AudioStealer: Extracting Audio Prompts via Shapley Value-Guided Query Search

Yingbin Jin[1*], **Xingjian Du**[2*], **Hanjun Luo**[3], **Zihao Wang**[4], **Haibo Hu**[1†], **Xiaofeng Wang**[4], **Xinfeng Li**[4†]
[1]Hong Kong Polytechnic University, [2]University of Rochester,
[3]New York University Abu Dhabi, [4]Nanyang Technological University

Abstract

As text-to-music models gain widespread adoption, the prompts used to guide these systems have become valuable intellectual property. This shift has given rise to a new form of attack: prompt stealing, aiming to reconstruct the high-value prompts that guide music generation. However, unlike prior work in text and image generation, prompt stealing in text-to-music systems faces unique challenges due to the entangled and diffuse nature of semantic representations in audio, which complicates decoupling textual tokens from acoustic outputs. To address these challenges, we present AudioStealer, the first study of prompt inversion in the audio domain. AudioStealer operates via a two-stage black-box attack framework: first, a heuristic search guided by audio-language embeddings identifies initial candidates; then, these candidates are refined using a game-theoretic strategy based on Shapley value estimation to attribute semantic contributions. Our method requires no direct access to the target model and relies solely on a shadow model, making it applicable. Through extensive experiments, we demonstrate that AudioStealer recovers prompts with high textual consistency to the ground truth, while regenerated audio maintains strong perceptual similarity to target recordings. These results expose vulnerabilities in the text-to-audio market ecosystem and underscore the need for intellectual property protections in generative audio technologies. Our code and dataset are openly available at https://github.com/kprisoner/AudioStealer.

1 Introduction

The rapid advancements in generative models have demonstrated unprecedented capabilities in high-quality multimedia content generation, attracting attention across various domains (OpenAI, 2023; Rombach et al., 2022; Agostinelli et al., 2023). Among these, text-to-music models such as Stable Audio (Evans et al., 2024), Suno (Suno, 2024), and MusicLM (Agostinelli et al., 2023) have introduced transformative changes to tasks like music composition and audio content creation by translating text-based prompts into sophisticated musical outputs. While these models lower the barriers to music creation and provide professional musicians with powerful tools, crafting effective prompts remains a complex and iterative process. Users often engage in extensive trial-and-error and the optimization can be both time-consuming and resource-intensive (Yuan et al., 2024; Sahoo et al., 2024; Schulhoff et al., 2024).

Consequently, high-quality prompts have emerged as valuable digital assets, leading to the rise of "prompt engineers" who specialize in designing high-quality prompts for purchase, and specialized marketplaces like PromptBase (PromptBase, 2025a) and Prompt AI (Prompt AI, 2025). The underlying business model is straightforward: customers browse sample generated audio clips, and they can purchase their associated prompt, which can then be adapted to generate similar music. On PromptBase, the top 50 sellers alone sold approximately 45K prompts in 2024, generating around $200K in revenue (PromptBase, 2025b). This underscores that these prompts are no longer just instructions, but significant intellectual property with substantial commercial value.

However, this growing commercial value has invited a new form of intellectual property theft: prompt stealing attacks. In such attacks, an adversary analyzes a music recording generated by a text-to-music model to infer its original prompt that can be reused to regenerate audio with similar style and structure. Unlike benign music captioning or description tasks, prompt stealing aims to recover prompts that are functionally effective for regeneration, enabling attackers to bypass the cost

*Equal contribution. 25053992r@connect.polyu.hk, diggerdu97@gmail.com
†Corresponding Author: haibo.hu@polyu.edu.hk, lxfmakeit@gmail.com

Findings of the Association for Computational Linguistics: ACL 2026, pages 14052–14067

of prompt engineering and replicate high-quality outputs at negligible effort. This raises serious concerns over the infringement of prompt creators' intellectual property rights and poses a critical threat to the integrity of text-to-music platforms' content ecosystems.

While prompt stealing has been explored in the field of text generation and image generation, the text-to-audio domain remains unstudied despite its growing adoption. Existing techniques cannot be directly migrated to the audio domain due to fundamental disparities in representation and architectural localization. Audio signals present unique challenges: concepts are not discrete tokens but continuous signals, and musical features are diffusely distributed and deeply entangled across the temporal-frequency domain.

To address these challenges, we propose **`AudioStealer`**, a two-stage prompt inversion framework designed to reconstruct plausible prompts from audio generated by a black-box text-to-music model. In the first stage, AudioStealer conducts a similarity-guided heuristic search that leverages contrastive audio-language embeddings to identify a high-quality candidate from an initial caption generated by a music-tagging model. In the second stage, to tackle semantic entanglement, we introduce a game-theoretic refinement strategy based on Shapley value (Shapley, 1952; Lundberg and Lee, 2017) estimation. This refinement stage quantifies the marginal contribution of individual semantic components and guides iterative prompt optimization.

Throughout both stages, AudioStealer relies exclusively on a local shadow model for querying, ensuring that the attack remains independent of the target generator's internal parameters. This design reflects realistic black-box threat scenarios in which direct access to the target model is unavailable or restricted. Through extensive experiments, we demonstrate that AudioStealer recovers prompts that are semantically faithful to the original inputs and capable of generating perceptually aligned audio, revealing a previously underexplored vulnerability in text-to-music systems.

Our key contributions are outlined below:

❶ ***A diverse prompt stealing benchmark dataset from leading text-to-music models.*** We introduce the first large-scale benchmark dataset for prompt stealing attacks, comprising audio-prompt pairs sampled from **four distinct text-to-music models**. By incorporating varied and leading model architectures, our dataset provides a challenging and realistic testbed for systematically evaluating and comparing the effectiveness and transferability of attack strategies.

❷ ***Proposal of a black-box prompt stealing attack.*** We propose a black-box attack method that combines music tagging-based initialization with similarity-guided heuristic search and refinement. By extracting conceptual tags from generated audio as an initial basis, the method employs heuristic search to iteratively optimize candidate prompts, enabling practical reconstruction of original prompts.

❸ ***Introduction of a Shapley value-based interpretability framework.*** To address the challenges of concept entanglement in audio, we introduce a Shapley value framework that quantitatively attributes the impact of each prompt component to audio similarity. This framework serves as a principled mechanism for prompt refinement.

2 Related Work

Prompt stealing, as a form of model inversion attack, has seen significant progress in other modalities, yet its application to audio remains nascent.

Prompt Stealing in the Text Domain. Sha & Zhang (Sha and Zhang, 2024) study prompt stealing against large language models, framing it as a classification or pattern-matching problem over discrete output tokens. This approach leverages the structured and discrete nature of language. Recent studies have developed optimized prompt leaking attacks (Hui et al., 2024) and comprehensive benchmarks (Wang et al., 2024) to evaluate susceptibility in LLM-integrated applications. However, music is a continuous and unstructured signal where concepts do not follow predictable patterns, making classification-based extraction significantly less effective in the audio domain.

