Supplementary Material

Appendix

In this supplemental document, we provide additional technical details, expanded analyses, and extended results that complement the main paper:

- **A.1:** We present a detailed derivation of flow matching, including its ODE formulation, connections to diffusion models, training objectives, and theoretical properties relevant to deterministic generative flows.
- **A.2:** We detail the truncated inversion-free editing procedure, analyze the role of temporal parameters $t_{\min}$ and $t_{\max}$, and discuss how different regions of the flow trajectory contribute to structure–detail trade-offs in music style transfer.
- **A.3:** We detail hyperparameters, preprocessing, dataset statistics, and evaluation pipeline configurations for reproducibility.
- **A.4:** We report further empirical studies, including text-length sensitivity, genre-specific transfer difficulty, and visual-encoder ablations, providing deeper insights into the behavior of multimodal style conditioning.
- **A.5:** We describe the subjective evaluation process, including interface design, rating criteria, and instructions provided to participants.
- **A.6:** We showcase additional visualizations, spectrogram comparisons, and audio examples illustrating multimodal control, melody preservation, and stylistic diversity.

A.1 Flow Matching

Flow matching provides a deterministic alternative to diffusion-based generative modeling by directly learning a time-dependent velocity field that transports a simple base distribution to the data distribution. This section summarizes the formulation and clarifies its theoretical connection to diffusion models.

Diffusion models and the probability flow ODE define a forward noising process as an Itô SDE:

$$dx_t = f(x_t, t)\, dt + g(t)\, dw_t, \quad t \in [0, 1], \tag{10}$$

where $x_t \in \mathbb{R}^d$ denotes the latent at time t, $f(\cdot, t)$ is a drift term, $g(t)$ is a scalar (or diagonal) diffusion coefficient, and w_t is a standard Wiener process. This process gradually perturbs clean data $x_1 \sim p_1$ into a simple prior $x_0 \sim p_0$. The corresponding reverse-time SDE is

$$dx_t = \big(f(x_t, t) - g(t)^2 \nabla_x \log p_t(x_t)\big)\, dt + g(t)\, d\bar{w}_t, \tag{11}$$

where $\bar{w}_t$ is a reverse-time Wiener process, p_t is the marginal density and $\nabla_x \log p_t$ is the score function.

Song et al. [58] showed that this reverse SDE shares the same marginal densities as the deterministic probability flow ODE:

$$\frac{dx_t}{dt} = f(x_t, t) - \tfrac{1}{2} g(t)^2 \nabla_x \log p_t(x_t), \tag{12}$$

where the right-hand side can be interpreted as a velocity field that transports p_0 to p_1. Thus, diffusion sampling can be viewed as integrating an ODE governed by an unknown velocity field involving the score.

Deterministic flows and flow matching [32, 35] avoid score estimation by directly learning a velocity field $v_\theta(x, t)$ such that the ODE:

$$\frac{dz_t}{dt} = v_\theta(z_t, t), \qquad z_0 \sim p_0, \tag{13}$$

where $z_t \in \mathbb{R}^d$ is the flow state at time t, and p_0 is a simple base distribution (e.g., Gaussian). The goal is push forward the distribution p_0 to the data distribution p_1. Instead of estimating $\nabla_x \log p_t$, the key idea is to synthesize reference trajectories for which the true velocity can be computed analytically. Given a coupling $\pi(x_0, x_1)$ between p_0 and p_1, define an interpolation path $z_t = \psi_t(x_0, x_1)$, for example the linear interpolation:

$$\psi_t(x_0, x_1) = (1 - t)x_0 + t x_1. \tag{14}$$

Along this path, the target (supervisory) velocity is:

$$v^\star(z_t, t) = \frac{d}{dt}\psi_t(x_0, x_1) = x_1 - x_0. \tag{15}$$

The flow matching loss trains v_θ to match this velocity:

$$\mathcal{L} = \mathbb{E}_{x_0, x_1, t}\Big[\big\| v_\theta\big(\psi_t(x_0, x_1), t\big) - v^\star\big(\psi_t(x_0, x_1), t\big)\big\|_2^2\Big], \tag{16}$$

where $x_0, x_1 \sim \pi$ and $t \sim \mathcal{U}(0, 1)$. Once trained, sampling reduces to integrating the deterministic ODE above, yielding a noise-free and typically more stable generation process. Rectified flows [35] extend this idea by choosing a nonlinear interpolation:

$$\psi_t^{(\alpha)}(x_0, x_1) = (1 - t^\alpha)x_0 + t^\alpha x_1, \quad \alpha > 1, \tag{17}$$

whose derivative $v_\alpha^\star(z_t, t) = \alpha t^{\alpha-1}(x_1 - x_0)$ "straightens" trajectories and improves sample quality at few integration steps. Conceptually, flow matching can be viewed as directly learning the velocity of a probability flow ODE, while diffusion learns scores that implicitly define that ODE. This links the two paradigms: diffusion models solve generative modeling through stochastic processes and score estimation, whereas flow matching solves it through supervised velocity learning and deterministic transport.

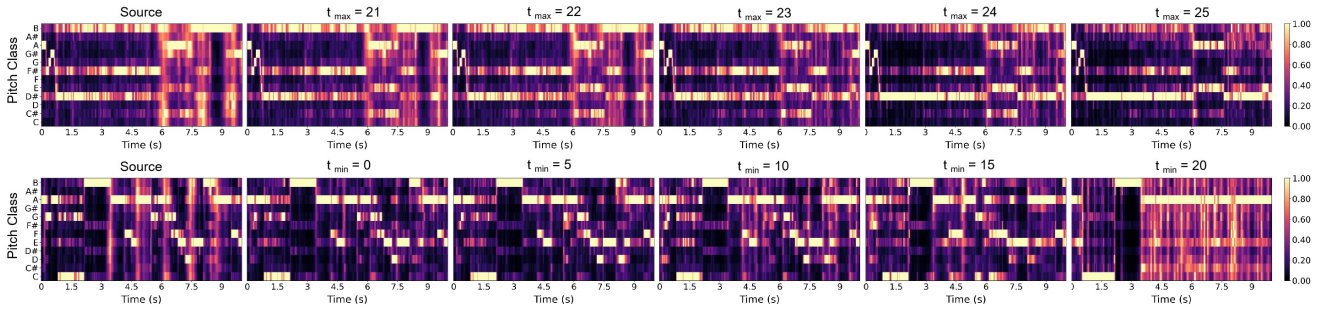

Figure 5. Chroma visualizations under different $t_{\max}$ and $t_{\min}$ configurations. Varying $t_{\max}$ produces minimal change, while changes in $t_{\min}$ strongly affect harmonic structure and melodic preservation.

