Diff4Steer: Steerable Diffusion Prior for Generative Music Retrieval with Semantic Guidance

Xuchan Bao[*] University of Toronto
Judith Yue Li Google Research
Zhong Yi Wan Google Research
Kun Su Google Research

Timo Denk Google DeepMind
Joonseok Lee Google Research, Seoul National University
Dima Kuzmin Google Research
Fei Sha Google Research

1 Introduction

Modern retrieval systems [13, 5], including those for music [11], often employ embedding-based dense retrieval system for candidate generation. These systems use a joint embedding model (JEM) [9, 6] to obtain deterministic representations of queries, known as seed embeddings, within a semantic space shared with the retrieval candidates. The seed embeddings provide the personalized starting point in the target embedding space for retrieving similar music via nearest neighbor search.

While JEM-based system provides computationally efficient retrieval solution, they are insufficient in modeling user's diverse and uncertain retrieval preference. First, JEM only supports user to express music preference or steer the retrieval results via specific modalities that the JEM is built on. Moreover, music discovery is inherently a task with many possible outcomes – there is not an one-to-one mapping between the query and seed embedding given the large uncertainty how a user's music preference can be fully specified. For example, "energetic rock music" could mean "punk rock" for some, or "hard rock" for others. Modeling user preference using deterministic seed embedding can lead to monotonous and inflexible recommendations [3]. In essence, for creative applications exploring the user's possible intention through allowing them to steer the retrieval system via instructions and returning many items that are highly likely relevant is crucial.

To better represent diversity and uncertainty in user's retrieval preference, we introduce a novel framework *Diff4Steer* (Figure 1) for music retrieval that leverages the strength of generative models for synthesizing potential directions to explore, implied by the generated "oracle" seed embeddings: a collection of vectors in the music embedding space that represent distribution of user's music preferences given retrieval queries. Concretely, our lightweight diffusion-based generative models give rise to a statistical prior on the target modality – audio in our application focus – for the music retrieval task. Furthermore, the prior can be steered by either image or text inputs, to generate samples in the audio embedding space learned by the pre-trained joint embedding model. They are then used to retrieve the candidates using nearest neighbor search. Given that constructing a large-scale multimodal dataset that contains the aligned multimodal data (steering info, source modality, target modality) is very difficult, we also leverage the generative models' ability on classifier-free guidance to help us to steer the models trained with bimodal alignment data, eliminating the need for expensive, data-hungry joint embedding training across all modalities.

While we have seen that diffusion-based generative approaches [17, 18, 14] can ensure diversity and quality in the embedding generation, in this work we investigate their performance on retrieval tasks. We demonstrate that our generative music retrieval framework achieves competitive retrieval and ranking metrics while introducing much-needed diversity. A comparison with deterministic

[*]Work done while intern at Google. Correspondence to judithyueli@google.com

Preprint. Under review.

