RL-Soundtrack: Reinforcement Learning for Video-Aware Sequential BGM Recommendation

Yan-Chi Lu **R14921045** **Ting-Wei Huang** **R14921058**

Yi-An Lai **R14942082** **Yun-He Lin** **R14942098**

National Taiwan University
https://github.com/YianLai0327/RL_final_project

Abstract

This study addresses the critical challenge of automated background music (BGM) recommendation for narrative-driven Vlogs, where the selected soundtrack must not only align with immediate visual content but also maintain long-term thematic coherence. Existing solutions face significant limitations: traditional retrieval-based methods typically rely on greedy algorithms that optimize for local alignment, behaving in a myopic manner that ignores global narrative structure; meanwhile, generative audio models often lack the fine-grained control required to select from fixed, copyright-compliant music libraries. To bridge this gap, we propose *RL-Soundtrack*, a reinforcement learning framework that formulates BGM selection as a finite-horizon Markov Decision Process (MDP). We employ Maskable Proximal Policy Optimization (Maskable PPO) [1] to train an agent capable of making sequential decisions within a constrained action space. By integrating multimodal grounding—utilizing Gemini [2] for narrative segmentation and ImageBind [3] for cross-modal alignment—our agent learns to perceive and adapt to narrative shifts. Experimental results demonstrate that RL-Soundtrack significantly outperforms greedy baselines in long-horizon scenarios, achieving a 12% increase in cumulative reward, which evidences its superior global planning capability. Furthermore, our method achieves a 13x speedup in inference time compared to linear-scan retrieval methods, leveraging the $O(1)$ complexity of the policy network to ensure high scalability for real-time applications.

1 Introduction

The proliferation of short-form video content (e.g., Instagram Stories, TikTok) has spurred significant advancements in automated background music (BGM) recommendation. Existing systems excel at *pointwise matching*—retrieving a single audio track that aligns with the static mood or visual semantics of a short clip. However, as content creation shifts towards narrative-driven formats like Vlogs, these *static* approaches are becoming increasingly insufficient. A typical travel Vlog is not a snapshot of a single emotion; it is a *temporal sequence* evolving from calm introductions to energetic transitions and reflective conclusions. This dynamic nature requires a BGM recommendation system that possesses *global awareness* and the ability to plan for long-term coherence, rather than merely optimizing for immediate relevance.

Current approaches to video soundtracking broadly fall into two categories: generative audio and retrieval-based methods, both of which face distinct limitations in this context. Generative models such as MusicGen [4], while capable of creating novel audio from text prompts, fundamentally address a problem of "creation" rather than "curation". Professional video editing often relies on fixed, high-quality *royalty-free libraries* to ensure copyright compliance and production standards. Generative models struggle to select from these discrete, pre-defined constraints. On the other hand, traditional Retrieval-based methods typically employ *greedy algorithms*. These methods select the best-matching track for the current segment independently, behaving in a *myopic* manner. By ignoring temporal dependencies, greedy strategies often result in disjointed transitions, repetitive loops, or poor pacing, failing to capture the narrative "flow" essential for high-quality video production.

To address these gaps, we propose *RL-Soundtrack*, a novel framework that reformulates BGM recommendation as a *Sequential Decision Process*. Unlike greedy heuristics, our approach utilizes Deep Reinforcement Learning (DRL) to train an agent capable of optimizing a cumulative return that balances local video-audio alignment with global thematic consistency. By modeling the problem as a Markov Decision Process (MDP), the agent learns to make trade-offs—sacrificing immediate match scores when necessary to preserve the long-term structure and pacing of the video.

Our contributions are threefold:

1. **Sequential Decision Formulation:** We are the first to formulate the fixed-library BGM recommendation task as an MDP, enabling the use of Maskable PPO to handle variable-length videos and validity constraints.

2. **Multimodal Narrative Grounding:** We design a robust observation space that integrates Gemini-based semantic segmentation for event detection [2] and ImageBind embeddings for cross-modal alignment [3], allowing the agent to perceive narrative shifts rather than arbitrary time windows.

3. **Efficiency and Scalability:** We demonstrate that our RL-based inference operates with constant time complexity $O(1)$, independent of the library size. This results in a 13x speedup compared to linear-scan greedy baselines, making our system highly scalable for real-time applications.