A.2 Truncated Inversion-Free Editing

A distinguishing aspect of FlowEdit [26] is its use of temporal truncation to control where semantic editing occurs along the rectified flow trajectory. Although this mechanism is not the conceptual core of rectified flow, it plays a practical and influential role in how editing strength is distributed across different noise levels. In this appendix, we explore the behavior of the two associated hyperparameters – $t_{\max}$ and $t_{\min}$ – and analyze how they shape the outcome of inversion-free music style transfer.

This design arises from the well-established observation in flow and diffusion models that different time regions encode different semantic granularities: high-noise (early) states capture global structure, while low-noise (late) states encode fine-grained details such as texture, timbre, and local harmonic variations. Prior analyses of rectified flows [32, 35] establish that the flow trajectory is not semantically uniform. FlowEdit operationalizes this property by allowing users to truncate editing to a subinterval $[t_{\min}, t_{\max}]$, thereby deciding how much global structure and how much fine-scale detail should be rewritten.

To clarify how this mechanism functions, we provide a more detailed description of FlowEdit's editing dynamics. During the semantic editing phase ($t \in [t_{\min}, t_{\max}]$), the noisy source state is reconstructed by:

$$Z^{\text{src}}_{t_i} = (1 - t_i)X^{\text{src}} + t_i N_{t_i}, \tag{18}$$

where X^{src} is the clean source latent, N_{t_i} is the noise sample associated with time t_i, and $Z^{\text{src}}_{t_i}$ denotes the reconstructed source trajectory at step t_i. A target-aligned perturbation is then constructed by:

$$Z^{\text{tar}}_{t_i} = Z^{\text{FE}}_{t_i} + Z^{\text{src}}_{t_i} - X^{\text{src}}, \tag{19}$$

where $Z^{\text{FE}}_{t_i}$ is the current edited latent at time t_i and $Z^{\text{tar}}_{t_i}$ inherits target semantics (through $Z^{\text{FE}}_{t_i}$) while maintaining the structural offset between the reconstructed and clean source. This construction injects target semantics while preserving the structural offset from the reconstructed source. Semantic deviation between target and source flows is captured by the velocity difference:

$$V^{\Delta}_{t_i} = V^{\text{tar}}(Z^{\text{tar}}_{t_i}, t_i) \; - \; V^{\text{src}}(Z^{\text{src}}_{t_i}, t_i), \tag{20}$$

where V^{tar} and V^{src} denote the velocity fields (i.e., flow derivatives) for the target and source conditions, respectively. And $V^{\Delta}_{t_i}$ measures the semantic shift required to steer the source trajectory toward the target style. This equation can be integrated into the edited state via:

$$Z^{\text{FE}}_{t_{i-1}} = Z^{\text{FE}}_{t_i} + (t_{i-1} - t_i)\, V^{\Delta}_{t_i}. \tag{21}$$

This step is the only part of the trajectory where the model is allowed to modify semantic content; therefore the interval $[t_{\min}, t_{\max}]$ fully determines the region where editing occurs.

Once the trajectory reaches $t_{\min}$, FlowEdit enters a second deterministic refinement pass. Here, no further semantic differences are applied; instead, the edited state is transported purely under the target flow so that it remains consistent with the target distribution while avoiding unnecessary perturbation of low-level details. This refinement is expressed by:

$$Z^{\text{tar}}_{t_{i-1}} = Z^{\text{tar}}_{t_i} + (t_{i-1} - t_i)\, V^{\text{tar}}(Z^{\text{tar}}_{t_i}, t_i), \tag{22}$$

for all $t_i \leq t_{\min}$. Unlike the editing phase, this step preserves the fine-scale structure already established, effectively "locking in" the edited content as it approaches $t = 0$.

When adapting this mechanism to audio-based style transfer, we observe clear domain-specific behavior. As shown in Fig. 5, varying $t_{\max}$ has little audible effect, consistent with the fact that early-time flow states encode coarse structural cues that typically do not require modification. In contrast, changes to $t_{\min}$ produce pronounced differences in melodic and stylistic outcomes. Since states near $t = 0$ contain fine-grained harmonic and temporal details, deeper editing in this region directly influences musically salient structure. The chroma visualizations confirm this: smaller $t_{\min}$ values increase editing depth but risk

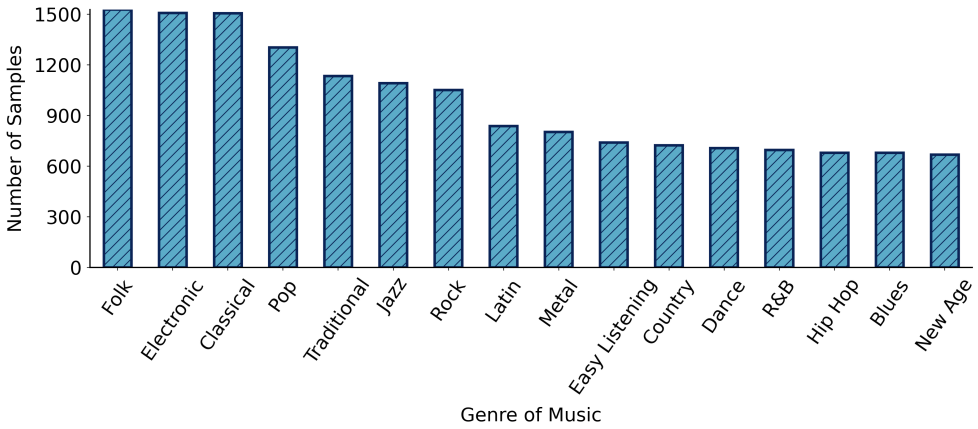

Figure 6. Distribution of the 16 annotated musical genres in our multimodal dataset (about $15k$ samples). The corpus is relatively balanced, with Folk, Electronic, and Classical containing slightly more samples and the remaining categories distributed evenly.

melody drift, whereas larger values better preserve pitch-class identity.

These findings support the broader theoretical consensus that temporal coordinates in rectified flows exhibit semantic non-uniformity: early steps correspond to global structure, and late steps encode local detail. For music, where fine-scale harmonic and melodic structure is crucial, $t_{\min}$ emerges as the dominant control parameter. Our truncated inversion-free editing therefore places primary emphasis on tuning $t_{\min}$, which provides a principled mechanism for balancing musical fidelity against stylistic transformation. In practice, we perform hyperparameter search on the validation set and select $t_{\max} = 23$ and $t_{\min} = 15$ as the optimal configuration.