Prompt Stealing in the Image Domain. Shen et al. (Shen et al., 2024) propose the first prompt stealing attack against text-to-image models, using an image captioning model to identify prompt subjects and a multi-label classifier to detect modifiers, progressively reconstructing the original prompt. Naseh et al. (Naseh et al., 2024) reframe the problem as image reproduction: they fine-tune a CLIP model on Midjourney-style data to extract keywords, train a multi-label classifier to detect image modifiers, and use GPT-4V to iteratively generate functionally equivalent prompts. Both approaches

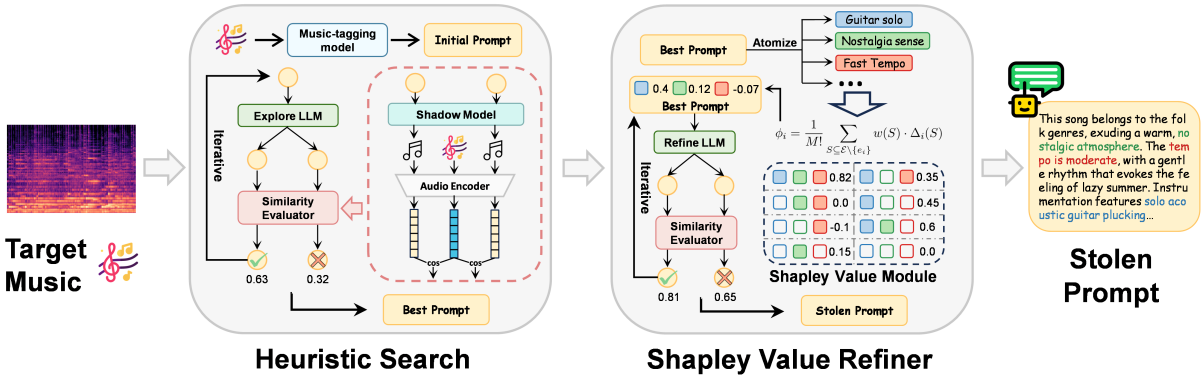

Figure 1: **AudioStealer** is a two-stage, black-box framework that integrates similarity-guided heuristic search for global exploration with Shapley value-guided refinement for precise semantic disentanglement.

rely on localized visual features specific to the image modality: the former depends on image-domain captioning and visual classification, while the latter relies on image-domain visual classifiers. Neither can be directly transferred to audio, where semantic features are diffusely distributed across the time-frequency domain rather than localized in any spatial structure.

Bridging to the Audio Domain. To verify the limitations of image-based methods, we adapt Shen et al.'s framework as the *Adapted T2I-Stealer* baseline in our experiments. Specifically, we preserve their overall "subject + modifier" structure but replace the subject generator with a music tagging system and the image-domain modifier detector with a music retrieval task. This adaptation serves to demonstrate that even when the T2I approach is maximally adapted, the absence of image-domain visual grounding and the presence of deep semantic entanglement in audio lead to substantially inferior performance compared to AudioStealer.

In summary, AudioStealer represents the first systematic study extending prompt stealing to audio domain and demonstrates for the first time the technical feasibility of audio prompt stealing.

3 Methodology

Our objective is to reconstruct a plausible input prompt that could have generated a given audio clip using a text-to-audio synthesis system. This task can be viewed as a form of prompt inversion (Fang et al., 2024), where the model's output is known, but the input that produced it is not. The challenge lies in performing this inversion without requiring extensive queries to the target model. This constraint is particularly critical in practical scenarios, where query access may be costly, rate-limited, or subject to surveillance.

Formally, let $\mathcal{P}$ denote the space of all possible natural language prompts and $\mathcal{X} \subset \mathbb{R}^L$ represents the domain of discrete audio signals of length L. A text-to-music generator is defined as a mapping function $G\colon \mathcal{P} \to \mathcal{X}$. Given a target audio sample $x_{tar} \in \mathcal{X}$, our objective is to find an optimal prompt $\hat{p} \in \mathcal{P}$ such that the synthesized audio $G(\hat{p})$ maximizes a similarity metric relative to x_{tar}.

3.1 AudioStealer Overview

The architecture of AudioStealer is designed to navigate the high-dimensional, continuous search space of audio prompts by addressing two primary technical barriers: the non-localization of acoustic features and the non-linear entanglement of semantic modifiers. Unlike text or image modalities, where concepts often occupy discrete tokens or spatial regions, audio concepts are diffusely distributed across the temporal-frequency domain. Consequently, AudioStealer formulates prompt recovery as an approximate inversion problem under a black-box similarity objective, rather than relying on static classification or independent modifier-decoupling strategies.

The core intuition of our two-stage approach lies in a "coarse-to-fine" optimization strategy. First, to overcome the challenge of non-localization, we employ a global heuristic search to anchor the semantic direction within the vast search space. Second, to address the non-linear entanglement where multiple modifiers interweave to shape the spectral structure, we introduce a game-theoretic attribution mechanism. This allows us to mathematically isolate and quantify the individual impact of each prompt element on the generated audio.

As illustrated in Figure 1, the AudioStealer pipeline consists of the following phases:

❶ ***Exploration Stage.*** A comprehensive search pro-

cedure inspired by Tree-of-Attacks with Pruning (TAP) (Mehrotra et al., 2024). This stage uses contrastive embeddings to guide the search towards a high-quality prompt candidate, identifying a stable semantic anchor.

❷ ***Exploitation Stage.*** A refinement phase that optimizes the candidate by leveraging Shapley value estimation. This allows us to quantify the marginal contribution of each prompt element to the overall audio alignment. The analysis then guides an LLM-based controller to iteratively improve the prompt.

To ensure the framework remains independent of the target generator's internal parameters, we query only a local shadow model G_{shadow} using music prompts as a proxy for evaluating alignment during both exploration and exploitation. The overall search process is depicted in Appendix A.1.

3.2 Exploration: TAP-like Heuristic Search

We begin the exploration process with an initial prompt $p_0 \in \mathcal{P}$ generated by OPENJMLA (Du et al., 2023b), a zero-shot music-tagging model based on Contrastive Language-Audio Pretraining (CLAP) (Wu et al., 2023; Chen et al., 2022). Although this caption may be coarse or imprecise, it typically captures salient musical attributes such as instrumentation, rhythm, and genre. Examples include phrases like "funk bass line" or "syncopated drums". These semantic anchors provide a valuable starting point that significantly narrows the otherwise vast and unstructured prompt space.

To effectively explore the space of prompt modifications, we adopt a strategy inspired by Tree-of-Attacks with Pruning (TAP), which is implemented as an iterative beam search. At each iteration k, we maintain a "beam" of the top-B most promising prompts, denoted $\{p_k^{(1)}, \ldots, p_k^{(B)}\}$, where B is the beam width. Each prompt $p_k^{(b)}$ is submitted to a large language model (a locally deployed LLM, we use Qwen3-14B (Yang et al., 2025)) together with a history of previously successful prompts. This guides the LLM to learn from past iterations and adapt its generation strategy to generate K creative variations, or "branches".

This generation step produces a new set of $B \times K$ candidate prompts. Each candidate is then used to synthesize an audio sample via $G(p)$, and its quality is evaluated by computing its similarity to the target audio x_{tar}. To formally define the similarity, let $E : \mathcal{X} \to \mathbb{R}^d$ be a pre-trained CLAP audio encoder that maps an audio signal to a d-dimensional embedding space. The similarity function $\text{sim}(\cdot, \cdot)$ is defined as the cosine similarity between the corresponding embeddings $\mathbf{e}_p = E(G(p))$ and $\mathbf{e}_{tar} = E(x_{tar})$:

$$\text{sim}(\mathbf{x}_{\text{tar}}, G(p)) = \frac{\mathbf{e}_p^\top \mathbf{e}_{\text{tar}}}{\|\mathbf{e}_p\|_2 \|\mathbf{e}_{\text{tar}}\|_2}. \quad (1)$$

This measure provides a continuous proxy for semantic and perceptual closeness, which is essential for guiding the search process.

For the selection step, all newly generated $B \times K$ candidates are ranked by their similarity scores, and the top-B prompts are selected to form the beam for the next iteration. This selection mechanism maintains a balance between exploring new variations and avoiding drifting away from promising regions of the prompt space.

Exploration terminates when one of the following conditions is met: (i) the similarity score exceeds a threshold t_1, indicating that the prompt is already closely aligned with the target audio; or (ii) the maximum number of search rounds N_1 is reached. This stage prioritizes broad exploration while laying the groundwork for the more fine-grained optimization that follows.

3.3 Exploitation: Shapley Value Refiner

The heuristic search stage yields a reasonably strong prompt p^{best}, which serves as the basis for the exploitation stage. To enable finer-grained control, we first employ an LLM to decompose p^{best} from a single string into a structured dictionary of atomic semantic units, or *elements*, denoted by the set $\mathcal{E} = \{e_1, \ldots, e_M\}$. These elements represent complete musical concepts and are organized into four distinct categories: (A) *Style & Genre*, (B) *Mood & Atmosphere*, (C) *Instrumentation & Timbre*, (D) *Structure & Tempo*. The set $\mathcal{E}$ defines a closed world for the refinement process, the goal is to identify the optimal subset of elements $S \subseteq \mathcal{E}$ that maximizes alignment, rather than introducing new semantic information.

