Repulsive Guidance for Memorization Mitigation in Text-to-Music Diffusion Models

Taehyeon Kim , Hangyeol Lee , Chang Wook Ahn * , Man-Je Kim *

Posted Date: 12 March 2026

doi: 10.20944/preprints202603.0982.v1

Keywords: music generation; memorization; repulsive guidance; attraction basin

 doi:10.20944/preprints202603.0982.v1

Article

Repulsive Guidance for Memorization Mitigation in Text-to-Music Diffusion Models

Taehyeon Kim [1], Hangyeol Lee [1], Chang Wook Ahn [1,*] and Man-Je Kim [2,*]

[1] AI Graduate School, Gwangju Institute of Science and Technology, Gwangju 61005, Republic of Korea
[2] Convergence of AI, Chonnam National University, Gwangju 61186, Republic of Korea
* Correspondence: cwan@gist.ac.kr (C.W.A.); jaykim0104@jnu.ac.kr (M.-J.K.)

Abstract

Recent progress in text-to-music generation has enabled high-quality audio synthesis from natural language prompts. However, such models are at risk of unintended replication, raising concerns regarding originality and intellectual property. While training-time mitigation strategies can address this issue, they typically require retraining or curated datasets, limiting their practicality for large-scale systems. Inference-time methods provide a more lightweight alternative but often involve a trade-off between fidelity and memorization risk. This work introduces Repulsive Guidance (RG), a systematic inference-time mitigation strategy that reduces memorization without disrupting the intended conditional guidance from the text prompt. RG operates by enforcing divergence between dual diffusion trajectories through a repulsive term applied only during early denoising steps, without reversing the conditional guidance from the prompt. Experiments on MusicBench with the TANGO model demonstrate that RG offers a complementary mitigation strategy, providing new insights into balancing fidelity and memorization risk.

Keywords: music generation; memorization; repulsive guidance; attraction basin

1. Introduction

Text-to-music generation has recently achieved significant progress, producing high-quality and stylistically coherent audio from natural-language descriptions. Models such as MusicGen [1], MusicLM [2], Moûsai [3], and Mustango [4] demonstrate that large-scale generative architectures can capture both semantic and musical structure, enabling controllable and expressive synthesis. As these models are increasingly deployed in creative applications, however, concerns regarding originality and intellectual property have become more prominent. When trained on large collections of copyrighted audio, diffusion-based generators may unintentionally reproduce training examples, leading to potential legal and ethical risks.

Despite this progress, concerns remain about originality and intellectual property, as models trained on copyrighted data risk unintended replication. Studies on audio latent diffusion models [5–7] show that replication arises more frequently when dataset size is limited or duplicated, and that sampling dynamics strongly influence whether a model converges toward memorized outputs. Evaluations of large-scale music generation systems such as MusicLM [2] and MusicLDM [8] similarly report that a nontrivial portion of generated audio exhibits near-duplicate similarity to training data. These findings motivate the development of techniques that can mitigate memorization while preserving the perceptual fidelity and semantic alignment of the generated content.

Training-time mitigation strategies, such as dataset curation or mixup-based augmentation [8], can reduce the likelihood of replication but require retraining or modifying large-scale datasets. In many practical scenarios, especially when working with pre-trained or proprietary models, retraining is infeasible. Unlike training-time approaches, inference-time mitigation approaches in latent diffusion models do not require retraining and can be grouped into trajectory-level and prompt-level

 doi:10.20944/preprints202603.0982.v1

interventions. Trajectory-level interventions [9,10] (Figure 1 (a)) explicitly steer denoising paths away from memorization-prone regions. [9] employs multiple classifier-free guidance (CFG) [11] weight schedules to address different causes of memorization whenever the similarity to a nearest training sample exceeds a threshold, but suffers from high latency due to per-step neighbor checks. In contrast, [10] showed that early or strong CFG induces an *attraction basin* in latent space, where trajectories converge and cause memorization, and demonstrated that applying the opposite CFG can mitigate memorization without explicit nearest-neighbor comparisons. Meanwhile, prompt-level interventions [12,13] (Figure 1 (b)) mitigate overfitting by perturbing trigger tokens, thereby weakening reliance on specific prompts and indirectly reshaping the attraction basin that drives memorization. However, whether such inference-time interventions transfer effectively to text-to-music diffusion, where long-range temporal and harmonic dependencies shape the generative process, remains an open question.