A.3 Experiments Setup

Train Configuration. For our method, we adopt Make An Audio 3 (MAA3) [70] as the backbone given its strong flow-based performance and efficient inference. The backbone is frozen, and only the cross-adapter is trained. Training is performed for 20 epochs with a learning rate of 1×10^{-5} and a batch size of 16. The melody-guided rate η is set to 1×10^{-4} with a step size of 2 and a weight λ_{chr} of 1. The dataset is split into training, validation, and test sets with an 8:1:1 ratio, with identical data splits and prompt settings for fair comparison. All models are implemented in PyTorch and trained on L40S GPUs.

Dataset. We construct a unified multimodal corpus by reorganizing and extending samples from MeLBench and MusicCaps into coherent triplets $\langle I, T, M \rangle$, where each triplet contains a visually descriptive image, a text prompt, and an audio clip. Because the original music captions from these datasets are often noisy, stylistically inconsistent, or overly literal, we apply an LLM-assisted refinement procedure to obtain clean genre annotations and structured semantic descriptions in two stages.

In the first stage, GPT-4 is prompted to classify each sample into one of sixteen predefined musical genres. Given a raw caption such as *"This song is a pop tune with a medium tempo, featuring an electric guitar,"* the model selects the most suitable genre from a fixed list, ensuring consistency by repeating the classification twice independently and accepting only samples with matching predictions. This procedure yields a reasonably balanced genre distribution: folk, electronic, and classical contain moderately more samples, followed by pop, while the remaining genres have roughly comparable counts (visualized in Fig. 6).

In the second stage, we convert every free-form caption into a structured representation that minimizes linguistic noise and improves conditioning stability. For instance, the caption above is rewritten as *"{clean e-guitar playing & all}, {live performance ambiance & all}"*, which disentangles instrumentation, ambience, and stylistic factors into a compact semantic form. This structured normalization helps the model align text, image, and music more reliably during style transfer. Finally, to focus on melodic and harmonic content, we apply Demucs to remove vocals from all audio tracks, keeping purely instrumental stems across the corpus. The resulting dataset provides multimodally aligned, genre-aware, and stylistically coherent samples suitable for fine-grained music style transfer.

Evaluation Metrics. To comprehensively assess the effectiveness of our multimodal music style transfer framework, we evaluate both objective and subjective aspects of the generated audio. Objective metrics quantify structural fidelity, stylistic alignment, and multimodal consistency, while subjective ratings capture human perception of audio quality, prompt relevance, and melodic similarity. Together, these metrics provide a holistic view of how well a model

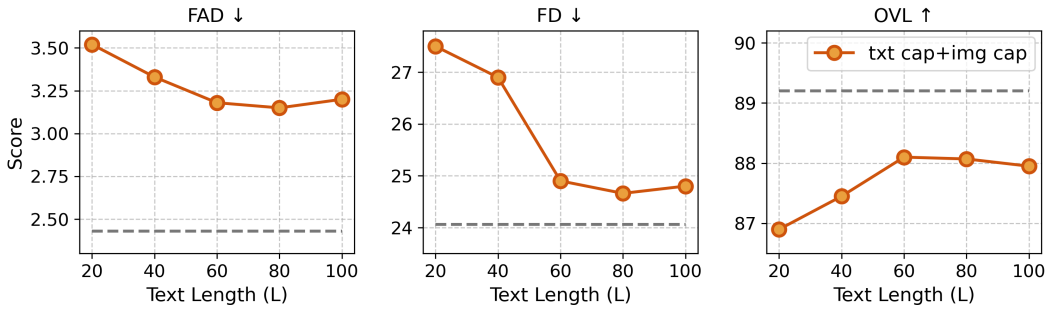

Figure 7. Effect of text length on multimodal conditioning. The orange curve (*"txt cap + img cap"*) uses the full textual input, meaning the model is conditioned on both the music-description caption and the image-description caption, while the gray dashed line corresponds to conditioning on the music prompt together with the image.

preserves musical identity while adapting to the target style.

1) Objective Evaluation. Our objective evaluation focuses on two key dimensions of music style transfer: (1) structural preservation – melody, harmony, and tonal integrity with respect to the source, and (2) stylistic consistency – alignment with the target domain and multimodal prompts. To this end, we employ five complementary metrics: FAD and FD for distributional style similarity, IMSM for cross-modal coherence, and F_0-PCC and Chroma-CQT for melody and harmonic preservation. Below, we briefly summarize each metric and its relevance to MST performance.

- **Fréchet Audio Distance (FAD).** FAD [24] extends the Fréchet Inception Distance (FID) [17] from images to audio by comparing the mean and covariance of embeddings extracted from a pretrained VGGish network [16] between generated and reference (target-style) audio sets. It quantifies overall perceptual distributional distance and correlates well with human judgments of quality.
- **Fréchet Distance (FD).** FD adopts the same Gaussian-assumption formulation as FAD but employs PANNs [25] as the embedding extractor, offering a complementary measure of distributional alignment between generated and target-style music.
- **Image–Music Similarity Metric (IMSM).** IMSM [4] measures multimodal consistency between generated music and its textual–visual prompts by combining CLIP-based image–text and CLAP-based audio–text similarities. The similarity matrix $\mathcal{A}_{\text{IMSM}} = A_{\text{CLIP}} A_{\text{CLAP}}^{\top}$ captures joint semantic alignment across modalities.
- **Fundamental Frequency Correlation (F_0-PCC).** To evaluate melody preservation, we compute the Pearson correlation coefficient [3] between the F_0 contours of source and transferred audio. This metric emphasizes pitch-contour consistency while being robust to timbral variation [28, 45].
- **Chroma-CQT Similarity.** Chroma features summarize energy across the 12 pitch classes (C–B) independent of octave, capturing harmonic and tonal information while being largely invariant to timbre. We extract chroma features using the Constant-Q Transform (CQT) [55] and compute the cosine similarity between source and transferred chromagrams. Higher Chroma-CQT values indicate better preservation of harmonic and tonal structure.

2) Subjective Evaluation. We employ three subjective metrics to assess perceptual quality: Overall Audio Quality (OVL), Relevance to Prompt (REL), and Melodic Consistency (MOS_{con}). All ratings are collected on a 0–100 scale, with 15 participants evaluating randomized samples to minimize bias.

- **Overall Audio Quality (OVL).** OVL measures the perceived realism and clarity of the generated music. Participants are asked to rate how natural, coherent, and artifact-free each audio clip sounds, following the MOS protocol.
- **Relevance to Prompt (REL).** REL quantifies how well the generated audio aligns semantically and stylistically with the conditioning inputs (image and/or text). Listeners examine the provided prompt and evaluate thematic, emotional, and stylistic correspondence.
- **Melodic Consistency (MOS_{con}).** MOS_{con} assesses the preservation of melodic structure between source and transferred audio. Raters listen to paired samples and score the perceptual similarity of their melodic contours, ignoring timbral or instrumentation differences.