To quantify the individual contribution of each element to the interwoven audio output, we adopt a game-theoretic approach based on Shapley values. For a given element $e_i \in \mathcal{E}$, its importance ϕ_i is defined as the average marginal gain across all possible subsets S that exclude e_i:

$$\phi_i = \frac{1}{M!} \sum_{S \subseteq \mathcal{E} \setminus \{e_i\}} w(S) \cdot \Delta_i(S), \quad (2)$$

where $w(S) = |S|!(M - |S| - 1)!$ is a combinatorial weight and the marginal gain $\Delta_i(S)$ is given by:

$$\Delta_i(S) = f(S \cup \{e_i\}) - f(S), \tag{3}$$

with $f(S) = \text{sim}(x_{tar}, G(\texttt{join}(S)))$ denoting the similarity between the target audio and the audio generated by joining elements in S into a prompt. Consistent with the exploration stage, the similarity is computed using the CLAP-based cosine distance between audio embeddings.

Exact computation of Equations (2) and (3) is intractable when the number of elements M is large, due to the exponential number of subsets. We therefore approximate ϕ_i using T Monte Carlo samples (Castro et al., 2009), where each sample corresponds to a random permutation of the element list. We initialize a structured memory matrix, denoted $\mathbf{A}^{(r)} \in \{0, 1\}^{T \times (M+1)}$. Each row i in $\mathbf{A}^{(r)}$ represents a specific combination, where an entry $\mathbf{A}^{(r)}_{i,j} = 1$ indicates the presence of element e_j, and the final column $\mathbf{A}^{(r)}_{i,M+1}$ records the resulting similarity score $\text{sim}(x_{tar}, G(p_{r,i}))$.

During each refinement iteration, the refine LLM receives a comprehensive context: the current p^{best}, its corresponding element set $\mathcal{E}$, estimated Shapley values $\{\phi_i\}$ and the memory matrix $\mathbf{A}^{(r)}$ as input. Based on this context, the LLM is instructed to propose a batch of new element combinations predicted to yield higher scores. This constraint encourages the model to reason explicitly about semantic synergy and redundancy. Once a revised prompt p' is generated, it would be evaluated, and the results are used to update the memory matrix $\mathbf{A}^{(r+1)}$. The global p^{best} is updated only upon finding a strictly higher score. This iterative refinement proceeds until the score exceeds threshold t_2 or reaches the maximum rounds N_2.

This two-stage approach balances breadth and precision. The heuristic search stage is effective for reaching a good initial approximation of the target. The Shapley-guided refinement stage, although more computationally intensive, provides targeted improvements by trimming uninformative content and reweighting key semantic components. Together, they mimic a human composer's workflow: starting with a high-level stylistic sketch, then gradually refining instrumentation, mood, and structure to achieve a polished final result. Further details are provided in Appendix A.

4 Prompt2Music Dataset

To systematically investigate and benchmark the vulnerabilities of text-to-music models to prompt stealing attacks, a specialized dataset is required. However, existing publicly available datasets, such as MusicCaps (Agostinelli et al., 2023), LP-MusicCaps (Doh et al., 2023), and MusicBench (Melechovsky et al., 2024), primarily offer audio-caption pairs. While recent prompt-to-music dataset, Wikimt-x (Wu et al., 2025) represents an advancement, it typically provides a one-to-one mapping, linking each prompt to a single audio output from one model. This structure is insufficient for a comparative benchmark, as it cannot be used to evaluate how the performance of a prompt stealing attack varies across different model architectures. Consequently, a dataset built for the specific purpose of cross-model comparison is needed.

To address this gap, we constructed a new dataset, **`Prompt2Music`**, tailored for prompt inversion research. The construction involved a multi-step data engineering pipeline. First, we designed a generation template based on established practices for music prompt engineering (Suno Prompts, 2025; Altorise Hub, 2025; Soundverse, 2025) to emulate high-efficiency prompts used in real-world scenarios. Each prompt was required to include four key components: (1) *Style & Genre*, (2) *Instrumentation & Timbre*, (3) *Mood, Atmosphere & Scene*, and (4) *Structure, Tempo & Dynamics*.

We invited music professionals to manually design 100 high-quality "master" prompts following this template. Using these as representative samples, we leveraged GPT-4o to programmatically expand the corpus to 5,000 diverse music prompts. A quality control step was applied in the process to filter out repetitive or overly simplistic prompts, ensuring the diversity of the final set. We then sampled the generated prompts and presented them to music professionals to audit the quality.

Subsequently, each of these structured prompts was fed into four distinct text-to-music models to synthesize the corresponding audio: Stable Audio Open (Evans et al., 2024), MusicGen (Copet et al., 2024), InspireMusic (Zhang et al., 2025), and ACE-Step (Gong et al., 2025). A fixed seed was used across all generation processes to ensure experimental reproducibility. Given the generally limited vocal generation capabilities of current open-source music models, all prompts in our dataset

were designed to be instrumental, excluding any vocal elements. Additionally, constrained by generation time and model capabilities, the duration of each audio clip was uniformly set to 30 seconds.

The final Prompt2Music dataset comprises 5,000 tuples, resulting in a total of 20,000 audio files. T-SNE visualization (Appendix B.2) confirms high semantic diversity across multiple musical dimensions. Each tuple contains a detailed source prompt and the four resulting audio clips generated by the different models. This one-to-many mapping from a single prompt to outputs from various models provides a robust foundation for comparative analysis. It serves as a critical resource for developing and benchmarking prompt stealing attack methodologies across diverse model architectures. Further details on the data construction and examples are provided in Appendix B.1.

5 Experiments

5.1 Experimental Settings

Dataset. All primary experiments in this section are conducted on a subset of our newly constructed Prompt2Music Dataset. For each experimental run, we randomly sample 100 prompt-audio pairs. The original, high-quality text prompt serves as the ground truth, which our attack framework and baseline methods aim to reconstruct by analyzing the corresponding generated audio. To assess the generalization capabilities, we also curated a prompt collection from three widely-used music caption datasets: MusicCaps, LP-MusicCaps, MusicBench, and one music prompt dataset, Wikimt-x. We randomly sampled 100 entries from each, using their provided caption/description fields. These 400 text snippets serve as the basis for our generalization study, testing the methods against prompts of varied styles and origins. Further details on the test datasets are in Appendix B.3.

Baseline Methods. We compare AudioStealer against four representative baselines that cover distinct technical approaches for prompt inversion.

- ***Zero-shot Tagger (OpenJMLA).*** We use the state-of-the-art music tagging system, OPEN-JMLA (Du et al., 2023b). Given a target audio, the model generates a set of descriptive tags, which are concatenated to form the prompt.
- ***Adapted T2I-Stealer.*** To evaluate the direct transferability of text-to-image prompt stealing method (Shen et al., 2024), we implement the Adapted T2I-Stealer including a subject generator(OpenJMLA) and a modifier detector based on CLAP. It simulates a multi-classifier through a music retrieval task using MTG-JAMENDO tag set (Bogdanov et al., 2019), appending identified tags to the subject description to form the prompt. Details are provided in Appendix C.4.
- ***Music Understanding LLM (Mu-LLaMA).*** As a specialized baseline, we employ Mu-LLaMA (Liu et al., 2023). We provide it with the target audio and the instruction, "Describe this piece of music with a detailed prompt that a text-to-music model could use to generate it." Its response serves as the reconstructed prompt.
- ***General Multimodal LLM (Qwen2-Audio).*** To represent powerful general-purpose models, we use Qwen2-Audio (Chu et al., 2024). It receives the same audio and instruction as Mu-LLaMA to infer the source prompt.

Evaluation Metric. We evaluate the performance of all methods from multiple perspectives, assessing both the reconstructed prompt's fidelity and the auditory quality of the music it generates.