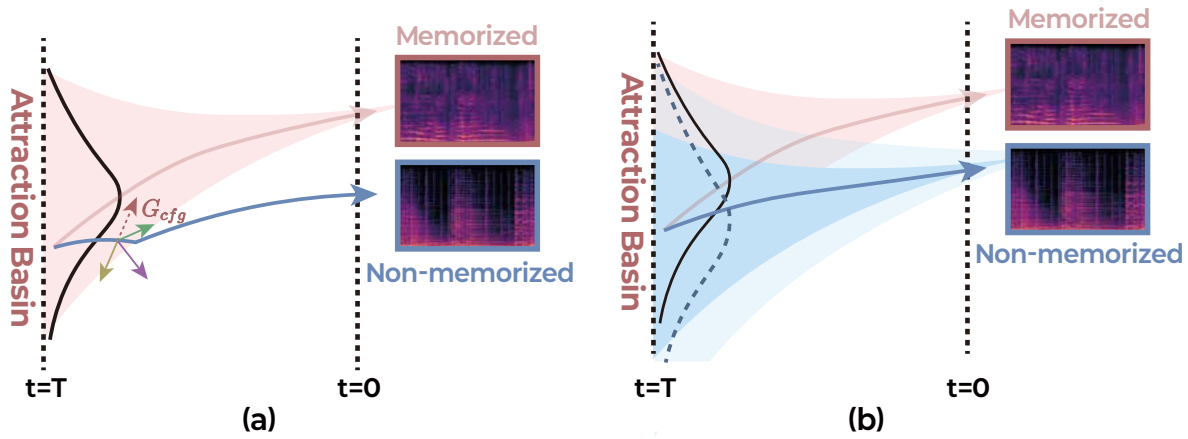

Figure 1. (a) Trajectory-level Intervention [9,10] (b) Prompt-level Intervention [12,13]

In this work, we investigate how inference-time memorization mitigation strategies can be adapted to text-to-music diffusion models while preserving perceptual fidelity. To this end, we propose *Repulsive Guidance (RG)*, an inference-time mitigation method that introduces an additional guidance component to counteract memorization without disrupting the conditional guidance from the text prompt. The main contributions of this work are summarized as follows:

- **Repulsive Guidance Formulation:** We introduce RG as a lightweight inference-time strategy that encourages controlled divergence between diffusion trajectories without requiring retraining.
- **Adaptation to Music Generation:** We analyze memorization in the music domain and highlight challenges that are not adequately addressed by inference-time mitigation methods developed for image generation.
- **Fidelity–Safety Trade-off Analysis:** Through experiments across dataset scales and parameter settings, we provide an empirical analysis of how RG balances perceptual fidelity and memorization risk in text-to-music generation.

The remainder of this paper is organized as follows. Section 2 reviews related work and provides background on memorization in diffusion models. Section 3 presents the proposed Repulsive Guidance method. Section 4 describes the experimental setup and evaluation method. Section 5 reports and analyzes the experimental results. Finally, Section 6 concludes the paper.

2. Related Work

2.1. Latent Diffusion Models and Classifier-Free Guidance

Let $\mathbf{x} \in \mathbb{R}^N$ denote a discrete audio waveform of length N samples. A latent diffusion model [14] first compresses the audio into a low-dimensional autoencoder latent $\mathcal{E}(\mathbf{x}) = \mathbf{z}$ through an encoder $\mathcal{E}$, and reconstructs it using a decoder $\mathcal{D}$ as $\mathcal{D}(\mathbf{z}) = \mathbf{x}$. The diffusion process is then applied in the latent

 doi:10.20944/preprints202603.0982.v1

space $\mathbf{z}$. Let $z_0 = \mathbf{z}$ denote the initial diffusion latent. During the forward diffusion process, Gaussian noise is gradually added over T timesteps so that the final latent z_T approaches an isotropic Gaussian distribution $\mathcal{N}(0, I)$. At an arbitrary timestep t, the marginal distribution is given by

$$q(z_t \mid z_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, z_0,\ (1 - \bar{\alpha}_t)\mathbf{I}\big), \tag{1}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\beta_t \in (0, 1)$, $t = 1, \ldots, T$, is the noise schedule controlling the amount of noise injected at each step. The reverse process is parameterized by a neural network ϵ_θ that predicts the noise $\boldsymbol{\epsilon}_t$ from the noisy latent z_t and the timestep t. The model is trained by minimizing

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{\varepsilon} \sim \mathcal{N}(0,I),\, t}\big[\big\|\boldsymbol{\varepsilon}_t - \varepsilon_\theta(z_t, t)\big\|_2^2\big], \tag{2}$$

where conditioning on a text prompt is incorporated by providing the prompt embedding as an additional input to $\varepsilon_\theta(\cdot)$.

This latent diffusion formulation has emerged as a central framework for text-to-music and text-to-audio generation, as operating in a compressed latent space substantially reduces computational cost while maintaining perceptual fidelity. Originally introduced for image synthesis, this approach has since been adapted to music domains in systems such as MusicLDM [8] and Mustango [4], demonstrating that latent-space diffusion scales effectively to large datasets and complex musical structures.

A key component enabling precise semantic control in these models is CFG [11]. CFG jointly trains conditional and unconditional objectives by randomly dropping conditioning during training, and at inference time it injects conditional information into the denoising trajectory by interpolating between unconditional and conditional noise predictions:

$$\hat{\varepsilon}_t = \varepsilon_u + w \cdot (\varepsilon_c - \varepsilon_u), \tag{3}$$

where ε_u and ε_c denote the unconditional and conditional noise estimates, respectively, and w is the guidance scale. Larger values of w generally strengthen prompt adherence but simultaneously reduce sample diversity and amplify over-fitting tendencies.