A.4 Additional Experiments

Effect of Text Length. To examine how textual length influences multimodal conditioning, we analyze caption length under two settings: (1) a text-only configuration, where the model receives a concatenation of the music description and the image caption, and (2) a combined text–image configuration, where the raw image serves as an additional conditioning modality. As shown in Fig. 7, increasing caption length in the text-only setting generally improves performance, since richer descriptions provide stronger semantic grounding. However, excessively long

Table 5. Comparison of different encoder configurations.

Encoder(s)	FAD↓	FD↓	IMSM↑	OVL↑	REL↑	MOS_{con}↑
CLIP + ViT	**2.43**	**24.06**	**0.828**	**89.27**	**88.13**	**89.20**
CLIP	2.68	25.33	0.807	87.93	86.47	88.27
ViT	3.12	26.47	0.792	86.67	85.73	87.07

Table 6. Variability across genre-to-genre transfers. Embedding similarity indicates stylistic similarity; lower values suggest closer styles and easier transfer.

Source → Target	Similarity	FAD↓	FD↓	IMSM↑	F_0-PCC↑	CCS↑	OVL↑	REL↑	MOS_{con}↑
Blues → Metal	0.57	2.87	25.18	0.794	0.425	0.880	87.60	87.27	87.67
Blues → Jazz	0.69	2.38	23.25	0.836	0.408	0.872	89.53	88.20	89.40
Blues → Classical	0.62	2.61	24.10	0.812	0.417	0.876	88.40	87.73	88.53
Rock → Metal	0.71	2.20	22.85	0.848	0.402	0.870	90.13	89.13	90.33
Rock → Jazz	0.59	2.78	24.95	0.803	0.430	0.884	87.07	86.47	87.80
Rock → Classical	0.60	2.73	24.40	0.815	0.423	0.879	87.67	86.93	88.07

captions introduce redundant or noisy information, leading to a slight degradation in performance. In contrast, the combined text–image configuration, evaluated using the optimal caption length selected via hyperparameter search, achieves consistently superior performance compared to all text-only variants, indicating that the visual modality supplies concise and unambiguous stylistic cues, and the model relies less on highly verbose text.

Effect of Visual Encoders. To examine the role of different visual encoders in our multimodal conditioning pipeline, we evaluate three configurations, as shown in Tab. 5, the system that uses both CLIP and ViT, the system that uses only CLIP, and a system that uses only ViT. The configuration employing both encoders achieves the strongest overall performance. We conjecture that this advantage arises from the complementary nature of the two architectures. CLIP tends to capture higher-level semantic and stylistic attributes, whereas ViT is more attuned to global spatial composition and visual structure. When combined, these distinct forms of information appear to provide richer visual conditioning signals, which in turn lead to more stable style transfer.

Variability Across Genre-to-Genre Transfers. We evaluate multiple source–target genre pairs to investigate how stylistic similarity influences transfer quality. As summarized in Tab. 6, genre pairs with higher similarity generally achieve better objective and subjective scores, while pairs with lower similarity exhibit moderate declines, particularly in target-oriented metrics such as FAD, FD, IMSM, OVL, and MOS_{con}. Despite these variations, the overall differences remain modest, indicating that the model maintains stable performance across diverse genre transitions.

A.5 User Study

To assess perceptual quality and multimodal alignment, we conducted a controlled subjective evaluation with fifteen volunteers from our institution, each compensated $10 USD. Participants rated three perceptual dimensions of the generated audio – overall quality, stylistic relevance, and melodic consistency. Each participant completed two evaluation sessions corresponding to our two conditioning modes: a text-only session, where only the caption was provided, and a text+image session, where both the visual reference and caption were shown. This setup enables a direct measurement of how visual cues influence human judgments of stylistic alignment. Each participant evaluated 20 samples per condition (40 total), presented in randomized order and freely replayable, with an average study duration of about 18 minutes.

The evaluation interface (Fig. 8) was implemented as a web-based tool. For each sample, the interface displayed the conditioning prompt(s) alongside two audio players: one for the source clip and one for the generated output. Participants assigned 0–10 scores for Overall Audio Quality (OVL), Style Relevance (REL), and Melodic Consistency using slider controls, and selected their participant ID prior to submission. All responses were automatically logged with timestamps and associated metadata.

A.6 Additional Cases

We present additional qualitative examples in Figs. 9–15 to illustrate how our model responds to different text–image prompts. Each case shows the input conditions and the resulting Mel-spectrograms, enabling visual inspection of

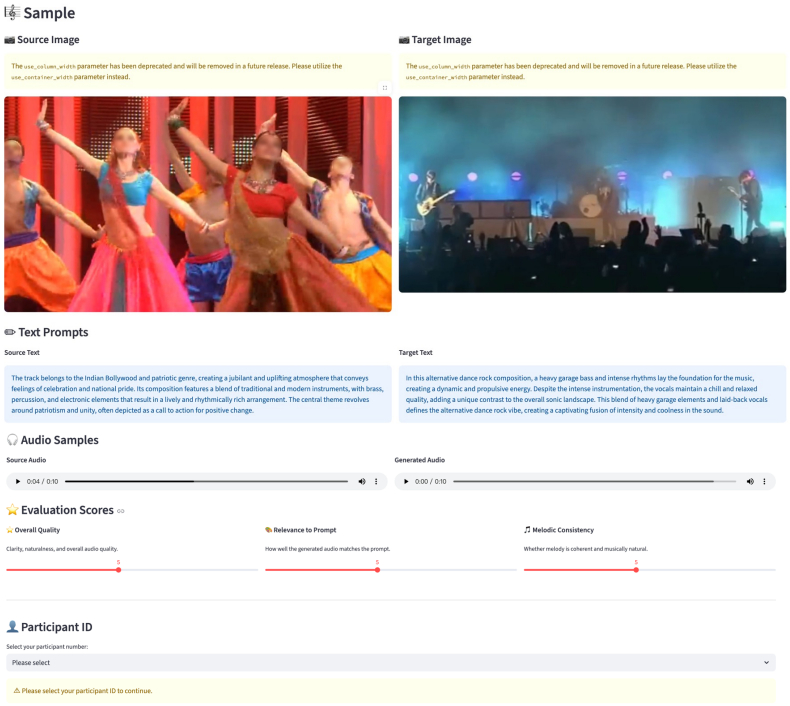

Figure 8. Participants are presented with the conditioning prompts (image, text, or both), along with the source audio and the generated audio. They rate each sample along three dimensions – Overall Audio Quality (OVL), Style Relevance (REL), and Melodic Consistency – using 0–10 sliders. A participant ID dropdown ensures proper tracking before submission.

stylistic changes and the preservation of melodic structure. These examples complement the main results and offer a clearer view of the model's multimodal editing behavior. For completeness, we also include the corresponding audio samples in the supplementary materials.