- ***Semantic Similarity.*** This metric quantifies the textual correspondence between the reconstructed prompt and the original ground-truth prompt. We use all-MiniLM-L6-v2's text encoder to generate the embeddings and compute the cosine similarity between the vector embeddings of the target and stolen prompts. (Wang et al., 2020; Reimers and Gurevych, 2019)
- ***Music Similarity.*** The music similarity is the cosine similarity between the embeddings of the target and the stolen music, which is a widely adopted metric to measure similarity between audios. (Barnett et al., 2024) We rely on CLAP's audio encoder to obtain an audio's embedding, in order to gauge perceptual and acoustic similarity. (Elizalde et al., 2024; Xiao et al., 2024)
- ***ByteCover3 Similarity.*** Since an attack should generate a new "version" of the target music, a high-quality Cover Song Identification (CSI) model provides an objective measure of this target. We calculate the metric with the state-of-the-art CSI model, ByteCover3 (Du et al., 2023a).

Acknowledging the limitations inherent in any automatic metric, we also conducted a human evaluation study to perform a comprehensive perceptual assessment. The human-rated similarity refers to the perceived similarity between target and stolen audios by end-users. Specifically, for

each target music and its corresponding stolen audio, ten domain experts are assigned to label it using a 5-level Likert-scale, ranging from "not similar at all" to "very similar." The detailed criteria for each level are stated in Appendix C.2. We randomly sample twenty-five pieces of music in each evaluation and report the mean value.

Models Setting. The performance and behavior of our two-stage framework are governed by several key hyperparameters. Theoretically, employing larger values for beam width B, branch factor K, and the number of rounds N_1, N_2, along with a higher Monte-Carlo budget T and no early stopping would likely yield superior results through a more exhaustive search. However, to maintain practical applicability, our default settings are chosen to strike a balance between effectiveness and resource usage. A detailed efficiency analysis is provided in Appendix D. In the following experiments, for the Exploration Stage, we set $B = 3$, $K = 4$, $t_1 = 0.7$ and $N_1 = 10$. For the Exploitation Stage, we set $T = 30$, $t_2 = 0.8$ and $N_2 = 5$. To ensure statistical significance, we executed each experiment using three different random seeds and reported the average values. All experiments were conducted on a server with Ubuntu 22.04 (CPU: AMD EPYC 7763, GPU: NVIDIA A100-80G). More details on the experimental settings are in Appendix C.

5.2 Main Results

Target	Method	Sem	Mus	BC3	Human
Inspire	OpenJMLA	42.7	64.3	24.4	2.56
	Adapted T2I	43.1	63.6	24.1	2.42
	Mu-LLaMA	40.7	61.1	23.1	2.48
	Qwen2-Audio	44.2	63.2	25.1	2.52
	AudioStealer	**59.4**	**70.1**	**27.8**	**3.32**
ACE-Step	OpenJMLA	44.2	65.1	21.5	2.38
	Adapted T2I	46.1	66.0	21.6	2.46
	Mu-LLaMA	45.8	64.6	21.8	2.62
	Qwen2-Audio	54.2	67.8	23.7	2.95
	AudioStealer	**59.0**	**69.4**	**25.9**	**3.28**
MusicGen	OpenJMLA	47.1	50.3	14.3	2.25
	Adapted T2I	48.3	49.8	14.8	2.32
	Mu-LLaMA	45.7	47.4	16.0	2.14
	Qwen2-Audio	53.7	55.0	17.1	2.45
	AudioStealer	**60.1**	**61.1**	**20.3**	**2.66**

(1) **Sem**(%): Semantic Similarity; **Mus**(%): Music Similarity; **BC3**(%): ByteCover3 Similarity.
(2) **Human**: Human-rated Similarity on 5-point Likert scale.

Table 1: Performance comparison across different target models and methods

Table 1 compares our method with various baselines against target models. As shown, while the performance exhibits some variance across the different targets, AudioStealer consistently demonstrates a robust performance advantage over the baselines on all four evaluation metrics. When InspireMusic serves as the target model, **AudioStealer achieves 59.4%, 70.1%, and 27.8%** for Semantic, Music and ByteCover3 similarity respectively. Thus, we selected it as the default target model for subsequent ablation studies. We also observe that all methods yield relatively lower Music Similarity scores on MusicGen (AudioStealer: 61.1% vs. 69.4%/70.1% on other models), which may be due to MusicGen's higher output diversity, making a direct acoustic match more challenging.

Notably, the ByteCover3 Similarity scores are numerically lower across the board. We attribute this to the non-linear nature of the metric: ByteCover3 utilizes a MaxMean matching mechanism to evaluate deep-level structural and melodic correspondence. Consequently, its scoring distribution operates on a different scale that is not directly comparable to other cosine similarity-based metrics.

Interestingly, the Adapted T2I-Stealer occasionally underperforms the standalone OpenJMLA, highlighting the inherent limitations of directly image frameworks to audio domain. This suggests that simply appending retrieved modifiers can exacerbate semantic confusion rather than providing guidance, as the entangled nature of audio concepts resists independent tag reconstruction.

Public dataset. We utilized four representative music-caption/prompt datasets to assess the model's performance in more realistic and diverse scenarios. As shown in Table 2, AudioStealer achieves the highest Music Similarity (69.0-70.3%) and ByteCover3 Similarity (27.7-28.6%) across all four datasets. However, performance on Semantic Similarity shows interesting dataset-dependent patterns. AudioStealer achieves notably superior Semantic Similarity on Wikimt-x, but falls slightly on LP-MusicCaps and MusicBench. This variance directly reflects the fundamental distinction between prompts (Wikimt-x, Prompt2Music) and descriptive captions (MusicCaps, LP-MusicCaps, MusicBench). The core target of prompt stealing attacks is to generate prompts that maximize audio regeneration fidelity, not textual paraphrasing. Critically, on caption-based dataset, AudioStealer still generates more acoustically faithful audio than all baselines, demonstrating that our method prioritizes perceptual fidelity over literal text matching.

Dataset	Method	Sem	Mus	BC3	Human
MusicCaps	OpenJMLA	42.3	66.4	24.7	2.82
	Adapted T2I	45.1	66.8	23.6	2.78
	Mu-LLaMA	41.1	63.3	23.0	2.56
	Qwen2-Audio	50.9	64.4	23.0	2.60
	AudioStealer	**54.6**	**70.2**	**27.7**	**3.26**
LP-Music	OpenJMLA	51.5	64.8	24.8	2.46
	Adapted T2I	49.1	65.9	24.5	2.50
	Mu-LLaMA	46.8	61.9	23.3	2.44
	Qwen2-Audio	**57.4**	63.6	23.4	2.38
	AudioStealer	55.7	**69.0**	**27.9**	**3.08**
MusicBench	OpenJMLA	50.4	64.6	24.3	2.56
	Adapted T2I	50.2	65.0	24.9	2.54
	Mu-LLaMA	45.1	62.4	23.9	2.30
	Qwen2-Audio	**55.3**	64.9	25.2	2.68
	AudioStealer	53.7	**70.3**	**28.4**	**3.25**
Wikimt-x	OpenJMLA	57.0	64.2	26.8	2.68
	Adapted T2I	56.4	64.6	26.9	2.65
	Mu-LLaMA	48.5	63.0	25.2	2.56
	Qwen2-Audio	57.1	64.2	26.1	2.44
	AudioStealer	**63.8**	**69.8**	**28.6**	**3.18**

(1) **Sem**, **Mus**, **BC3** in percentage (%)

Table 2: Performance comparison across public datasets

Furthermore, the human evaluation provides an important complementary perspective to the automatic metrics. Across both model-specific and public datasets, AudioStealer consistently achieves higher Human Similarity scores, indicating that its regenerated audio is perceived by listeners as more similar to the target music in overall style, structure, and musical coherence. For completeness, we report the standard deviation in Appendix C.3.

5.3 Ablation Studies

We conducted a series of ablation studies to analyze the contribution of each component within our framework. Specifically, we assessed the performance of the Exploration Stage and the Exploitation Stage in isolation, and evaluated the sensitivity of AudioStealer to different choices for its core components: the shadow model, the audio encoder, and the basis model.