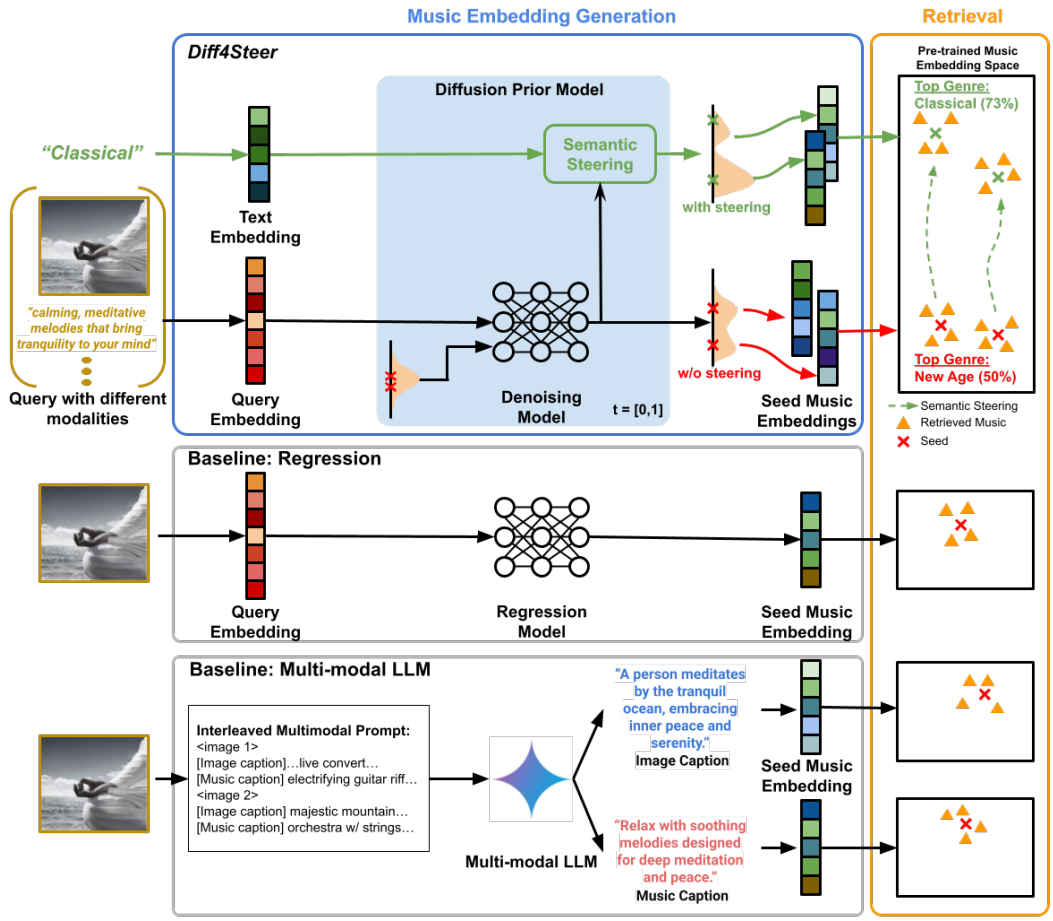

Figure 1: Overall diagram of our generative retrieval framework for cross-modal music retrieval, with comparison to the regression and multi-modal LLM baselines.

regression methods shows that *Diff4Steer* achieves superior retrieval metrics. This is thanks to the higher quality of the generated embedding, which reflects the underlying data distribution, as well as incorporating uncertainty in modeling user preferences.

2 Approach

Music Embedding Diffusion Prior Following the EDM [10] formulation, our diffusion prior is parametrized by a denoiser neural network $D(\tilde{z}_{\mathrm{m}}, \sigma, q)$, which learns to predict the clean music embedding z_{m} given a noisy embedding $\tilde{z}_{\mathrm{m}} = z_{\mathrm{m}} + \epsilon\sigma$, noise level σ and cross-modal query q, by minimizing the ℓ_2 loss:

$$L(\theta; \mathcal{D}) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),\sigma\sim\eta(\sigma),\{z_{\mathrm{m}},q\}\in\mathcal{D}} \left[\lambda(\sigma) \cdot \|D_\theta(\tilde{z}_{\mathrm{m}}, \sigma, q) - z_{\mathrm{m}}\|^2\right], \tag{1}$$

where ϵ draws from a standard Gaussian, η is a training distribution for σ, λ is the loss weighting, and $\mathcal{D}$ denotes training dataset with paired (z_{m}, q) examples. Sampling is performed by solving the stochastic differential equation (SDE)

$$d\tilde{z}_{\mathrm{m},t} = \left[(\dot{\sigma}_t/\sigma_t)\,\tilde{z}_{\mathrm{m},t} - 2(\dot{\sigma}_t/\sigma_t) D_\theta\left(\tilde{z}_{\mathrm{m},t}, \sigma_t, q\right)\right] dt + \sqrt{2\dot{\sigma}_t\sigma_t}\, dW_t, \tag{2}$$

from $t = 1$ to 0 with noise schedule σ_t and initial condition $\tilde{z}_{\mathrm{m},1} \sim \mathcal{N}(0, \sigma_{t=1})$, using a first-order Euler-Maruyama solver.