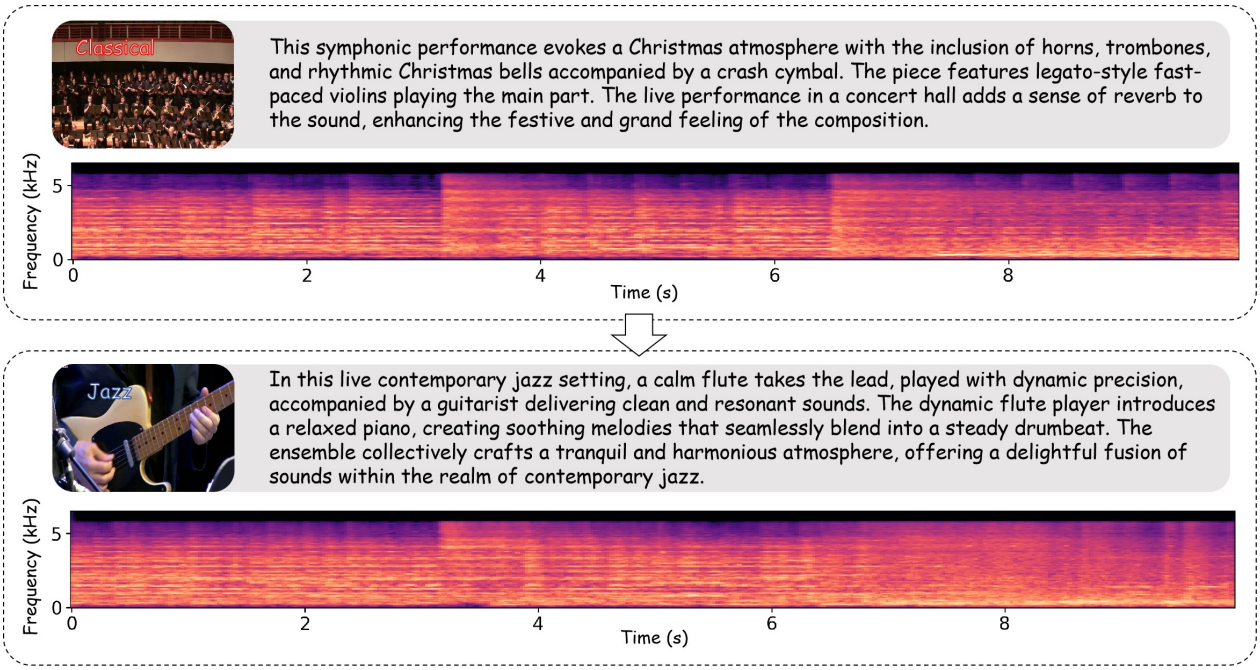

Figure 9. The Classical to Jazz style transfer. The figure shows the visual–text prompt with the source classical mel-spectrogram, followed by the same prompt paired with the generated jazz mel-spectrogram.

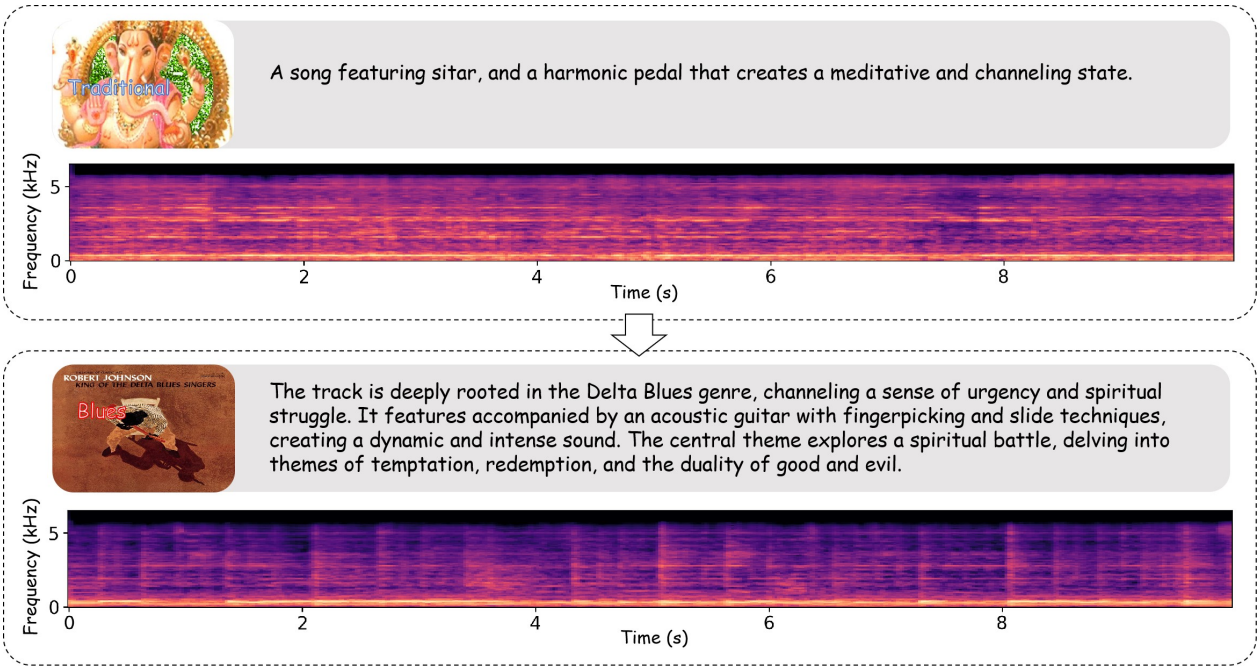

Figure 10. The Traditional to Blues style transfer.