Table 3 shows the individual outputs of two stages in AudioStealer. Stage 1 alone achieves a high Music similarity of 68.9% as a representative black-box optimization. However, its lower Semantic similarity (53.1%) suggests that a pure search process, while acoustically effective, tends to produce prompts that are semantically cluttered or imprecise. Conversely, Stage 2 alone with the initial prompt struggles to improve upon the simple initial prompt, resulting in weaker performance. These findings highlight that the two stages are complementary and indispensable, demonstrating the synergy between exploration and refinement.

Method	Sem(%)	Mus(%)	BC3(%)
Stage1 Only	53.1	68.9	26.7
Stage2 Only	45.5	65.8	25.9
AudioStealer	**59.4**	**70.1**	**27.8**

Table 3: Ablation study results

We further evaluate the robustness of AudioStealer across various architectural choices for its core components, summarized in Figure 2.

- ***A) Shadow Model Comparison.*** Figure reveals interesting transferability patterns across different shadow models. When the shadow model is identical to the target (InspireMusic), AudioStealer achieves the upper bound performance (Sem: 63.8%, Mus: 77.6%, BC3: 36.4%). Notably, when employing models with different architectures as proxies, such as MusicGen and ACE-Step, the attack maintains competitive performance with only marginal degradation. This finding supports the viability of AudioStealer in practical black-box scenarios.
- ***B) Audio encoder Comparison.*** We consider four audio encoders in our experiments. CLAP achieves the best performance (Sem: 59.4%, Mus: 70.1%, BC3: 28.3%), followed closely by Wav2CLIP. This near-identical performance can be attributed to their shared core structure: inspired by CLIP contrastive learning framework. In contrast, MERT shows degraded performance, which may be because of its focus on intra-modal music understanding tasks.
- ***C) Basis model Comparison.*** We compared the effectiveness of three different basis models for generating the initial prompt: Mu-LLaMA, Qwen2-Audio, and OpenJMLA. The attack with a prompt from Qwen2-Audio (Sem: 59.1%, Mus: 70.6%, BC3: 28.3%) yields results nearly on par with using OpenJMLA (Sem: 59.4%, Mus: 70.7%, BC3: 27.8%). Meanwhile, using Mu-LLaMA as the basis model leads to a bit drop in performance. This suggests that the two-stage refinement framework can effectively recover from suboptimal initial prompts, though a better starting point does provide a modest advantage.

6 Conclusion

This work presents **AudioStealer**, the first systematic framework for prompt stealing in text-to-

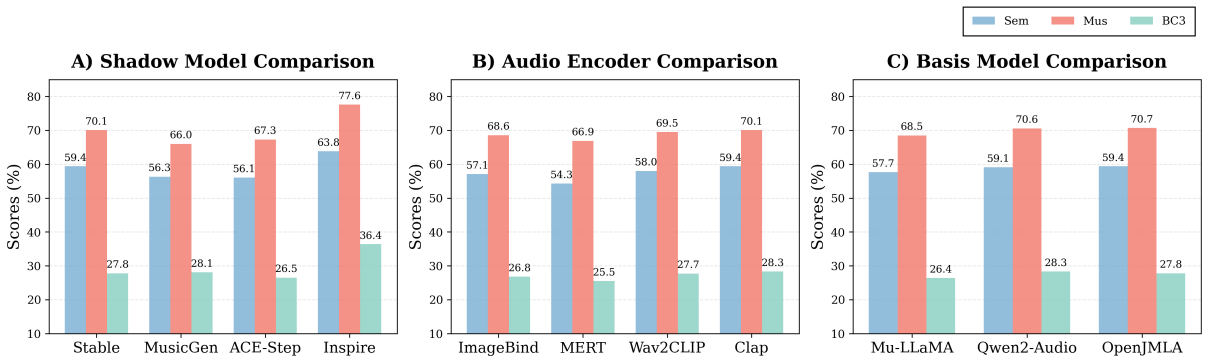

Figure 2: Component ablation analysis. We evaluate the impact of (A) different shadow models for generating audio, (B) different audio encoders for computing similarity, and (C) different basis models for producing initial prompts.

music models. By addressing the unique challenges of semantic entanglement and concept dispersion in audio, we demonstrate that it is possible to recover high-quality prompts using only black-box access. Our two-stage approach combines heuristic search with a principled Shapley value-based refinement strategy, enabling effective and interpretable prompt reconstruction. Experimental results confirm the practicality and accuracy of our method, revealing a significant and previously unexplored security risk in generative audio systems. These findings call for immediate attention to intellectual property protection and robust defense mechanisms in text-to-audio generation platforms.

Limitations

While our study focuses on English-language prompts and predominantly Western musical styles, it does not examine culturally diverse genres, instruments. Nevertheless, we believe the core attack mechanisms, contrastive audio-language retrieval and Shapley value-guided refinement, are broadly applicable across languages and musical types, assuming the availability of appropriate embedding spaces. Constructing multilingual and multicultural datasets, and evaluating the attack's effectiveness in underrepresented audio domains, remains an important direction for future research.

Additionally, although we evaluate reconstruction fidelity and semantic accuracy, our work does not fully explore the broader societal and legal implications of prompt stealing, including questions of copyright infringement, fair use, and the economic impact on emerging prompt marketplaces. We emphasize that technical findings must be contextualized through interdisciplinary dialogue involving policymakers, content creators, and platform providers to establish fair and enforceable norms. An investigation of regulatory frameworks and responsible disclosure practices is beyond the scope of this paper and left for future work.

Ethical Considerations

For the human evaluation experiments involving participants, we strictly adhered to all ethical guidelines. All participants were clearly informed of the research purpose, procedures, potential task difficulty, and data usage; they provided signed informed consent and were notified of their right to withdraw at any time. To protect privacy, all participant data has been fully anonymized. We provided fair compensation for their time and expertise. All study procedures were reviewed and approved by the Institutional Review Board (IRB) of the primary contributors' university.

Our study focuses on identifying security vulnerabilities in text-to-music systems to promote the development of robust IP protection mechanisms. The framework is designed as a black-box attack that relies on a local shadow model, minimizing direct interaction with and potential disruption of commercial platforms. We are committed to responsible disclosure and hope our findings encourage the community to establish fair and enforceable norms for generative audio technologies.

Acknowledgments

This work was supported by the Ministry of Science and Technology of the People's Republic of China (Grant No: 2025YFE0200100), the Research Grants Council (Grant No: 15209922 and 15210023), and the Innovation and Technology Fund (Grant No: ITS-140-23FP), Hong Kong SAR, China.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Altorise Hub. 2025. Best Prompts for AI Music Generators: Beginner to Advanced Level. https://altorise.com/hub/best-prompts-for-ai-music-generators/. Accessed: 2025-07-22.

Julia Barnett, Hugo Flores Garcia, and Bryan Pardo. 2024. Exploring musical roots: Applying audio embeddings to empower influence attribution for a generative music model. *arXiv preprint arXiv:2401.14542*.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the shapley value based on sampling. *Computers & operations research*, 36(5):1726–1730.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022*, pages 646–650. IEEE.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.

Xingjian Du, Zijie Wang, Xia Liang, Huidong Liang, Bilei Zhu, and Zejun Ma. 2023a. Bytecover3: Accurate cover song identification on short queries. *arXiv preprint arXiv:2303.11692*.

Xingjian Du, Zhesong Yu, Jiaju Lin, Bilei Zhu, and Qiuqiang Kong. 2023b. Joint music and language attention models for zero-shot music tagging. *arXiv preprint arXiv:2310.10159*.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2024. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340. IEEE.

Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. 2024. Fast timing-conditioned latent audio diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12652–12665.

Hao Fang, Yixiang Qiu, Hongyao Yu, Wenbo Yu, Jiawei Kong, Baoli Chong, Bin Chen, Xuan Wang, Shu-Tao Xia, and Ke Xu. 2024. Privacy leakage on dnns: A survey of model inversion attacks and defenses. *arXiv preprint arXiv:2402.04013*.

Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. 2025. Ace-step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*.