Classifier-free Guidance (CFG) [8] is used to enhance the alignment of the sampled music embeddings to the cross-modal inputs. During training, the condition q is randomly masked with a zero

vector with probability p_{mask}, such that the model simultaneously learns to generate conditional and unconditional samples with shared parameters. At sampling time, the effective denoiser D'_θ is an affine combination of the conditional and unconditional versions

$$D'_\theta(\tilde{z}_\text{m}, \sigma, q) = (1+\omega)D_\theta(\tilde{z}_\text{m}, \sigma, q) - \omega D_\theta(\tilde{z}_\text{m}, \sigma, \mathbf{0}), \tag{3}$$

where ω denotes the CFG strength, which boosts alignment with q when $\omega > 0$. $\omega = -1.0$ indicates unconditional generation.

Additional text steering can be applied when the underlying music embedding space is that of a text-music JEM [9]. In such case, the JEM provides a text encoder $E_\text{t} : T \to z_\text{t}$ with a text-music similarity measure via the vector dot product $\langle z_\text{t}, z_\text{m} \rangle$. This allows us to incorporate (potentially multiple) text steering signals by modifying the denoising function *at sampling time*:

$$D''_\theta\left(\tilde{z}_\text{m}, \sigma, q, [z_{\text{t},1}, z_{\text{t},2}, ...]\right) = D'_\theta(\tilde{z}_\text{m}, \sigma, q) + \sum_n k_n \nabla_{\tilde{z}_\text{m}} \left\langle D'_\theta(\tilde{z}_\text{m}, \sigma, q), z_{\text{t},n} \right\rangle, \tag{4}$$

where k_n is the strength for the n-th text steer signal $z_{\text{t},n}$. Each k_n can be positive/negative depending on whether features described by corresponding texts are desirable/undesirable in the samples. We note that such steering comes in addition to explicit condition q, which may itself contain text inputs.

3 Experimental settings

3.1 Tasks and Datasets

In our retrieval experiments, we use our diffusion prior model to perform several downstream tasks simultaneously, namely image-to-music retrieval, text-to-music retrieval and image-to-music retrieval with text steering. For image-to-music tasks the query embedding is CLIP [16]. For text-to-music retrieval or text steering, text is encoded via MuLan text embedding and incorporated as a steering condition to steer the seed embedding generation using genre or music caption.

YouTube 8M (YT8M) [1] is a dataset originally developed for the video classification task, equipped with video-level labels. We use the 116K music videos in this dataset to generate (music, image) pairs by extracting 10s audios and randomly sampling a video frame in the same time window. This dataset is primarily used for training.

We use two other expert-annotated datasets for evaluation. First, **MusicCaps (MC)** [2] is a collection of 10s music audio clips with human-annotated textual descriptions. We extend the dataset with an image frame extracted from the corresponding music video. **MelBench (MB)** [4] is another collection of images paired with matching music caption and music audio annotated by music professionals.

3.2 Model and training

We use a 6-layer ResNet with width of 4096 as the backbone of the denoising model. For classifier-free guidance, we use a condition mask probability $p_{\text{mask}} = 0.1$, in order to simultaneously learn the conditional and unconditional denoising model under shared parameters. We train the denoising model on paired image and music embeddings from the YT8M music videos. We use the Adam [12] optimizer under cosine annealed learning rate schedule [15] with peak rate 10^{-5}. Our model has 282.9M parameters in total and can fit into one TPU. We train our model for 2M steps, which takes around two days on a single TPU v5e device.

3.3 Baselines

MuLan. [9] As a text-music JEM, MuLan enables text-to-music retrieval through a nearest neighbor search based on the dot product similarity between a text query and candidate music embeddings.

Regression model. We train a regression baseline model that maps the query embeddings (CLIP image embedding) to MuLan audio embeddings deterministically using the same architecture as the diffusion model (excluding noise).