2.2. Memorization in Text-to-Audio Diffusion Models

Memorization in diffusion-based generators is increasingly recognized as a systematic behavior shaped by finite training data and sampling dynamics, rather than a rare anomaly [9,10]. In text-to-music generation, for example, MusicLDM [8] reported replication-related concerns and proposed beat-synchronous mixup strategies to improve novelty, suggesting that diffusion-based music generators can converge toward memorization-prone regions of the training distribution.

To quantify and diagnose such behavior, recent work has developed evaluation and attribution tools within audio-generation pipelines. [5] analyzed memorization in audio latent diffusion models [15] trained on AudioCaps [16], showing that observed replication varies with dataset scale and the choice of similarity metrics. [6] proposed a model-agnostic evaluation protocol based on multiple music similarity measures, and [7] introduced an embedding-based attribution approach that links generated outputs to specific training audio samples. Collectively, these studies position memorization as a fundamental challenge for safe and trustworthy audio generative modeling. Although primarily developed in the broader audio-generation context, these findings naturally extend to text-to-music systems, where long-range harmonic and rhythmic coherence can render memorization even more perceptually pronounced.

 doi:10.20944/preprints202603.0982.v1

Algorithm 1 RG for Text-to-Music Diffusion Models

Input: Prompt $\mathbf{c}$, diffusion steps T, guidance scale w, repulsion strength λ
Output: Generated latent $z_0^{(k^*)}$

Initialize dual trajectories $z_T^{(1)}, z_T^{(2)} \sim \mathcal{N}(0, I)$
Initialize risk scores $s^{(1)} \leftarrow 0$, $s^{(2)} \leftarrow 0$
Set transition rule for τ_1, τ_2
for $t = T$ **down to** 1 **do**
  Compute CFG terms $\Delta\varepsilon_t^{(1)}, \Delta\varepsilon_t^{(2)}$
  Compute normalized repulsive vector g_t (Equation (5))
  for $k = 1$ **to** 2 **do**
    Compute repulsion scale β_k (Equation (6))
    Compute guided noise $\hat{\varepsilon}_t^{(k)}$ (Equation (7))
    Accumulate risk scores $s^{(k)} \leftarrow s^{(k)} + \|\Delta\varepsilon_t^{(k)}\|_2$
    Update trajectory $z_{t-1}^{(k)}$ via DDPM
  end for
end for
Select output $k^* = \arg\min_k s^{(k)}$
return Generated latent $z_0^{(k^*)}$

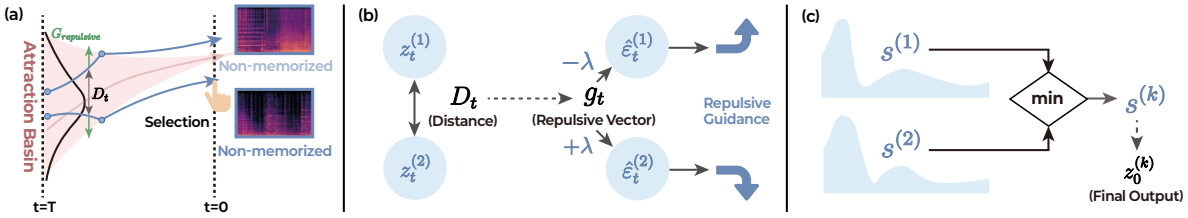

Figure 2. Overview of the proposed Repulsive Guidance (RG). (a) Dual-trajectory sampling begins from two independently drawn noise latents. (b) At each timestep, the repulsive direction is computed from the distance between trajectories and modulated by a time-dependent repulsion scale. (c) The cumulative CFG magnitude is tracked for both trajectories, and the final output is selected as the sample with the lower accumulated risk.

3. Proposed Method

As outlined in Algorithm 1, we propose RG, an inference-time strategy that mitigates memorization by enforcing divergence between two diffusion trajectories. Starting from independent Gaussian noise, two latent samples $z_t^{(1)}$ and $z_t^{(2)}$ evolve in parallel. At each timestep t, we compute their CFG directions $\Delta\varepsilon_t^{(k)}$ and accumulate their ℓ_2 norms $s^{(k)}$ as a proxy for memorization risk, motivated by prior observations that memorized samples exhibit consistently larger CFG magnitudes. During early denoising steps ($t > \tau_k$), where τ_k denotes a trajectory-specific transition point, the latent distance D_t is used to normalize a repulsive vector g_t, which is then applied with opposite signs to $\varepsilon_t^{(1)}$ and $\varepsilon_t^{(2)}$. The resulting adjusted directions are used in the DDPM [17] update, explicitly counteracting trajectory collapse and encouraging controlled divergence. After sampling completes, the trajectory with the lower accumulated risk is selected as the final output $z_0^{(k)}$.

