset scale. With $N = 48$ training pairs and a frozen UNet, the model cannot learn new photographic concepts. It can re-weight an existing prior toward the training distribution. We therefore expect changes at the level of palette, composition density, and subject framing—not at the level of generating entirely new visual content. The figures match this prediction: the model reliably re-weights the SD prior toward portrait photography and artist-typical palettes, but it does not synthesize genuinely novel structured content such as a correctly-sized group or legible custom typography. Where the target identity is a re-weighting of existing concepts, the system succeeds; where it requires composing a new concept, it does not.

7 Limitations

- **Per-artist text, not per-song text.** Our text channel does not vary across songs of the same artist. A more granular variant would add per-track descriptions (mood, lyrical theme); this is left for future work because curating it for 60 songs by hand is non-trivial and letting an LLM caption the song risks leaking the answer through training-time captions.
- **No human evaluation.** The proposal anticipated ranking experiments. Within the project scope we report only authors’ qualitative observations.
- **Three artists, small dataset.** Generalisation claims to unseen artists are not supported. We trade scale for control: every mode sees the same images, audio, hyperparameters, and seeds, so the comparison across modes is internally valid.
- **Single audio window.** We extract a 10 s center clip per track; intra-song variation (e.g. bridge vs. chorus) is invisible to the model.
- **SD 1.5 native resolution.** SD 1.5 was pretrained at 512×512 ; generating at 256×256 keeps wall time and VRAM reasonable but slightly degrades absolute image quality. The relative comparison between modes is unaffected.

8 Future Work

The limitations above point to several concrete next steps, roughly ordered from cheapest to most involved.

More and more diverse data. The clearest bottleneck is scale. Our 48 training pairs span three artists in adjacent pop/hip-hop/K-pop idioms, which lets the model memorize three artist templates but not learn a general audio→visual mapping. Expanding to dozens of artists across genres with sharply different cover conventions (e.g. metal, jazz, classical, electronic, country) would test whether the audio channel can drive *style* rather than just retrieve a per-artist look, and would let us hold out entire artists at test time to measure genuine generalization rather than per-song variation within seen artists. More songs per artist would also reduce the overfitting that currently collapses LE SSERAFIM’s five-member compositions to one or two figures.

Longer and better-regularized training. We trained each mode for only 50 epochs (~ 600 steps) as a compute compromise. Longer schedules with a learning-rate warmup and cosine decay, EMA of the LoRA weights, and light data augmentation (horizontal flips, small crops) would likely sharpen the outputs and reduce the saturation and duplicated-feature artifacts we observed. A systematic sweep over LoRA rank, the number of audio tokens K , and the CFG dropout rate would clarify how much capacity the audio pathway actually needs.

Per-song audio-to-text captioning. Our text channel is per-artist and therefore constant across an artist’s songs. A natural extension is to caption each track—with a music tagging or audio-captioning model such as a CLAP-based zero-shot tagger or an audio-language model—to obtain per-song descriptors (tempo, mood, key instruments, lyrical theme) and feed those alongside the audio embedding. This directly targets the question the original proposal raised: does *automatically generated* per-song text outperform a static artist profile, and does it complement or merely duplicate the information already in the CLAP embedding? Care is needed to avoid leakage, since a caption derived from the same audio is not an independent signal.

Stronger backbones and higher resolution. Moving from SD 1.5 at 256×256 to SD 2.1 or SDXL at native resolution would raise absolute image quality and, importantly, improve text rendering, which matters for the wordmark-driven covers (LE SSERAFIM) where SD 1.5’s inability to spell was a visible failure mode. A model with a dedicated typography-aware pipeline, or an explicit ControlNet/IP-Adapter conditioning path, could let the system place a legible artist name without garbling.

Quantitative and human evaluation. Our study is deliberately qualitative, but a larger dataset would make automatic metrics meaningful again: CLAP audio $\leftrightarrow$ image-caption similarity, CLIP image $\leftrightarrow$ metadata similarity, and FID against the held-out covers, each computed per condition to quantify the modality contributions we currently describe by eye. The proposal’s planned human study—ranking which generated cover best matches a song’s mood and identity—would provide the subjective ground truth that no embedding-based metric fully captures, and is the most important missing piece for turning these observations into measured claims.