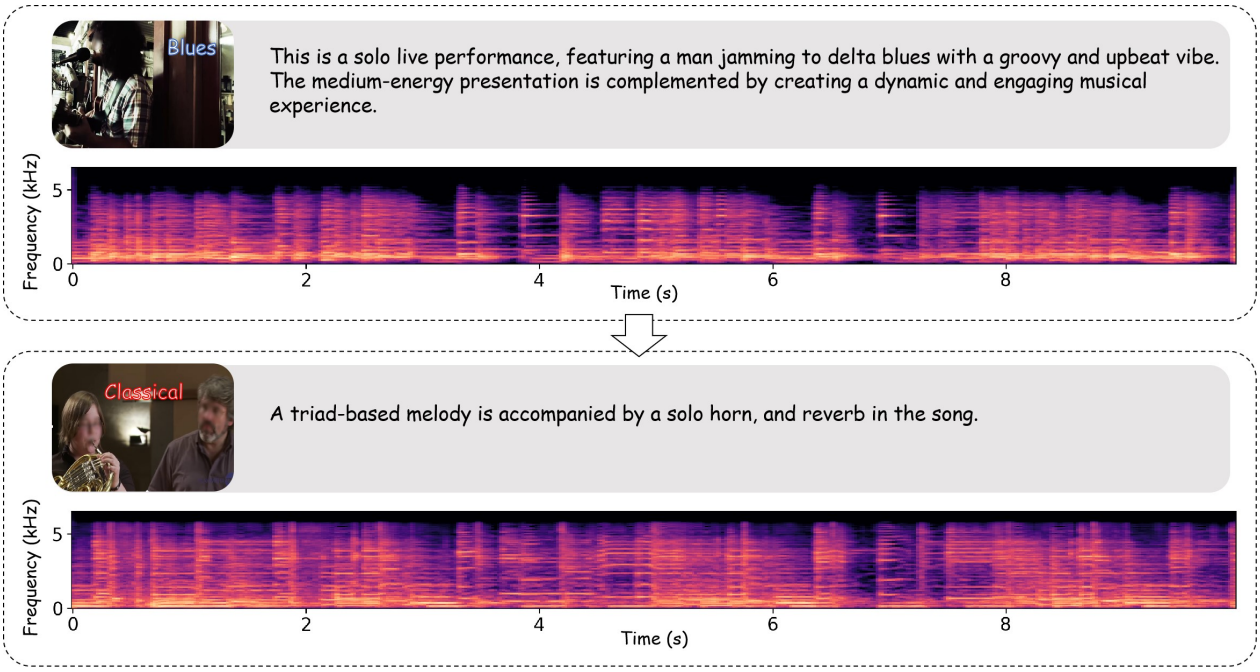

Figure 11. The Blues to Classical style transfer.

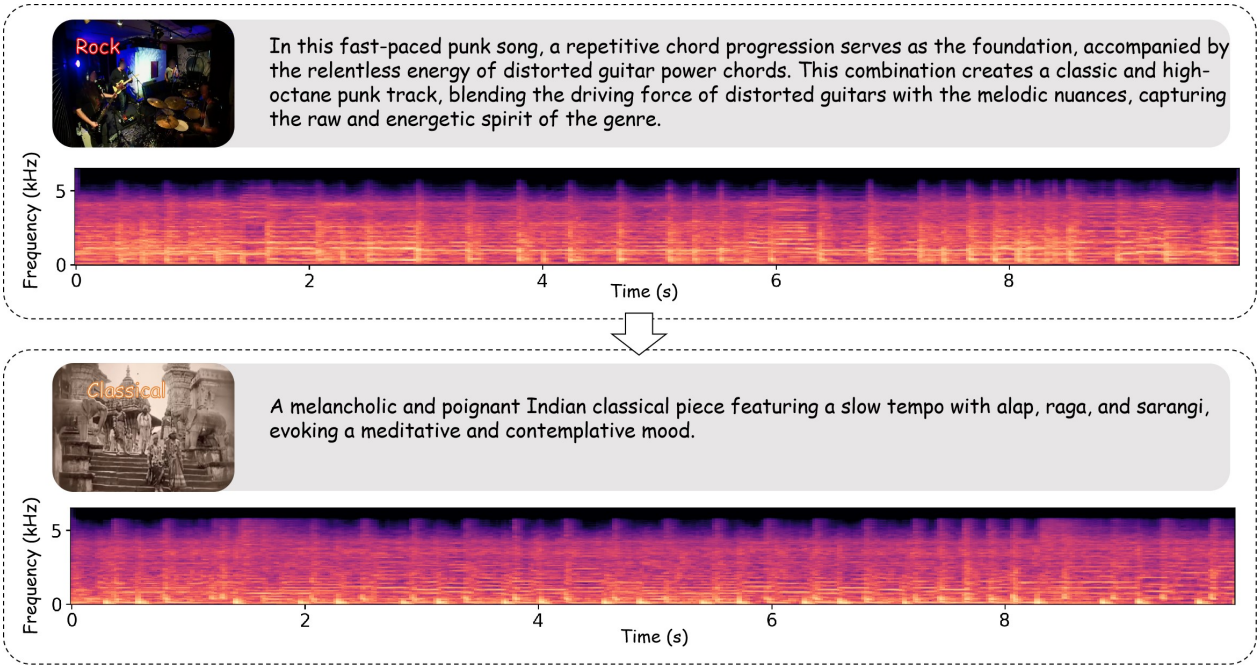

Figure 12. The Rock to Classical style transfer.

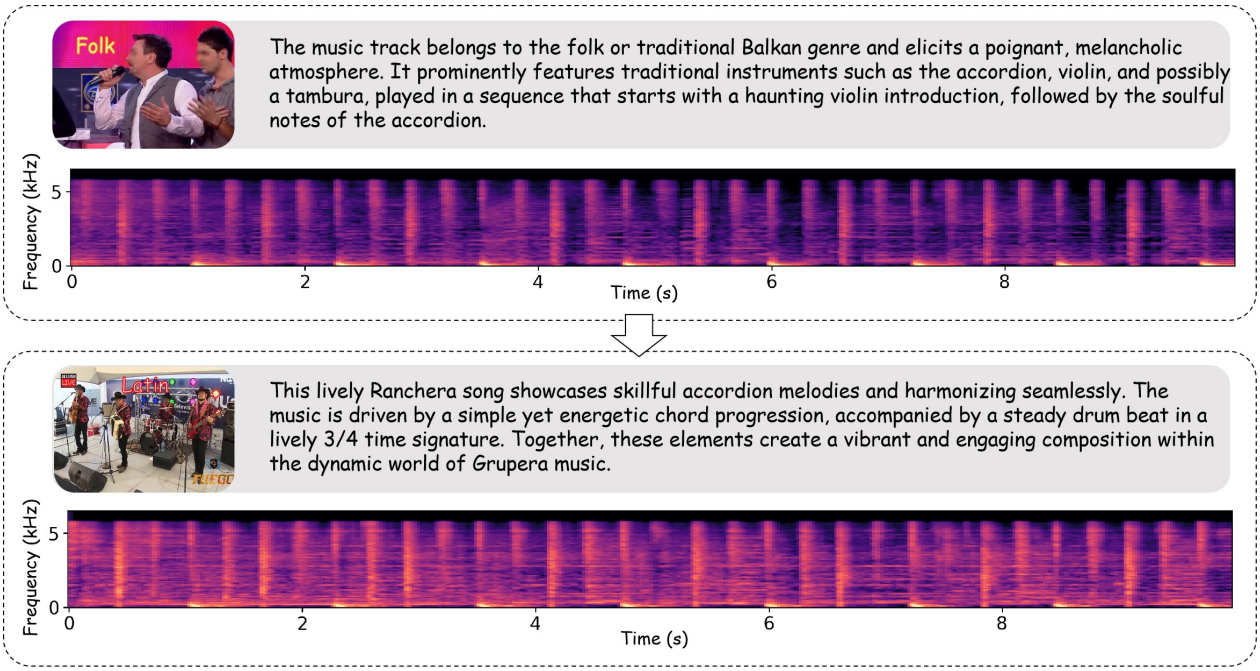

Figure 13. The Folk to Latin style transfer.