3.1. Repulsive Direction Calculation

To prevent the two latent trajectories from collapsing into the same attraction basin, we introduce an explicit repulsive interaction during the denoising process. At each timestep t, we first measure the distance between the two latent states:

$$D_t = \|z_t^{(2)} - z_t^{(1)}\|_2. \tag{4}$$

 doi:10.20944/preprints202603.0982.v1

This distance is then used to define a repulsive direction that encourages separation between the trajectories. Specifically, we compute a normalized repulsion vector pointing from $z_t^{(1)}$ to $z_t^{(2)}$, normalized by the distance with a small constant ε for numerical stability:

$$g_t = \frac{z_t^{(2)} - z_t^{(1)}}{D_t + \varepsilon}. \tag{5}$$

The normalized vector g_t is applied with opposite signs to the two trajectories in subsequent updates, explicitly pushing them apart in latent space. This design ensures that even when the initial noise samples are close, the corresponding denoising paths are actively encouraged to diverge, reducing the likelihood of trajectory collapse.

3.2. Repulsive Scale Calculation

The normalized repulsive vector g_t is weighted by a scaling factor β_k, which determines both the magnitude and the direction of the repulsion applied to each trajectory $z_t^{(k)}$. The value of β_k depends on the trajectory index $k \in \{1, 2\}$ and whether the denoising step has passed its transition point. Formally,

$$\beta_k = \begin{cases} (-1)^k \lambda, & \text{if } t > \tau_k, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where λ controls the strength of the repulsive force. The transition point τ_k denotes the timestep at which repulsion is disabled and is set to the first local minimum of the CFG norm, following [10]. In practice, it is detected online during sampling, although a static schedule may be used as a simplified alternative. This design restricts repulsion to early denoising steps, when trajectory collapse is most likely to occur.

3.3. Trajectory Update with Repulsive Guidance Term

At each timestep t, the final denoising direction for each trajectory $z_t^{(k)}$ is obtained by combining three components: the unconditional noise prediction $\varepsilon_u^{(k)}$, the CFG term $w\,\Delta\varepsilon_t^{(k)}$, and the repulsive adjustment $\beta_k g_t$. The combined update is given by

$$\hat{\varepsilon}_t^{(k)} = \varepsilon_u^{(k)} + w\,\Delta\varepsilon_t^{(k)} + \beta_k g_t. \tag{7}$$

In this formulation, the CFG component $\Delta\varepsilon_t^{(k)} = \varepsilon_c^{(k)} - \varepsilon_u^{(k)}$ maintains alignment with the input prompt, while the repulsive term $\beta_k g_t$ encourages the two trajectories to diverge in latent space. By applying repulsion only when needed and without altering the intended conditional direction, this update prevents the trajectories from collapsing into memorized regions during denoising.

3.4. Sample Selection by Risk Minimization

Prior work has shown that memorized outputs tend to exhibit larger CFG magnitudes throughout sampling [10,13]. Following this observation, we measure the memorization risk of each trajectory by accumulating the ℓ_2 norm of its CFG term over all timesteps. For trajectory $k \in \{1, 2\}$, the risk score is defined as

$$s^{(k)} = \sum_{t=1}^{T} \|\Delta\varepsilon_t^{(k)}\|_2. \tag{8}$$

A lower value of $s^{(k)}$ indicates weaker attraction toward training examples and therefore a safer trajectory. Once denoising is complete, the final output $z_0^{(k)}$ is chosen as the sample associated with the lower accumulated risk.

 doi:10.20944/preprints202603.0982.v1

4. Experiments

4.1. Experimental Designs

We evaluate the proposed RG under controlled settings using the MusicBench dataset [4] with the TANGO backbone [15], which is known to exhibit memorization behavior [5]. MusicBench extends MusicCaps [2] by adding musically relevant control signals, enriched captions, and systematic audio–text augmentations, resulting in about 53K text–audio pairs. To assess scalability, we use both a 1K subset and the full dataset, which enables a systematic comparison of data size effects on memorization and model robustness.

As baselines, we include (1) No Mitigation, representing the unmitigated TANGO model, (2) prompt-level intervention methods [12], including random token replacement & addition (RT), caption word repetition (WR), and random number addition (RNA), which were originally proposed to reduce overfitting to trigger tokens, and (3) Opposite Guidance (OG) [10], a trajectory-level intervention that reverses the CFG direction to avoid the attraction basin. Anti-Memorization Guidance [9] is not included, as its per-step nearest-neighbor checks against the training set both require direct access to training data at inference and incur substantial computational costs on large-scale datasets like MusicBench. Instead, we adopt OG as a representative trajectory-level baseline.

We train the TANGO [15] backbone for 40 epochs on four NVIDIA A100 GPUs using AdamW [18] optimizer with a learning rate of 4.5e-5 and a batch size of 24. All mitigation methods are applied only at inference time during sampling. Inference is performed with 200 denoising steps on a single A100 GPU, and each configuration is repeated 10 times with different random seeds. The trajectory update follows [15] with $w = 3.0$, while $\lambda = 1.0$ is empirically validated as a balance point across experiments.

4.2. Evaluation Metrics

To assess the fidelity and safety of generated music, we employ the following evaluation metrics:

Fidelity Metrics: To assess the fidelity of the generated music, following prior work [4,8], we adopt the Fréchet Distance (FD) metric computed with two different audio embeddings. Specifically, FD_{pann} is calculated using embeddings from a PANNs [19] classifier, while FD_{vgg} employs embeddings from VGGish [20]. Using both provides complementary views of fidelity, and lower FD values indicate that the generated audio is closer to real data in the perceptual embedding space.