Bo Hui, Haolin Yuan, Neil Gong, Philippe Burlina, and Yinzhi Cao. 2024. Pleak: Prompt leaking attacks against large language model applications. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pages 3600–3614.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2023. Music understanding llama: Advancing text-to-music generation with question answering and captioning. *arXiv preprint arXiv:2308.11276*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of attacks: Jailbreaking black-box llms automatically. *arXiv preprint arXiv:2312.02119*.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*.

Ali Naseh, Katherine Thai, Mohit Iyyer, and Amir Houmansadr. 2024. Iteratively prompting multimodal llms to reproduce natural and ai-generated images. *CoRR*, abs/2404.13784.

OpenAI. 2023. ChatGPT. https://chat.openai.com/ (Accessed: 2025-06-05).

Prompt AI. 2025. Prompt AI. https://prompti.ai/. (Accessed: 2025-06-05).

PromptBase. 2025a. PromptBase. https://promptbase.com/. (Accessed: 2025-06-05).

PromptBase. 2025b. PromptBase Leaderboard. https://promptbase.com/leaderboard/(Accessed: 2025-06-05).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2024. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.

Zeyang Sha and Yang Zhang. 2024. Prompt stealing attacks against large language models. *arXiv preprint arXiv:2402.12959*.

Lloyd S. Shapley. 1952. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA.

Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. 2024. Prompt stealing attacks against {Text-to-Image} generation models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 5823–5840.

Soundverse. 2025. Best Prompts for Music Generator AI. https://www.soundverse.ai/blog/article/best-prompts-for-music-generator-ai. Accessed: 2025-07-22.

Suno. 2024. Introducing Suno v4. https://suno.com/blog/v4. Accessed: 2025-06-05.

Suno Prompts. 2025. How To Prompt Suno: Suno Prompts Guide. https://howtopromptsuno.com. Accessed: 2025-06-05.

Junlin Wang, Tianyi Yang, Roy Xie, and Bhuwan Dhingra. 2024. Raccoon: Prompt extraction benchmark of llm-integrated applications. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13349–13365.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Shangda Wu, Zhancheng Guo, Ruibin Yuan, Junyan Jiang, Seungheon Doh, Gus Xia, Juhan Nam, Xiaobing Li, Feng Yu, and Maosong Sun. 2025. Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages. *arXiv preprint arXiv:2502.10362*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Feiyang Xiao, Jian Guan, Qiaoxi Zhu, Xubo Liu, Wenbo Wang, Shuhan Qi, Kejia Zhang, Jianyuan Sun, and Wenwu Wang. 2024. A reference-free metric for language-queried audio source separation using contrastive language-audio pretraining. *arXiv preprint arXiv:2407.04936*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, Ziyang Ma, Liumeng Xue, Ziyu Wang, Qin Liu, Tianyu Zheng, Yizhi Li, Yinghao Ma, Yiming Liang, Xiaowei Chi, and 16 others. 2024. Chatmusician: Understanding and generating music intrinsically with llm. *arXiv preprint arXiv:2402.16153*.

Chong Zhang, Yukun Ma, Qian Chen, Wen Wang, Shengkui Zhao, Zexu Pan, Hao Wang, Chongjia Ni, Trung Hieu Nguyen, Kun Zhou, Yidi Jiang, Chaohong Tan, Zhifu Gao, Zhihao Du, and Bin Ma. 2025. Inspiremusic: Integrating super resolution and large language model for high-fidelity long-form music generation. *arXiv preprint arXiv:2503.00084*.

A Methodology Details

A.1 Implementation and Model Specifications

The overall search process of AudioStealer is depicted in Alg. 1.

OpenJMLA Setup. The model is run in evaluation mode. We utilize Llama-2-7b-hf as the tokenizer. All input audio files are first resampled to a 16kHz sampling rate. The resampled audio is then converted into a log-mel spectrogram. The spectrograms are normalized and then padded or cropped to a fixed length of 2,992 frames to serve as the model input.

Algorithm 1: AUDIOSTEALER: Two-Stage Prompt Extraction from a Text-to-Audio Generator

Input: Target audio x_{tar}; generator G_{shadow}; similarity metric sim; thresholds t_1, t_2; maxima N_1, N_2; beam width B; branch factor K; Monte-Carlo budget T
Output: Recovered prompt $\hat{p}$

// Stage 1: Exploration (TAP-like Heuristic Search);
$p_0 \leftarrow \text{OPENJMLA}(x_{\text{tar}})$
$p^{\text{best}} \leftarrow p_0,\ s^{\text{best}} \leftarrow \text{sim}(\mathbf{x}_{\text{tar}}, G_{\text{shadow}}(p_0))$;
Beam $\leftarrow \{p_0\}$;
for $k = 0, \ldots, N_1 - 1$ **do**
  Candidates $\leftarrow$ Expand(Beam, K, LLM);
  Beam $\leftarrow$ TopB(Candidates, B, sim);
  $p_{\text{current}}, s_{\text{current}} \leftarrow$ First(Beam);
  if $s_{current} > s^{\text{best}}$ **then**
    $p^{\text{best}}, s^{\text{best}} \leftarrow p_{\text{current}}, s_{\text{current}}$;
  end
  if $s^{\text{best}} \geq t_1$ **then break**;
end
// Stage 2: Exploitation (Shapley Value Refiner);
decompose p^{best} into element set $\mathcal{E} = \{e_1, \ldots, e_M\}$;
initialize memory matrix $\mathbf{A}^{(0)} \leftarrow \text{Initialize}(\mathcal{E}, T)$;
for $r = 0, \ldots, N_2 - 1$ **do**
  $\{\phi_i\}_{i=1}^{M} \leftarrow \text{EstimateShapley}(\mathbf{A}^{(r)})$;
  $\{p'\} \leftarrow \text{LLM.Refine}(\mathcal{E}, \{\phi_i\}, \mathbf{A}^{(r)}, p^{\text{best}})$
  $p'_{\text{best}}, s'_{\text{best}} \leftarrow \text{FindBest}(\{p'\}, \text{sim})$;
  update $\mathbf{A}^{(r+1)}$ **if** $s'_{best} > s^{\text{best}}$ **then**
    $p^{\text{best}}, s^{\text{best}} \leftarrow p'_{\text{best}}, s'_{\text{best}}$
  end
  if $s^{\text{best}} \geq t_2$ **then break**;
end
return $\hat{p} \leftarrow p^{\text{best}}$

Clap Setup. The CLAP model serves as the core audio encoder for calculating audio embeddings and similarity. The model is run in evaluation mode. The implementation uses the laion/clap-htsat-fused model as default setting. Input audio is loaded as a mono-channel signal using librosa and resampled to the model's required 48kHz sampling rate.

Stable Audio Open Setup. We use the stable-audio-open-1.0 model as the shadow model for generating audio from text prompts. The generated audio is normalized, rearranged, and saved as a 16-bit integer WAV file. The sample rate is 44.1kHz.

A.2 LLM Prompt Templates

The following are the specific LLM prompt templates designed for our attack framework. The Prompt Variation template facilitates the Exploration Stage. The Refinement Prompt and Decomposition Prompt are subsequently used during the Exploitation Stage to iteratively improve upon promising candidates.

Prompt Variation for Exploration

You are a creative music prompt optimizer. Your task is to generate variations of music prompts while maintaining core musical elements that lead to higher similarity scores.
Below are the prompts selected from previous iterations and their corresponding similarity scores (0-1 where 1 is perfect):
ITERATION HISTORY: `{iteration_history}`
CURRENT:
Prompt to optimize: `{current_prompt}`
Current Score: `{current_score}`
DYNAMIC STRATEGY GUIDANCE:
If stagnation is detected or score is lower than 0.4, make bold, creative changes... Otherwise, make targeted modifications...
Analyze the previous iterations to understand what elements lead to higher or lower similarity scores. Format your response in JSON with two elements:
- **Improvement**: A brief explanation of your prompt modification strategy based on the history and current score.
- **Prompt**: Your new variation of the music prompt.