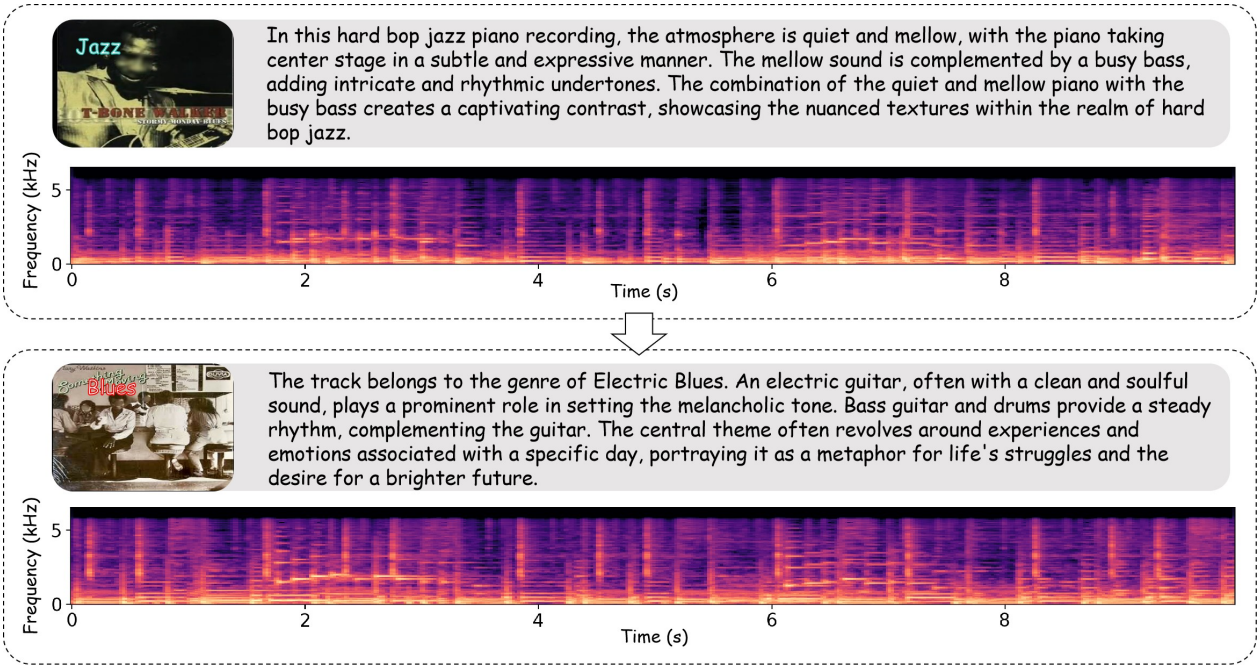

Figure 14. The Jazz to Blues style transfer.

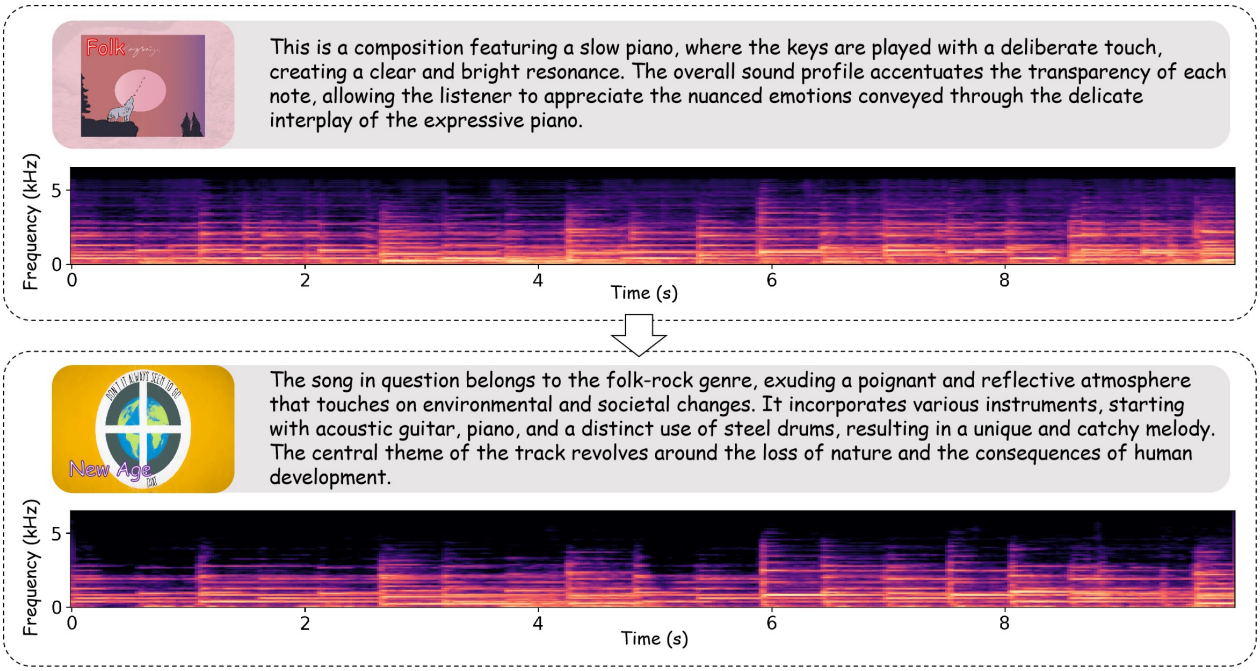

Figure 15. The Folk to New Age style transfer.