Safety (Similarity) Metrics: We compute similarity using CLAP [21] audio embeddings, following the previous studies [5,8]. Given embeddings of a generated sample E^g and a training sample E^t, similarity is defined as cosine similarity $SIM(E^g, E^t) = \frac{E^{g\top} \cdot E^t}{\|E^g\| \|E^t\|}$. We report SIM_{Avg}, the mean similarity across all training samples, and SIM_{90}, the proportion of generated samples with similarity above 0.9. While SIM_{90} highlights highly memorized cases, SIM_{Avg} captures broader distributional overlap with training data, offering a complementary view of memorization risk. Lower values indicate reduced risk of replication.

5. Results

5.1. Main Results

Table 1 reports results on MusicBench with both the 1K subset and the full training set. The *No Mitigation* baseline exhibits consistently high similarity and the highest SIM_{90} in both regimes, confirming that TANGO can produce near-duplicate outputs under memorization-prone settings. Among the prompt-level interventions, WR and RNA slightly reduce similarity but degrade fidelity (FD_{pann}, FD_{vgg}) on the full training set. RT is more effective in lowering SIM_{90}, but this comes at the cost of substantial fidelity loss, revealing the inherent fidelity–safety trade-off. OG reduces SIM_{90} but increases SIM_{Avg}, suggesting that guidance reversal can suppress extreme near-duplicate cases while pushing samples toward a more training-like manifold in terms of average proximity. Moreover, OG degrades fidelity, particularly on the full training set. This indicates that reversing CFG from the prompt may incur fidelity degradation.

 doi:10.20944/preprints202603.0982.v1

Table 1. Evaluation of generation fidelity and similarity for RG compared with baselines on MusicBench (1K and Full). Bold values indicate the best performance, and underlined values indicate the second best.

	1K Size MusicBench Dataset				Full Size MusicBench Dataset			
Model	FD_{pann}	FD_{vgg}	SIM_{Avg}	SIM_{90}	FD_{pann}	FD_{vgg}	SIM_{Avg}	SIM_{90}
No Mitigation	41.41	4.41	0.5345	0.1592	28.44	2.65	0.4589	0.0432
RT	51.25	6.19	0.5372	**0.0238**	39.84	5.01	0.4734	**0.0151**
WR	41.35	4.40	0.5348	0.1589	28.85	2.75	<u>0.4578</u>	0.0414
RNA	**41.18**	4.42	0.5341	0.1545	28.68	2.77	0.4588	0.0426
OG	<u>41.19</u>	4.43	0.5355	<u>0.1479</u>	28.68	2.81	0.4604	<u>0.0395</u>
RG ($\lambda = 0.0$)	41.87	**4.22**	<u>0.5283</u>	0.1584	<u>28.14</u>	<u>2.63</u>	<u>0.4578</u>	0.0415
RG ($\lambda = 1.0$)	41.81	<u>4.32</u>	**0.5279**	0.1521	**28.05**	**2.59**	**0.4564**	0.0421

In contrast, RG exhibits a different trade-off behavior. Across both dataset scales, RG reduces similarity while preserving fidelity close to the baseline. The only observable degradation appears in FD_{pann} under the 1K setting, with no consistent fidelity decline in other configurations, suggesting that the impact on perceptual quality is minimal and condition-specific. Notably, RG consistently lowers SIM_{Avg} across all settings and achieves a slight reduction in SIM_{90}. Given that SIM_{90} captures highly memorized, near-duplicate cases, whereas SIM_{Avg} reflects broader distributional overlap with the training data, these results indicate that RG effectively mitigates diffuse replication tendencies rather than solely suppressing extreme memorization instances.

5.2. Ablation Studies

In Table 1, we compare two variants of RG, namely $\lambda = 0.0$, which applies only risk-based trajectory selection without repulsion, and $\lambda = 1.0$, which incorporates the repulsive term during sampling. Both variants maintain fidelity metrics close to the baseline across dataset scales. However, incorporating the repulsive term consistently results in lower SIM_{Avg} values, indicating that explicit trajectory separation provides additional mitigation effects beyond trajectory selection alone. These patterns persist in both the 1K and full-data regimes, highlighting the stability and scalability of RG with respect to dataset size.

To provide qualitative evidence of the repulsive term's effect, Figure 3 illustrates spectrogram pairs generated from an identical text prompt under different repulsive scales. When $\lambda = 0.0$, the two trajectories tend to converge toward highly similar outputs. In contrast, with $\lambda = 1.0$, the trajectories evolve toward perceptually distinct realizations, indicating that the repulsive mechanism actively counteracts trajectory collapse. These qualitative observations align with the quantitative results, demonstrating that RG promotes controlled divergence between sampling paths, thereby reducing replication risk while preserving meaningful diversity.