Multi-modal Gemini. The multi-modal Gemini serves as a strong baseline for our image-to-music retrieval tasks. We leverage a few-shot interleaved multi-modal prompt that given an image it can generate image caption or matching music caption. Specifically, *Gemini-ImageCap* encodes the

Table 1: FMD and MISCS of the generated music embeddings for YT8M, MC and MB datasets (image2music). Across all the datasets, our diffusion model outperforms the deterministic model in both embedding quality (FMD) and diversity (MISCS).

Method	FMD ↓			MISCS ↓		
	YT8M	MC	MB	YT8M	MC	MB
Regression	0.480	0.507	0.518	1.000	1.000	1.000
Diff4Steer (ours)	**0.161**	**0.152**	**0.172**	**0.805**	**0.391**	**0.404**

generated image caption into a MuLan text embedding for retrieving candidate audio embeddings. *Gemini-MusicCap* encodes the generated music caption into a MuLan text embedding for retrieving candidate audio embeddings.

3.4 Evaluation Metrics

Embedding Quality. We use two metrics to measure the quality of generated music embeddings: Fréchet MuLan Distance (FMD) and mean intra-sample cosine similarity (MISCS). FMD is inspired by Fréchet Inception Distance (FID) [7] and measures the similarity of a set of generated music embeddings to a population of real music embeddings in distribution.

Music-image Alignment (M2I). Assessing alignment between generated music embeddings and input images is challenging due to their distinct domains. Leveraging the shared text modality in CLIP and MuLan, we use text as a bridge for evaluating music-image (M2I) alignment following Chowdhury et al. [4] . This approach eliminates the need for paired data and instead requires a set of images and a separate set of texts. By encoding texts into both CLIP and MuLan embeddings, M2I is calculated as the average of the product of two cosine similarities.

Retrieval Metrics. We evaluate retrieval results using three metrics. First, we report recall@K (R@K), a standard metric in information retrieval. However, image-to-music or text-to-music retrieval is inherently subjective, often featuring one-to-many mappings. Thus, recall@K alone is insufficient, and we also report diversity using mean intra-sample cosine similarity (MISCS) and triplet accuracy (TA) to provide a more comprehensive evaluation.

4 Results and Discussion

In this section, we present experimental results that demonstrate: (1) our *Diff4Steer* model effectively generates high-quality seed embeddings using a diffusion-based approach; (2) *Diff4Steer* achieves competitive retrieval performance compared to other cross-modal retrieval methods and significantly improves retrieval diversity, and enables effective and personalized steering of seed embeddings during inference.

4.1 Quality of the Generated Seed Embeddings

Table 1 presents a comparison of embedding quality between *Diff4Steer* and the regression baseline for both image-to-music and text-to-music tasks across multiple datasets. Results show that our diffusion prior model consistently exhibits significantly lower FMD, indicating higher quality and greater realism in generated MuLan audio embeddings compared to the baseline. In addition, the diffusion model achieves significantly lower MISCS scores across all datasets, indicating that it allows us to generate diverse samples, which is impossible with a regression model.

There is a dynamic relationship between classifier-free guidance (CFG) strength ω and the quality and diversity of embeddings generated by our diffusion model. With a guidance strength of $\omega = -1.0$, corresponding to unconditional samples, FMD initially deteriorates, then improves, and eventually gets worse again with excessively high ω. Conversely, diversity consistently decreases with increasing ω, highlighting the inherent trade-off between embedding quality and diversity.

Table 2: Music retrieval results of our model and various baselines, evaluated on MC and MB.