2 Related Work

2.1 From Pointwise to Sequential Music Recommendation

Background music recommendation has traditionally been formulated as a *pointwise retrieval* task. Conventional methods typically treat video segments as independent queries, selecting the best-matching track for each timeframe in isolation. While effective for short-form content, these greedy approaches are myopic—they optimize for immediate relevance while neglecting the temporal dependencies and thematic consistency required for long-form narratives. Conversely, recent advancements in Generative Audio, such as MusicGen [4], have enabled the synthesis of high-fidelity music from textual descriptions. However, these generative approaches face significant *library constraints* in professional workflows, as they cannot select from specific, copyright-cleared catalogues. Our work bridges this gap by formulating BGM curation as a *sequential decision process*. Unlike pointwise retrieval, our RL-based approach optimizes a trajectory of choices, ensuring that the selected sequence maintains global coherence while adhering to the constraints of a fixed music library.

2.2 Multimodal Grounding for Video Analysis

Effective BGM recommendation relies on the precise alignment of visual and auditory modalities. Early works often relied on handcrafted features or separate embedding spaces, which limited their ability to capture semantic nuances. The emergence of *joint multimodal embeddings*, such as ImageBind [3], has revolutionized this field by projecting video, audio, and text into a shared latent space, enabling direct arithmetic comparison across modalities. Furthermore, to move beyond arbitrary time-windowing, modern approaches utilize Large Multimodal Models (LMMs) for semantic understanding. In our framework, we leverage Gemini [2] to perform narrative-driven segmentation, identifying distinct "events" within a video. This combination of ImageBind for feature alignment and Gemini for temporal segmentation provides a robust, semantically grounded observation space for the RL agent.

2.3 RL in Recommender Systems

Deep Reinforcement Learning (DRL) has emerged as a robust paradigm for recommender systems, particularly for its capacity to model sequential decision-making processes and optimize for long-term cumulative rewards rather than localized, immediate clicks. Early research in this domain primarily focused on value-based methods or standard actor-critic architectures to capture evolving user preferences over time. Among these, Proximal Policy Optimization (PPO) [5] has gained significant traction due to its ability to ensure stable policy updates and prevent catastrophic performance drops through its clipped objective function.

However, applying standard PPO to specialized tasks like BGM recommendation presents unique challenges, particularly regarding the validity of action spaces across different time steps. For instance, in our sequential decision process, the "CONTINUE" action is logically invalid at the first step ($t = 1$) or immediately after a track has reached its conclusion. To address this, we integrate Maskable PPO [1] regarding invalid action masking in policy gradient algorithms. By explicitly masking out illegal transitions within the policy network such as selecting to continue a non-existent or finished track, the agent is forced to explore only the valid action space.

3 Problem Formulation

We formalize the task of automatic video soundtrack generation as a sequential decision-making process. Unlike static information retrieval, where a single query maps to a single result, soundtrack generation requires constructing a temporal sequence of background music (BGM) tracks that align with a visually evolving narrative while maintaining musical continuity and thematic coherence.

3.1 System Modeling

We formally define the entities within our system to establish a rigorous basis for optimization. Let the video dataset be denoted as $\mathcal{V}$, where each video $V \in \mathcal{V}$ is a sequence of T temporal segments, $V = \{v_1, v_2, \dots, v_T\}$. Each segment v_t is projected into a multi-modal embedding space. Each segments v_t is characterized by a visual embedding $\mathbf{e}_t^v \in \mathbb{R}^{d_1}$ and a textual caption embeddings $\mathbf{e}_t^{v,c} \in \mathbb{R}^{d_2}$.

Similarly, the music library is defined as $\mathcal{M} = \{m_1, m_2, \dots, m_M\}$. Each track m_i is characterized by an audio embedding $\mathbf{e}_i^m \in \mathbb{R}^{d_1}$ and a textual description embedding $\mathbf{e}_i^{m,c} \in \mathbb{R}^{d_2}$, residing in the same vector spaces as their video counterparts to facilitate semantic comparison. Furthermore, to capture the structural properties of music, we extract a vector of acoustic features ϕ_i (e.g., BPM, energy, spectral centroid) for each track.

We model the quality of a soundtrack using three fundamental properties:

- Alignment ($\mathcal{A}$): The degree of semantic correspondence between the video segment v_t and the concurrent music m_{a_t}. This is quantified via cosine similarity in the shared embedding space: $\cos(\mathbf{e}_t^v, \mathbf{e}_{a_t}^m)$.
- Smoothness ($\mathcal{S}$): The auditory transition quality between consecutive steps $t-1$ and t. This comprises both semantic continuity $\cos(\mathbf{e}_{t-1}^m, \mathbf{e}_t^m)$ and acoustic stability $\Delta(\phi_{t-1}, \phi_t)$.
- Coherence ($\mathcal{C}$): The global thematic consistency of the selected tracks across the entire episode, preventing jarring style shifts.