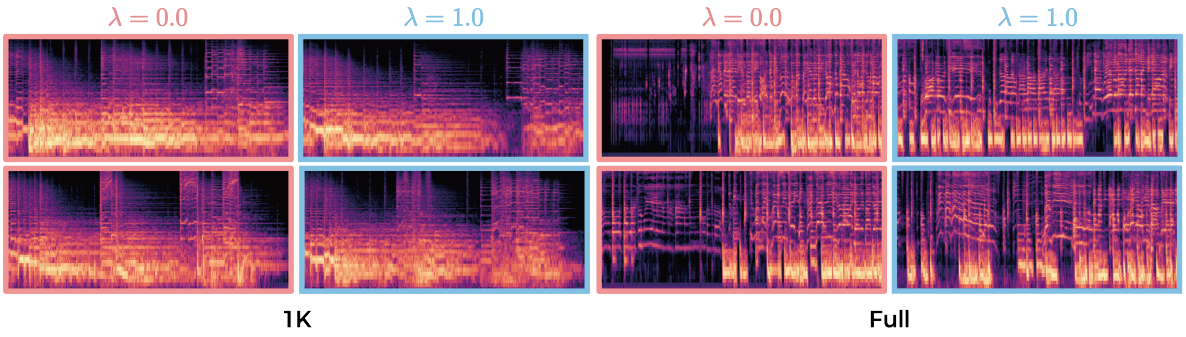

Figure 3. Examples of generated spectrogram pairs under different repulsive scales λ. From the same prompt, $\lambda = 0.0$ (pink) yields similar outputs, whereas $\lambda = 1.0$ (blue) produces more divergent results, showing that repulsive guidance enhances diversity by avoiding the shared attraction basin.

 doi:10.20944/preprints202603.0982.v1

5.3. Trade-Off Analysis with Static Transition Point

Following [10], we further evaluate RG under a static transition schedule defined as $\tau_{\text{static}} = T \times (1 - r_t)$, where the application rate $r_t \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Figure 4 presents the results on the 1K dataset. Panel (a) illustrates the Pareto relationship between FD_{vgg} and SIM_{Avg}. As r_t increases, OG progressively reduces similarity. However, this reduction is accompanied by a marked increase in FD_{vgg}, indicating substantial fidelity degradation. In contrast, RG remains close to the baseline across both metrics, occupying a stable region near the lower-left area of the trade-off space.

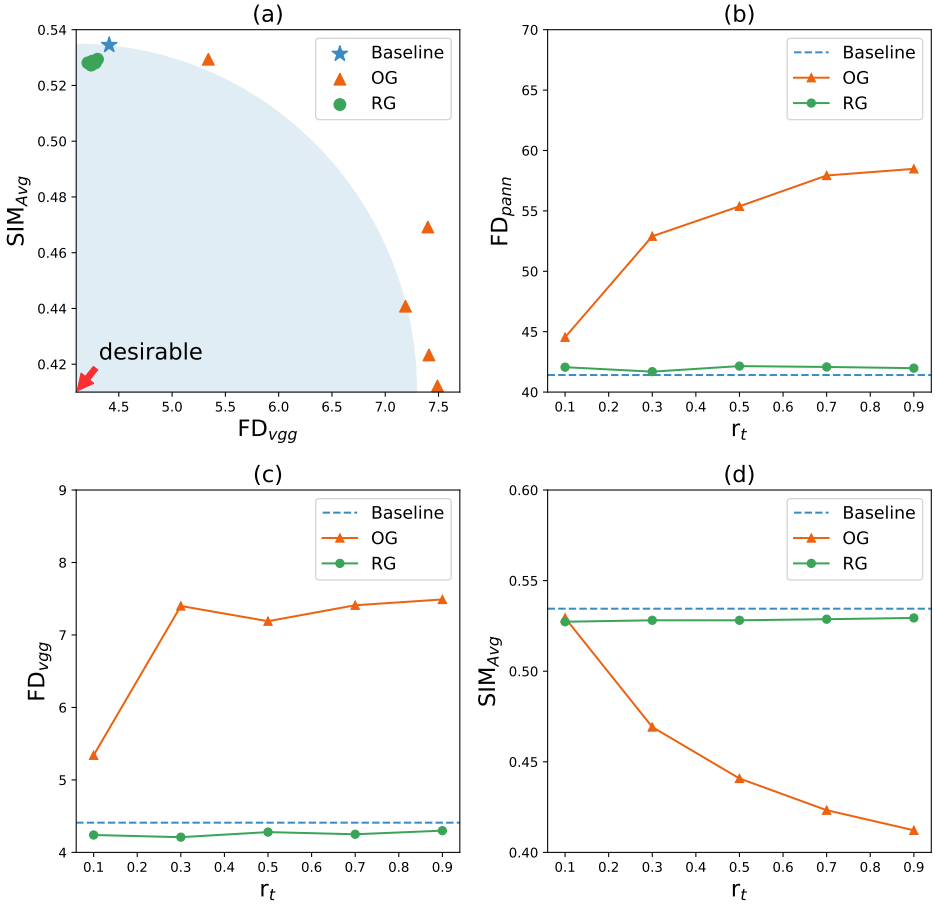

Figure 4. Comparison of RG and OG on MusicBench (1K) across application time rates r_t. (a) Pareto trade-off between FD_{vgg} and SIM_{Avg}. (Lower-left indicates better fidelity–safety.) (b) FD_{pann} values across r_t. (c) FD_{vgg} values across r_t. (d) SIM_{Avg} values across r_t.