Method	Input	MC w/ Images				MB			
		R@100	R@10	M2I	TA	R@100	R@10	M2I	TA
Gemini-ImageCap	image	0.215	0.055	89.12	0.488	0.162	0.036	90.32	0.685
Gemini-MusicCap	image	0.210	0.049	84.48	0.521	0.145	0.026	88.09	0.695
Regression	image	0.129	0.026	**96.21**	0.646	0.165	0.032	**95.79**	0.724
Diff4Steer (ours)	image	**0.334**	**0.105**	89.69	**0.778**	**0.341**	**0.086**	90.28	**0.836**
Regression (txt)	genre	0.378	0.103	**90.63**	0.838	0.147	0.016	**92.20**	0.739
Diff4Steer (ours)	genre	**0.389**	**0.108**	88.02	**0.855**	**0.165**	**0.019**	89.65	**0.762**
Regression (txt)	caption	0.419	**0.131**	**90.72**	0.871	0.380	**0.086**	**91.40**	0.872
Diff4Steer (ours)	caption	**0.435**	0.127	87.79	**0.877**	**0.384**	0.085	89.67	**0.876**
Diff4Steer (ours)	image + genre	0.425	0.165	**91.91**	0.889	0.384	0.090	**94.47**	0.883
Diff4Steer (ours)	image + caption	**0.536**	**0.184**	91.56	**0.915**	**0.488**	**0.141**	93.19	**0.916**

Table 3: Image-to-music evaluation on MB with genre diversity metrics.

ω	R@10 ↑	TA ↑	MISCS ↓	$\mathcal{H}$@10 ↑	$\mathcal{H}$@20 ↑	$\mathcal{H}$@50 ↑
-1.0	0.004	0.505	0.338	1.876	2.196	2.395
5.0	0.082	0.822	0.710	1.049	1.183	1.284
9.0	**0.089**	0.832	0.772	0.920	1.030	1.113
11.0	0.087	0.833	0.789	0.881	0.988	1.066
15.0	0.085	**0.834**	0.807	0.843	0.945	1.019

4.2 Embedding-based Music Retrieval

We show embedding-based music retrieval results in Table 2. The image CFG strength is an important hyperparameter, and we tune it using the FMD score, based on the YT8M evaluation split. For the remaining evaluations in this paper, we set the image guidance strength to be 19.0.

High-quality embeddings leads to high recall. A key finding from Table 2 is that our *Diff4Steer* model has significantly higher recall and triplet accuracy, compared to the regression and multi-modal Gemini baselines. This underscores the value of our approach for music retrieval applications. Notably, while the regression model has the highest M2I in the image-to-music task, it falls short in standard retrieval metrics. This observation, along with the FMD results in Section 4.1, highlights the crucial role of high-quality seed embeddings in achieving optimal retrieval performance.

Modality gap may harm retrieval results. For the multi-modal Gemini baselines, the image-to-music embedding generation is broken down to multiple stages. We use text (image or music captions) as an intermediate modality, thereby introducing potential modality gap. As shown in Table 2, despite the power of the general-purpose LLMs, multi-modal Gemini baselines have worse retrieval performance than our *Diff4Steer* model, likely due to the loss of information with the modality gap. Additionally, our model offers a significantly lighter weight solution in terms of training consumption and latency compared to multi-modal foundation models.

One model for all modality. Notably, our *Diff4Steer* model demonstrates competitive performance on genre-to-music and caption-to-music retrieval tasks (the second and third groups in Table 2) despite not being trained on paired text and music data. This is achieved by unconditionally generating audio embeddings guided by text-music similarity. Compared to the regression baseline, *Diff4Steer* achieves superior results on most retrieval and ranking metrics, especially on the tasks that involve higher retrieval uncertainty, *e.g.*, genre-to-music retrieval.

Text steering improves recall. Furthermore, we explore the extent to which text steering helps with retrieval. In addition to the image input, we also provide our diffusion model with the genre label or ground truth caption at inference time. The last group in Table 2 shows that when steered with the additional textual information, the diffusion prior achieves significantly higher recall and triplet accuracy.