3.2 Multi-Objective Optimization

The soundtrack curation problem is formulated as a multi-objective optimization task. We seek a policy π that generates an action sequence $A = (a_1, \dots, a_T)$ to maximize a weighted sum of alignment, smoothness, and coherence, subject to soft constraints on music switching frequencies.

Let w_{align}, w_{smooth}, w_{switch}, and w_{theme} denote the importance weights for each objective. The global objective function $\mathcal{J}$ is:

$$\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} \left(w_{\text{align}} \mathcal{A}(v_t, a_t) + w_{\text{smooth}} \mathcal{S}(a_{t-1}, a_t) - \mathcal{P}(a_t) \right) + w_{\text{switch}} \mathcal{R}_{\text{switch}} + w_{\text{theme}} \mathcal{C}(A) \right]$$

Here, $\mathcal{P}(a_t)$ penalizes unconventional repetition, and $\mathcal{R}_{\text{switch}}$ is a global reward ensuring the frequency of music changes matches a video-specific target budget. This formulation necessitates a solution that balances immediate local rewards (alignment) with long-horizon planning (coherence and budget adherence), rendering greedy approaches insufficient.

3.3 Markov Decision Process (MDP) Characterization

This problem formulation naturally satisfies the Markov property, as the suitability of a chosen track at time t depends explicitly on the immediate visual context and the musical context established at $t-1$. Consequently, we cast the soundtrack selection process as a finite-horizon Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$:

State Space ($\mathcal{S}$) The state formulation is critical for enabling the agent to make context-aware decisions. At decision step t, the state $\mathbf{s}_t$ must encapsulate both the current visual stimulus and the auditory context established by previous actions. We define the state as the concatenation of the current video segment embedding and the embedding of the music track selected in the previous step:

$$\mathbf{s}_t = \left(\mathbf{e}^{v}_{i,t} ; \mathbf{e}^{m}_{t-1} \right)$$

For the initial step $t = 1$, we initialize $\mathbf{e}^{m}_{0}$ as a zero vector $\mathbf{0}$. This design allows the agent to perceive the *flow* of the soundtrack; by observing $\mathbf{e}^{m}_{t-1}$, the agent can choose a subsequent track that is acoustically compatible, thereby maximizing the smoothness objective, or detect when a sharp visual transition warrants a change in musical style.

Action Space ($\mathcal{A}$) We define a discrete action space:

$$\mathcal{A} = \{1, \ldots, N\} \cup \{\textsc{Continue}\}$$

The agent may either select a new track m_i from the library or execute a CONTINUE action to prolong the current track. Crucially, this space is state-dependent. The CONTINUE action is only valid if a track is currently playing and has not reached its end. Conversely, if $t = 1$ or the previous track duration D_i has been exhausted, the CONTINUE action must be masked. We implement this using valid action masking, which explicitly prevents the agent from sampling invalid actions during both training and inference.

4 Methodology

The proposed framework, illustrated in Figure 1, leverages Reinforcement Learning to traverse the semantic-audio space defined above. The pipeline begins with a Splitting module that decomposes raw video into segments, utilizing Gemini for scene-cutting timing detection during training. The core innovation lies in the RL-Soundtrack agent's ability to optimize long-horizon objectives that greedy retrieval methods fail to address.

4.1 RL-Soundtrack Agent

The decision-making core is powered by Maskable Proximal Policy Optimization (Maskable PPO) [1]. Standard policy gradient methods assume a static action space; however, our environment requires strict adherence to dynamic constraints (e.g., the CONTINUE action is invalid if the current track has finished).

To enforce these constraints within the differentiable policy optimization, we intervene at the logit level. Let $\mathbf{l}_t$ denote the raw logits output by the policy network and $\mathbf{m}_t \in \{0, 1\}^{|\mathcal{A}|}$ be a binary