Panels (b)–(d) provide a more detailed view across individual metrics. For OG, both FD_{pann} and FD_{vgg} increase sharply as r_t grows, while SIM_{Avg} decreases. This pattern reflects an explicit fidelity–safety trade-off driven by stronger intervention. However, RG exhibits minimal variation in fidelity metrics across all r_t values, while maintaining a moderate reduction in similarity relative to the baseline. These results indicate that RG achieves similarity mitigation without inducing large perceptual distortion.

Figure 5 reports the same analysis on the full MusicBench dataset. The overall trends remain consistent. OG again demonstrates a monotonic reduction in similarity accompanied by substantial fidelity degradation as r_t increases. By contrast, RG preserves fidelity metrics near baseline levels across all application rates and maintains stable similarity values, showing no evidence of escalating distortion under stronger scheduling.

 doi:10.20944/preprints202603.0982.v1

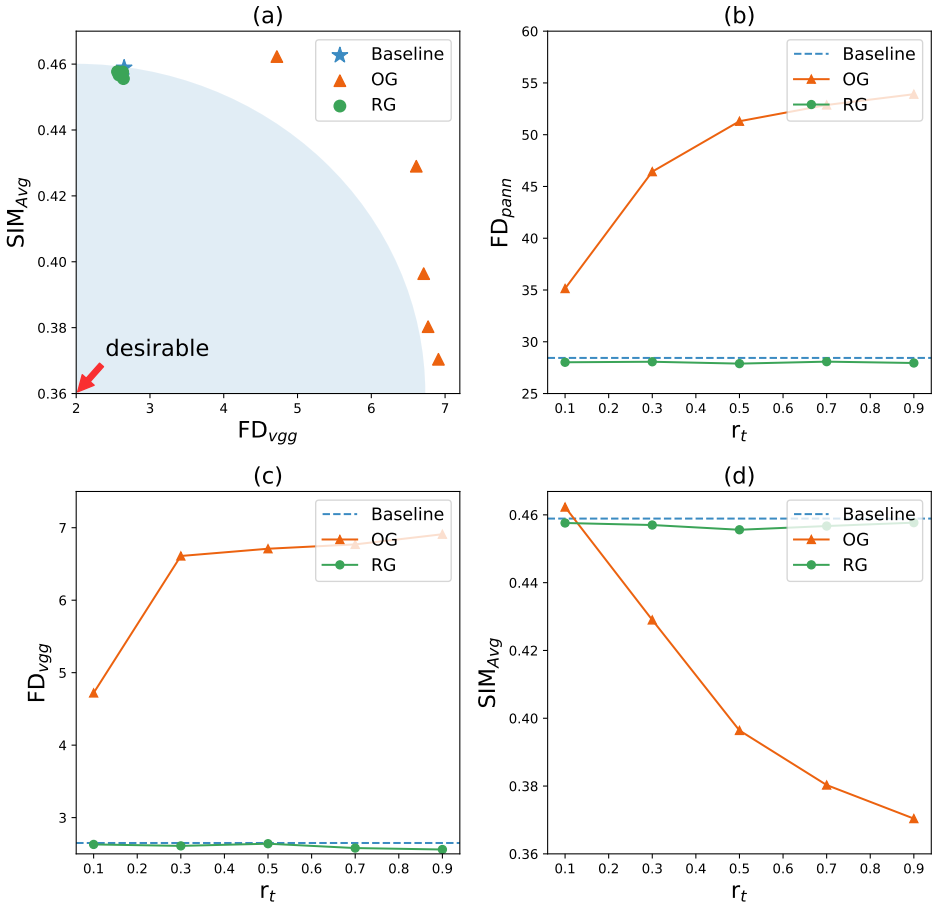

Figure 5. Comparison of RG and OG on MusicBench (Full) across application time rates r_t. (a) Pareto trade-off between FD_{vgg} and SIM_{Avg}. (Lower-left indicates better fidelity–safety.) (b) FD_{pann} values across r_t. (c) FD_{vgg} values across r_t. (d) SIM_{Avg} values across r_t.

Taken together, Figures 4 and 5 suggest that RG mitigates memorization by guiding sampling trajectories toward a stable region of the fidelity–safety landscape. Unlike approaches that invert or counteract conditional guidance, RG introduces controlled divergence between parallel trajectories while preserving the primary conditional direction. This mechanism reduces memorization without sacrificing prompt alignment or perceptual quality. Importantly, this behavior generalizes across dataset scales and a wide range of scheduling configurations, indicating that RG functions as a stable and practically deployable inference-time mitigation strategy.