Input Image	Ground Truth Genre	GS=5.0	GS=9.0	GS=15.0
	**Classical**	**Entropy**: 1.307 **Top-3 Genre**: Classical (64%) Folk Acoustic (10%) Jazz (6%)	**Entropy**: 1.127 **Top-3 Genre**: Classical (68%) Folk Acoustic (10%) Easy Listening (6%)	**Entropy**: 0.969 **Top-3 Genre**: Classical (74%) Folk Acoustic (8%) Easy Listening (6%)
	**Folk Acoustic**	**Entropy**: 1.614 **Top-3 Genre**: Folk Acoustic (44%) Rock (20%) New Age (14%)	**Entropy**: 1.621 **Top-3 Genre**: Folk Acoustic (36%) New Age (26%) Rock (22%)	**Entropy**: 1.581 **Top-3 Genre**: Folk Acoustic (34%) Rock (30%) New Age (22%)
	**New Age**	**Entropy**: 2.142 **Top-3 Genre**: New Age (30%) World Traditional (22%) Classical (8%)	**Entropy**: 1.489 **Top-3 Genre**: New Age (50%) World Traditional (24%) Classical (10%)	**Entropy**: 1.100 **Top-3 Genre**: New Age (64%) World Traditional (22%) Pop (6%)

Figure 2: Given an input image and various guided strengths (GS), we generate seed embeddings and retrieve their nearest music piece in MB. We show entropy and the probabilities of Top-3 genres. A higher entropy indicates more diverse music genres of retrieved music pieces.

4.3 Retrieval Diversity

Diff4Steer generates diverse seed embeddings, as quantified in Table 3. For each image, we generate 50 seed embeddings and measure diversity using MISCS and entropy ($\mathcal{H}@K$, with $K \in \{10, 20, 50\}$), calculated on the distribution of ground-truth genres in retrieved music pieces. Varying guided strengths ω during inference effectively modulates this diversity. Unconditional generation ($\omega = -1.0$) yields the lowest MISCS and highest entropy in recommended genres. Increasing GS initially decreases embedding diversity, with retrieval metrics peaking around $\omega = 9.0$ before declining.

Figure 2 illustrates retrieval diversity using three representative input images. With strong image-music correspondence (Top), the entropy is notably lower, reflecting a dominant genre (Classical). Increasing image guidance further amplifies this effect. Conversely, weaker correspondences (Middle, Bottom) show varied entropy changes with increased guidance, sometimes resulting in a dominant genre (Bottom), sometimes maintaining a balance (Middle). In both scenarios, our model generally retrieves music from accurate genres.

5 Conclusion and limitations

We introduce a novel generative music retrieval framework featuring a diffusion-based embedding-to-embedding model. By generating non-deterministic seed embeddings from cross-modal queries, our approach improves the quality and diversity of music retrieval results. Our model ensures semantic relevance and high quality, while text-based semantic steering allows user personalization. Extensive evaluations, including personalized retrieval experiments and human studies, show our method's superiority over existing alternatives.

While promising, our framework has limitations as well. High computational demands of diffusion sampling may impede real-time retrieval, and any issues with pre-trained JEMs, such as information loss or underrepresented items, naturally extend to our framework. Additionally, reliance on large, potentially biased training datasets may introduce biases into retrieval results. Future work should address these challenges to improve the retrieval effectiveness of music recommender systems.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

[3] A. Anderson, L. Maystre, I. Anderson, R. Mehrotra, and M. Lalmas. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of the web conference 2020*, pages 2155–2165, 2020.

[4] S. Chowdhury, S. Nag, J. KJ, B. V. Srinivasan, and D. Manocha. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] P. Covington, J. K. Adams, and E. Sargin. Deep neural networks for youtube recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016. URL https://api.semanticscholar.org/CorpusID:207240067.

[6] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang. Clap: Learning audio concepts from natural language supervision, 2022.

[7] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[8] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[9] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. Ellis. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*, 2022.

[10] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[11] D. M. Kim, K. Kim, K.-H. Park, J.-H. Lee, and K.-M. Lee. A music recommendation system with a dynamic k-means clustering algorithm. *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 399–403, 2007. URL https://api.semanticscholar.org/CorpusID:17603549.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Y. Koren, R. M. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42, 2009. URL https://api.semanticscholar.org/CorpusID:58370896.

[14] H. Liu, Q. Tian, Y. Yuan, X. Liu, X. Mei, Q. Kong, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023.

[15] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[18] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023.