Refinement Prompt for Exploitation

You are an expert Music Prompt Strategist. Your goal is to refine and optimize the current best prompt to discover one or more new combinations of elements that will achieve a higher score. The focus is on refinement, not complete overhaul.
CORE PRINCIPLES
1. You MUST and ONLY select elements from the Element Universe to construct new prompts. You are STRICTLY PROHIBITED from inventing, deriving, or introducing any elements not on this list.
2. The new combinations you generate MUST NOT be identical to any of the combinations listed in the Top Evaluations, Bottom Evaluations, or Recent Evaluations sections below.
3. Each new proposal in your output MUST contain at least one element from EACH of the four categories: ***Style & Genre***; ***Mood, Atmosphere & Scene***; ***Instrumentation & Timbre***; ***Structure, Tempo & Dynamics***.
Analyze all the information provided below. Then, strictly following the specified JSON format, return a JSON list containing `{variants_to_generate}` unique proposals that you believe have the highest probability of surpassing the current best score.
CURRENT BEST STATE
Current Best Prompt: `{best_prompt}`
Current Best Score: `{best_score}`
Elemental Composition: `{best_elements}`
ELEMENT UNIVERSE
Full Universe: `{element_universe}`
Contribution Values: `{shapley_values}`
HISTORICAL DATA
Top, Bottom Evaluations and Recent Evaluations.

Decomposition Prompt for Exploitation

You are an expert music prompt analyzer. Your task is to break down a music prompt into a structured list of semantically complete atomic elements.
CORE PRINCIPLES

1. Each element must be a complete, meaningful musical concept. Do NOT break down elements further if they lose their core musical meaning (e.g., "melancholic piano melody" is one element).
2. Group all identified elements into the 4 standard categories provided. Every conceptual part of the original prompt MUST be assigned to one, and only one, category. Do not omit any parts.
3. Do not add any elements that were not present in the original prompt.
EXAMPLE
Return ONLY a valid JSON object with the exact structure shown in the example above.

B Data Construction

B.1 Prompt2Music Dataset Details

The core objective of Prompt2Music is to create a dataset of 5000 high-quality, structured prompt-music pairs. To achieve this, we designed and implemented a multi-stage data engineering pipeline that combines human musical expertise with the large language model GPT-4o.

To ensure the professionalism and effectiveness of the prompts, we first designed a structured generation template based on established best practices in music prompt engineering. This template requires each prompt to comprehensively cover the following four core dimensions:

1. Style & Genre: Establish the fundamental musical framework and aesthetic tradition of the composition, define the overarching conventions that govern its melodic, harmonic, and rhythmic characteristics.
2. Instrumentation & Timbre: Specify the selection of sound-generating sources: encompassing acoustic, electronic, or synthetic instruments, and detail unique sonic textures, tonal colors.
3. Mood, Atmosphere & Scene: Defines the emotional core and environmental narrative, translating abstract acoustic signals into specific psychological states and evocative settings.
4. Structure, Tempo & Dynamics: Dictates the temporal organization and formal arrangement of the piece, including the rhythmic pace, musical sections, and the variations in intensity and volume.

Following this template, we invited music professionals to manually craft 100 high-quality "master" prompts to serve as high-standard seeds for the expansion process. Using these samples and the template as reference, We provided the following meta-prompt to GPT-4o to expand the set into a diverse corpus of 5,000 prompts:

Prompt Template of Prompt Generation

You are a world-class `{AI Music Prompt Engineer}`. Your core expertise is crafting prompts that guide text-to-music models to produce high-quality, stylistically specific music. Your task is to generate unique and detailed prompt for music tracks. Each prompt must be a single, cohesive paragraph and meticulously structured to provide a clear and evocative creative direction.
For each prompt, you must seamlessly weave together the following elements:
1. ***Style & Genre***: Establish the fundamental musical framework and aesthetic tradition of the composition, define the overarching conventions that govern its melodic, harmonic, and rhythmic characteristics.
2. ***Instrumentation & Timbre***: Specify the selection of sound-generating sources—encompassing acoustic, electronic, or synthetic instruments, and detail unique sonic textures, tonal colors.
3. ***Mood, Atmosphere & Scene***: Defines the emotional core and environmental narrative, translating abstract acoustic signals into specific psychological states and evocative settings.
4. ***Structure, Tempo & Dynamics***: Dictates the temporal organization and formal arrangement of the piece, including the rhythmic pace, musical sections, and the variations in intensity and volume.
EXAMPLE

In the process of generating the corpus, we applied a quality control script for filtering. This script was primarily used to remove prompts that were overly simplistic or highly repetitive, thereby ensuring the diversity and complexity of the final set of 5000 prompts. To further guarantee the reliability of the expansion, we randomly sampled the generated prompts and invited music professionals to audit their musical logic and descriptive accuracy.

We then selected four representative open-source models, each with different architectural focuses and generation capabilities: Stable Audio Open, MusicGen, InspireMusic, and ACE-Step. To ensure full experimental reproducibility, all audio generation processes adhered to the following fixed parameters:

- **Random Seed**: A fixed random seed was used for all generation processes.
- **Duration**: The duration of all generated audio clips was uniformly set to 30 seconds.
- **Content Constraint**: All prompts were for instrumental music only, excluding any vocals. (This decision was based on the limitations of most current open-source models in synthesizing vocals and to ensure the research focus remained on the reverse engineering of

musical elements.)

The final Prompt2Music dataset consists of 5000 data tuples, totaling 20,000 audio files.

B.2 Prompt2Music Statistics

To demonstrate the semantic diversity of our Prompt2Music dataset, Figure 3 presents a t-SNE visualization of all 5,000 prompts using embeddings from all-MiniLM-L6-v2. Colors represent local point density: yellow indicates isolated prompts while dark blue indicates dense semantic clusters.

The visualization reveals multiple distinct clusters distributed across the embedding space, confirming that our dataset covers diverse musical concepts spanning genre, mood, instrumentation, and tempo. The dispersed distribution with no dominant cluster validates that Prompt2Music provides broad semantic coverage suitable for evaluating prompt stealing attacks across varied musical domains.

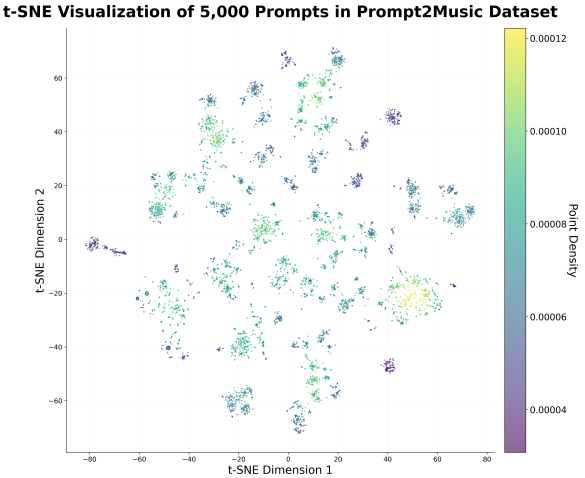

Figure 3: t-SNE visualization of 5,000 text prompts in Prompt2Music dataset.

B.3 Test Datasets Details

All our primary experiments were conducted on a subset of the Prompt2Music dataset. For each experimental run, we randomly sampled 100 prompt-audio pairs as the ground truth.

To assess the robustness and generalization capabilities of our attack framework, we curated a collection of 400 text descriptions from four well-known public music datasets. These texts vary in style, structure, and origin, representing a broader and more diverse "real-world" scenario. Unlike the structured prompts in Prompt2Music, these texts are often descriptive: post-hoc descriptions of existing music. Testing against this data allows us to understand how our methods perform on non-standardized and varied inputs.

Subsequently, we used these text descriptions to synthesize corresponding audio clips on the unified InspireMusic model, adhering to the same parameters detailed in Appendix A.1. These newly formed caption-music pairs then served as the ground truth for our "Different Dataset" experiments.