6. Conclusion

This work proposed Repulsive Guidance (RG), an inference-time approach designed to mitigate memorization in text-to-music diffusion models. By encouraging controlled divergence between dual diffusion trajectories and selecting the path with lower accumulated risk, RG provides a practical mechanism for mitigating unintended replication. The method preserves prompt alignment and perceptual fidelity while offering a complementary perspective on the broader fidelity–safety trade-off observed in generative models. Experimental results on MusicBench demonstrate that RG reliably reduces similarity to training data with minimal impact on audio quality, and that its behavior remains stable across dataset scales and application schedules. Future work includes deepening the theoretical understanding of CFG-based risk indicators, extending RG toward more adaptive or content-aware variants, and improving computational efficiency through strategies such as shared-trajectory sampling.

 doi:10.20944/preprints202603.0982.v1

Author Contributions: Conceptualization, T.K.; methodology, T.K.; software, T.K. and H.L.; validation, M.-J.K. and C.W.A.; formal analysis, T.K. and H.L.; investigation, T.K. and H.L.; resources, C.W.A.; data curation, T.K.; writing—original draft preparation, T.K.; writing—review and editing, H.L., M.-J.K. and C.W.A.; visualization, T.K.; supervision, M.-J.K.; project administration, C.W.A.; funding acquisition, M.-J.K. and C.W.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25398164); the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01842, Artificial Intelligence Graduate School Program (GIST) & the Artificial Intelligence Convergence Innovation Human Resources Development (RS-2023-00256629)); and the ITRC (Information Technology Research Center) Support Program (IITP-2026-RS-2024-00437718).

Data Availability Statement: The data presented in this study are openly available in [4].

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RG	Repulsive Guidance
OG	Opposite Guidance
CFG	Classifier-Free Guidance
FD	Fréchet Distance
FD_{pann}	Fréchet Distance computed using PANNs embeddings
FD_{vgg}	Fréchet Distance computed using VGGish embeddings
SIM_{Avg}	Average cosine similarity to training samples
SIM_{90}	Proportion of samples with cosine similarity > 0.9
DDPM	Denoising Diffusion Probabilistic Model
CLAP	Contrastive Language-Audio Pretraining
PANNs	Pretrained Audio Neural Networks
LDM	Latent Diffusion Model

References

1. Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; Défossez, A. Simple and Controllable Music Generation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
2. Agostinelli, A.; Denk, T.I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. MusicLM: Generating Music from Text. *arXiv preprint arXiv:2301.11325* **2023**.
3. Schneider, S.; Clark, P.; Lei, B.; Tyschenko, S. Moûsai: Efficient Text-to-Music Diffusion Models. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2024.
4. Melechovsky, J.; Guo, Z.; Ghosal, D.; Majumder, N.; Herremans, D.; Poria, S. Mustango: Toward Controllable Text-to-Music Generation. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2024, pp. 8286–8309.
5. Bralios, D.; Wichern, G.; Germain, F.G.; Pan, Z.; Khurana, S.; Hori, C.; Le Roux, J. Generation or Replication: Auscultating Audio Latent Diffusion Models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 1156–1160.
6. Batlle-Roca, R.; Liao, W.H.; Serra, X.; Mitsufuji, Y.; Gómez, E. Towards Assessing Data Replication in Music Generation With Music Similarity Metrics on Raw Audio. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 2024, pp. 1004–1011.
7. Barnett, J.; Garcia, H.F.; Pardo, B. Exploring Musical Roots: Applying Audio Embeddings to Empower Influence Attribution for a Generative Music Model. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 2024, pp. 360–368.

 doi:10.20944/preprints202603.0982.v1

8. Chen, K.; Wu, Y.; Liu, H.; Nezhurina, M.; Berg-Kirkpatrick, T.; Dubnov, S. MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 1206–1210.
9. Chen, C.; Liu, D.; Xu, C. Towards Memorization-Free Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8425–8434.
10. Jain, A.; Kobayashi, Y.; Shibuya, T.; Takida, Y.; Memon, N.; Togelius, J.; Mitsufuji, Y. Classifier-Free Guidance inside the Attraction Basin May Cause Memorization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 12871–12879.
11. Ho, J.; Salimans, T. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598* **2022**.
12. Somepalli, G.; Singla, V.; Goldblum, M.; Geiping, J.; Goldstein, T. Understanding and Mitigating Copying in Diffusion Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
13. Wen, Y.; Liu, Y.; Chen, C.; Lyu, L. Detecting, Explaining, and Mitigating Memorization in Diffusion Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
14. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.
15. Ghosal, D.; Majumder, N.; Mehrish, A.; Poria, S. Text-to-Audio Generation Using Instruction-Guided Latent Diffusion Model. In Proceedings of the ACM International Conference on Multimedia (ACM MM), 2023, pp. 3590–3598.
16. Kim, C.D.; Kim, B.; Lee, H.; Kim, G. AudioCaps: Generating Captions for Audios in the Wild. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 119–132.
17. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020.
18. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), 2019.
19. Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M.D. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2020**, *28*, 2880–2894.
20. Hershey, S.; Chaudhuri, S.; Ellis, D.P.W.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN Architectures for Large-Scale Audio Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 131–135.
21. Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; Dubnov, S. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.