References

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. 3, 6, 7

[2] Mark A Bartsch and Gregory H Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, 2005. 5

[3] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009. 12

[4] Sanjoy Chowdhury, Sayan Nag, KJ Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26826–26835, 2024. 2, 6, 7, 12

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 4

[6] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720, 2023. 3, 6, 7

[7] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*, 2019. 6

[8] Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Junhao Huang, Conghui He, Dahua Lin, and Jiaqi Wang. Songcomposer: A large language model for lyric and melody generation in song composition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7108–7127, 2025. 7

[9] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1, 2

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[11] S. Forsgren and H. Martiros. Riffusion - stable diffusion for real-time music generation, 2022. 7

[12] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3590–3598, 2023. 6

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[14] Eric Grinstein, Ngoc QK Duong, Alexey Ozerov, and Patrick Pérez. Audio style transfer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 586–590. IEEE, 2018. 3

[15] Bing Han, Junyu Dai, Weituo Hao, Xinyan He, Dong Guo, Jitong Chen, Yuxuan Wang, Yanmin Qian, and Xuchen Song. Instructme: an instruction guided music edit framework with latent diffusion models. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 5835–5843, 2024. 3

[16] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017. 12

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 12

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[19] Chih-Fang Huang, Yu-Shian Lian, Wei-Po Nien, and Wei-Hua Chieng. Analyzing the perception of chinese melodic imagery and its application to automated composition. *Multimedia Tools and Applications*, 75(13):7631–7654, 2016. 2

[20] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*, 2022. 3

[21] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023. 2, 3, 7

[22] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024. 3

[23] Junyan Jiang, Gus G Xia, Dave B Carlton, Chris N Anderson, and Ryan H Miyakawa. Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 516–520. IEEE, 2020. 3

[24] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proc. Interspeech*, pages 2350–2354, 2019. 6, 12

[25] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recogni-

tion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 6, 12

[26] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19721–19730, 2025. 2, 3, 7, 8, 10

[27] Max WY Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, et al. Efficient neural music generation. *Advances in Neural Information Processing Systems*, 36:17450–17463, 2023. 7

[28] Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 12

[29] Ruibin Li, Ruihuang Li, Song Guo, and Lei Zhang. Source prompt disentangled inversion for boosting image editability with diffusion models. In *European Conference on Computer Vision*, pages 404–421. Springer, 2024. 3

[30] Sizhe Li, Yiming Qin, Minghang Zheng, Xin Jin, and Yang Liu. Diff-bgm: A diffusion model for video background music generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27348–27357, 2024. 2

[31] Sifei Li, Yuxin Zhang, Fan Tang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Music style transfer with time-varying inversion of diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 547–555, 2024. 1, 2, 3, 6, 7

[32] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. 3, 9, 10

[33] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning*, pages 21450–21474. PMLR, 2023. 1, 2, 6

[34] Huadai Liu, Jialei Wang, Xiangtai Li, Rongjie Huang, Yang Liu, Jiayang Xu, and Zhou Zhao. Medic: Zero-shot music editing with disentangled inversion control. *arXiv preprint arXiv:2407.13220*, 2024. 3

[35] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. 3, 9, 10

[36] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 564–572, 2024. 1, 2, 6, 7

[37] Michele Mancusi, Yurii Halychanskyi, Kin Wai Cheuk, Eloi Moliner, Chieh-Hsin Lai, Stefan Uhlich, Junghyun Koo, Marco A Martínez-Ramírez, Wei-Hsiang Liao, Giorgio Fabbro, et al. Latent diffusion bridges for unsupervised musical audio timbre transfer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 1

[38] Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using ddpm inversion. In *International Conference on Machine Learning*, pages 34603–34629, 2024. 2, 3, 6, 7

[39] Giorgio Mariani, Irene Tallini, Emilian Postolache, Michele Mancusi, Luca Cosmo, and Emanuele Rodolà. Multi-source diffusion models for simultaneous music generation and separation. In *International Conference on Learning Representations*, 2024. 2

[40] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 3

[41] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3

[42] Mubert-Inc. Mubert ai music generator — royalty free music, 2023. 7

[43] Aashiq Muhamed, Liang Li, Xingjian Shi, Suri Yaddanapudi, Wayne Chi, Dylan Jackson, Rahul Suresh, Zachary C Lipton, and Alex J Smola. Symbolic music generation with transformer-gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 408–417, 2021. 3

[44] Ziqian Ning, Huakang Chen, Yuepeng Jiang, Chunbo Hao, Guobin Ma, Shuai Wang, Jixun Yao, and Lei Xie. Diffrhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. *arXiv preprint arXiv:2503.01183*, 2025. 7

[45] Xinlei Niu, Jing Zhang, and Charles Patrick Martin. Hybridvc: Efficient voice style conversion with text and audio prompts. In *Proc. Interspeech*, pages 4368–4372, 2024. 6, 12

[46] Stephen E Palmer, Karen B Schloss, Zoe Xu, and Lilia R Prado-León. Music–color associations are mediated by emotion. *Proceedings of the National Academy of Sciences*, 110 (22):8836–8841, 2013. 2

[47] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[48] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021. 3

[49] 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*, pages 8748–8763. PMLR, 2021. 4

[50] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 4

[51] Adhika Sigit Ramanto and Nur Ulfa Maulidevi. Markov chain based procedural music generator with user chosen mood compatibility. *International Journal of Asia Digital Art and Design*, 21(1):19–24, 2017. 2

[52] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *International Conference on Machine Learning*, pages 4364–4373. PMLR, 2018. 2

[53] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *International Conference on Learning Representations*, 2025. 7, 8

[54] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8050–8068, 2024. 7

[55] Christian Schörkhuber and Anssi Klapuri. Constant-q transform toolbox for music processing. In *7th sound and music computing conference, Barcelona, Spain*, pages 3–64. SMC, 2010. 6, 12

[56] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 4

[57] Jiahao Song and Yu-Zhao Wang. Musflow: Multimodal music generation via conditional flow matching. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10200–10209, 2025. 2

[58] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 9

[59] Inc. Suno. Suno, 2023. AI music generation platform. 1

[60] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR, 2016. 1

[61] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. In *International Conference on Machine Learning*, 2025. 7, 8

[62] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357, 2023. 3

[63] Duncan Williams, Victoria J Hodge, Lina Gega, Damian Murphy, Peter I Cowling, and Anders Drachen. Ai and automatic music generation for mindfulness. In *2019 AES International Conference on Immersive and Interactive Audio: Creating the Next Dimension of Sound Experience*. York, 2019. 2

[64] Yuxuan Wu, Yifan He, Xinlu Liu, Yi Wang, and Roger B Dannenberg. Transplayer: Timbre style transfer with flexible timbre control. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[65] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9452–9461, 2024. 2, 3

[66] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023. 3

[67] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*, 2017. 2

[68] Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, et al. Yue: Scaling open foundation models for long-form music generation. *arXiv preprint arXiv:2503.08638*, 2025. 7

[69] Sicheng Zhao, Yaxian Li, Xingxu Yao, Weizhi Nie, Pengfei Xu, Jufeng Yang, and Kurt Keutzer. Emotion-based end-to-end matching between image and music in valence-arousal space. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2945–2954, 2020. 2

[70] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024. 4, 6, 11