Below are each public dataset we used:

- **MusicCaps**: MusicCaps is a large-scale dataset of approximately 5,500 music-text pairs with captions written by human experts. We utilized its *caption* field. The text in this field typically consists of concise, high-quality paragraphs focusing on the music's mood, genre, and key instruments.
- **LP-MusicCaps-MSD**: This dataset is built by a LLM-Based pseudo music captioning model. We extracted content from the *caption_writing* field.
- **MusicBench**: MusicBench expands MusicCaps's caption by including music features of chords, beats, tempo, and key that are extracted from the audio and describing these music features using text templates. We used the *main_caption* field, which are augmented text prompts.
- **Wikimt-x**: Wikimt-x is a multimodal MIR benchmark dataset, primarily from 20th-century Western music. We chose the *description* field, which is generalized overview, excluding identifiable details.

C Experimental Setups

C.1 Evaluation Metric

To comprehensively evaluate the performance of the attack, we use three distinct metrics that assess the similarity from different perspectives: the textual fidelity of the prompt, the acoustic similarity of the generated audio, and the structural correspondence.

Semantic Similarity This metric quantifies the textual correspondence between the reconstructed prompt and the original ground-truth prompt. The process involves using the text encoder from all-MiniLM-L6-v2 to generate vector embeddings for both the stolen prompt and the ground-truth prompt. The final similarity score is the cosine similarity calculated between these two text embeddings.

Music Similarity This metric measures the perceptual and acoustic similarity between the tar-

get audio and the audio generated from the stolen prompt. The implementation uses the laion/clap-htsat-fused model. First, both audio files are loaded and resampled to a 48kHz sample rate. Then, the pre-trained CLAP model generates a 512-dimensional embedding vector for each audio file. The final score is the cosine similarity between the two audio embeddings, which is a widely adopted metric for this purpose. For reproducibility, a fixed random seed (42) is used throughout the process.

ByteCover3 Similarity To provide a more objective measure of musical correspondence, we treat the stolen audio as a "cover version" of the target audio and evaluate them using a state-of-the-art Cover Song Identification (CSI) model, ByteCover3. The model processes both the target audio and the stolen audio to extract a corresponding embedding vector for each. The final similarity is calculated as the cosine similarity between these two embeddings.

C.2 Human-rated Similarity Criteria

The detailed criteria for each level are stated in Table 4.

C.3 Standard Deviation of Human Ratings

To assess the consistency and reliability of our human evaluation, we report the standard deviation (SD) of human-rated similarity scores in Table 5. For each audio pair, ten domain experts provided ratings on a 5-point Likert scale, and we computed both the mean and standard deviation.

The standard deviations across all experimental conditions remain relatively consistent, indicating stable inter-annotator agreement. This consistency suggests that the human evaluation protocol provided sufficient clarity for experts to make reliable perceptual judgments.

Notably, AudioStealer consistently achieves both higher mean similarity scores and comparable or lower standard deviations compared to baseline methods, demonstrating that its regenerated audio not only aligns more closely with target music but also produces more consistent perceptual quality across different judges. This pattern holds across both the Prompt2Music cross-model experiments and evaluations on public datasets, confirming the robustness of our findings.

C.4 Details of Adapted T2I-Stealer

The Adapted T2I-Stealer baseline is designed to evaluate the direct transferability of existing image-based prompt stealing frameworks, specifically the dual-branch architecture consisting of a subject generator and a modifier detector. Its implementation is tailored to address the unique structural differences between visual and acoustic signals.

- **Subject Generator**: We utilize OpenJMLA to function as the subject generator. It extracts the primary musical elements and foundational descriptions from the target audio, providing a core semantic anchor.
- **Modifier Detector**: In the image domain, modifier detectors often use Transformer decoders with learnable label embeddings as queries to attend to localized visual features. However, music concepts are diffusely distributed across spectrogram patches rather than being localized in specific spatial coordinates. To adapt to this, we replace the localized query mechanism with a CLAP-based retrieval task. This approach simulates a multi-label classifier by matching the global audio embedding against a predefined set of musical

Level	Description
1 Not similar at all	The musical pieces differ entirely in core style and genre, evoke contrasting moods, and use completely different lead instruments with no shared tempo or structure.
2 Slightly similar	The musical pieces share a minor common genre element or mood, but differ significantly in lead instruments and tempo, with no overlapping structure.
3 Moderately similar	The musical pieces align in core genre or mood, but differ in lead instruments and lack shared supporting elements, with varying tempo or structure.
4 Quite similar	The musical pieces match in core genre, share similar lead instruments, and include some supporting elements, though tempo or minor structural details differ.
5 Very similar	The musical pieces are nearly identical in core style, genre, and mood, feature the same lead and supporting instruments, and align in tempo, structure, and dynamics.

Table 4: Human evaluation criteria for music similarity.

Category	Target / Dataset	Method	Human Rating (Mean ± SD)
Prompt2Music (Cross-Model)	Inspire	OpenJMLA	2.56 ± 0.92
		Adapted T2I	2.42 ± 1.02
		Mu-LLaMA	2.48 ± 0.98
		Qwen2-Audio	2.52 ± 0.96
		AudioStealer	**3.32 ± 0.80**
	ACE-Step	OpenJMLA	2.38 ± 0.96
		Adapted T2I	2.46 ± 1.08
		Mu-LLaMA	2.62 ± 0.98
		Qwen2-Audio	2.95 ± 0.94
		AudioStealer	**3.28 ± 0.82**
	MusicGen	OpenJMLA	2.25 ± 0.90
		Adapted T2I	2.32 ± 1.02
		Mu-LLaMA	2.14 ± 1.02
		Qwen2-Audio	2.45 ± 0.96
		AudioStealer	**2.66 ± 0.93**
Public Datasets (on Inspire)	MusicCaps	OpenJMLA	2.82 ± 0.96
		Adapted T2I	2.78 ± 0.98
		Mu-LLaMA	2.56 ± 1.04
		Qwen2-Audio	2.60 ± 1.00
		AudioStealer	**3.26 ± 0.88**
	LP-MusicCaps	OpenJMLA	2.46 ± 0.94
		Adapted T2I	2.50 ± 0.92
		Mu-LLaMA	2.44 ± 1.06
		Qwen2-Audio	2.38 ± 1.08
		AudioStealer	**3.08 ± 0.90**
	MusicBench	OpenJMLA	2.56 ± 0.95
		Adapted T2I	2.54 ± 0.91
		Mu-LLaMA	2.30 ± 0.98
		Qwen2-Audio	2.68 ± 0.94
		AudioStealer	**3.25 ± 0.85**
	Wikimt-x	OpenJMLA	2.68 ± 0.96
		Adapted T2I	2.65 ± 0.92
		Mu-LLaMA	2.56 ± 0.98
		Qwen2-Audio	2.44 ± 0.97
		AudioStealer	**3.18 ± 0.81**

Table 5: Comprehensive statistics of human-rated similarity scores. For Public Datasets, all samples were synthesized using the unified InspireMusic model as the target.

modifiers. We adopt the standardized MTG-JAMENDO tag list as our modifier universe. To identify relevant attributes, we calculate the prediction probability for each tag. Following the established protocols of image-based methods, we set a confidence threshold of 0.6. Only tags with a probability score exceeding this threshold are identified as modifiers.

- **Stolen Prompt Construction**: To generate the final output, we adopt the identical concatenation operation as described in the reference image work (PromptStealer). The identified modifiers are appended to the subject description to form the complete stolen prompt.

D Efficiency Analysis

Figure 4 presents the efficiency behavior of AudioStealer on the Prompt2Music dataset, where both the target and shadow models are set to InspireMusic to eliminate potential interference from model architecture differences. The convergence pattern demonstrates that AudioStealer reaches near-optimal performance within 80 queries to the local shadow model.

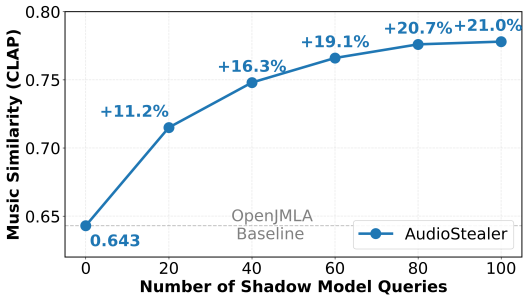

Figure 4: Efficiency analysis of Music Similarity scores across shadow model queries.