9 Conclusion

We presented a controlled qualitative study of the effect of textual artist context on audio-conditioned album-cover generation. By holding architecture, data, training, and seeds fixed across four context conditions and contrasting them with a no-training caption baseline, we isolate the contribution of each modality. Our headline finding is that, at this data scale, audio conditioning supplies what the caption baseline cannot—per-song variation and learned artist identity—while textual context acts as a complementary modulator of palette and typography rather than as the primary driver of the generated cover.

References

- Burak Can Biner, Farrin Marouf Sofian, Umur Berkay Karakaş, Duygu Ceylan, Erkut Erdem, and Aykut Erdem. SonicDiffusion: Audio-driven image generation and editing with pretrained diffusion models. *arXiv preprint arXiv:2405.00878*, 2024.
- Joong Ho Choi, Geonyeong Choi, Ji-Eun Han, Wonjin Yang, and Zhi-Qi Cheng. Music2P: A multi-modal ai-driven tool for simplifying album cover design. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM), Demo Track*, 2024.
- Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2CLIP: Learning robust audio representations from CLIP. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

A Held-out songs and seeds

Table 2: Held-out songs and the per-song seed used across all five generation columns.

Artist	Stem	Seed
Ariana Grande	stuckwithu	100
Ariana Grande	wecantbefriends	101
Ariana Grande	yesand	102
Ariana Grande	problem	103
Drake	inmyfeelings	104
Drake	laughnowcrylater	105
Drake	niceforwhat	106
Drake	passionfruit	107
Drake	themotto	108
LE SSERAFIM	antifragile	109
LE SSERAFIM	perfectnight	110
LE SSERAFIM	unforgiven	111

B Artist profile strings

The full text passed to SD’s CLIP text encoder in each mode is shown below; the source is `artist_profiles.json`.

Ariana Grande — style. *polished contemporary pop and R&B, expressive high-register vocals, romantic and introspective themes, glamorous editorial portraiture, soft luminous lighting, pastel pink, silver and monochrome palettes, elegant minimal album-cover composition.*

Ariana Grande — biography. *American singer and actress from Boca Raton, Florida; began on Broadway and Nickelodeon before releasing her 2013 debut album Yours Truly through Republic Records; known for a four-octave vocal range and a career spanning pop, R&B and electronic influences.*

Drake — style. *Toronto-rooted hip-hop and contemporary R&B, melodic rap and singing, confident and vulnerable moods, nocturnal city atmosphere, cinematic minimalism, dark blue, black and gold palettes, stark symbolic album-cover composition.*

Drake — biography. *Canadian rapper, singer and former actor from Toronto; first became known through Degrassi before independently releasing mixtapes and building an influential music career; co-founded the Toronto collective and record label OVO Sound.*

LE SSERAFIM — style. *high-energy K-pop, dance-pop and R&B influences, fearless self-confidence and resilience themes, fashion-forward group imagery, bold choreography, sleek editorial photography, high-contrast black, white and vivid accent colors, dynamic graphic album-cover composition.*

LE SSERAFIM — biography. *five-member multinational girl group formed by Source Music under HYBE, consisting of Sakura, Kim Chaewon, Huh Yunjin, Kazuha and Hong Eunhae; debuted in 2022; the name LE SSERAFIM is an anagram of IM FEARLESS and expresses determination to move forward without being shaken by public judgment.*

C Hyperparameter summary

Table 3: Training and inference hyperparameters, identical across modes.

Setting	Value
Base model	Stable Diffusion 1.5
Audio encoder	LAION-CLAP laion/clap-htsat-unfused
LoRA rank / α	8 / 16
LoRA target modules	to_q, to_k, to_v, to_out.0
Audio tokens K	8
Training epochs	50
Batch size	4
Optimizer	AdamW, lr 1×10^{-4} , wd 1×10^{-4}
CFG dropout p_{drop}	0.1
Mixed precision	fp16
Inference scheduler	DPM-Solver multistep
Inference steps	30
Guidance scale	5.0
Image size	256×256
Audio clip	10 s center, 48 kHz mono

D Reproducibility

All code, audio–cover pairs, profile strings, and the held-out split are in the project repository at <https://github.com/andreali31/AudioImageGen>. To reproduce: install `requirements.txt`, run `prepare_data.py` to build the cache, run `run_context_experiments.sh` to train the four modes, then `caption_baseline.py`, then `evaluate_heldout.py`, and finally `make_paper_figures.py`. The Colab notebook `run_experiments_colab.ipynb` stitches all six steps together.