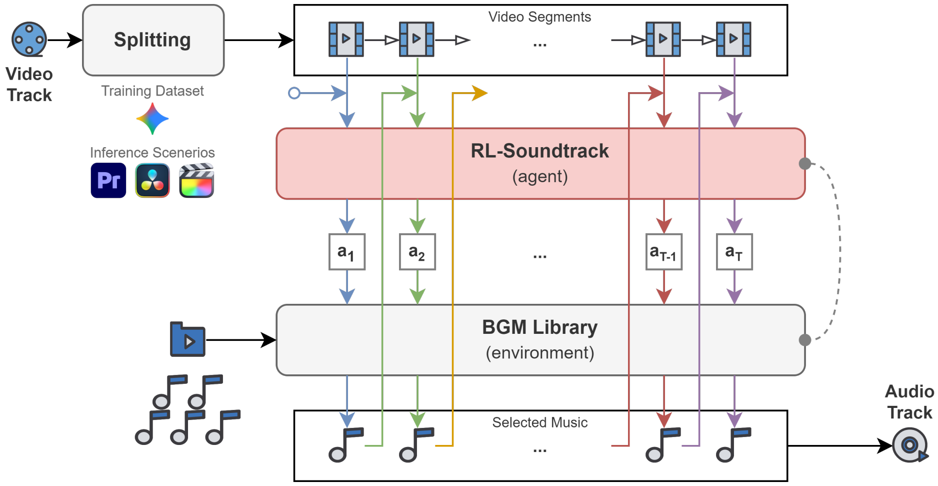

Figure 1: RL-assisted BGM recommendation architecture. The agent perceives a multi-modal state and optimizes a trajectory of musical selections subject to validity masking.

validity mask where $m_{t,i} = 1$ indicates a valid action. We construct the masked logits $\tilde{\mathbf{l}}_t$ via a conditional gating operation:

$$\tilde{l}_{t,i} = \begin{cases} l_{t,i} & \text{if } m_{t,i} = 1 \\ -\infty & \text{if } m_{t,i} = 0 \end{cases}$$

In practice, $-\infty$ is implemented as a sufficiently large negative constant (e.g., -10^8) to ensure numerical stability. The final policy distribution is obtained by applying the softmax function to these modified logits: $\pi(a_t|\mathbf{s}_t) = \text{softmax}(\tilde{\mathbf{l}}_t)$. This mechanism guarantees that the probability mass assigned to invalid actions is exactly zero, restricting the agent's exploration strictly to the valid decision manifold.

4.2 Reward Engineering

The efficacy of the method stems from its reward structure, which translates the aesthetic criteria of film scoring into scalar feedback signals. The reward function $R(\mathbf{s}_t, a_t)$ is decomposed into local and global terms.

Local Dynamics The step-wise reward aggregates alignment and smoothness. The alignment term r_t^{align} computes the centered cosine similarity between $\mathbf{e}_t^v$ and $\mathbf{e}_t^m$ relative to a dataset reference $\mathbf{e}_{\text{ref}}^v$, mitigating the hubness problem in high-dimensional embedding spaces. The smoothness term r_t^{smooth} penalizes semantic continuity $\cos(\mathbf{e}_t^m, \mathbf{e}_{t-1}^m)$ and acoustic dissonance $\Delta(\phi_{t-1}, \phi_t)$ but is modulated by a video-aware gating factor g_t. This gate relaxes penalties during sharp visual transitions, allowing the agent to learn that abrupt musical changes are permissible during scene cuts.

Global Constraints At the terminal step T, the agent receives sparse rewards for trajectory-level coherence. We model the switching budget r^{switch} using a conditional Gaussian $p(C \mid V)$, penalizing deviations from the predicted optimal pacing. Furthermore, thematic coherence r^{theme} is calculated by comparing individual track embeddings against the episode's mean embedding centroid. This multi-objective reinforcement signal forces the agent to balance immediate semantic matching with the long-term planning required to construct a coherent auditory narrative.

4.3 Multimodal Grounding

To enable the RL agent to make decisions based on high-level narrative context rather than arbitrary time intervals, we implement a multimodal grounding pipeline that processes both visual and acoustic signals into a unified semantic space.

4.3.1 Narrative-Driven Video Segmentation

Traditional background music recommendation systems often slice videos into fixed-length windows (e.g., every 5 seconds). However, narrative transitions in Vlogs rarely align with fixed time steps. We employ Gemini 2.5 flash as a semantic video encoder. Instead of processing static frames, we feed the raw video stream alongside an audio energy trace (loud/quiet segments) into the model. The model is prompted to segment the video timeline based on "events" (e.g., *00:00-00:15: Energetic travel montage*, *00:15-00:45: Speaking to camera*). This ensures that the RL agent's decision points (t) correspond to actual narrative shifts.

4.3.2 Signal-Constrained Audio Captioning

To represent the music library semantically, we generate rich text descriptions for each track. A common issue with Large Multimodal Models (LMMs) is hallucination—describing a slow song as "energetic" due to lack of grounding. To mitigate this, we introduce a Signal-Constrained Captioning module.

Before prompting the LMM, we extract the BPM (Beats Per Minute) via onset strength envelopes. This physical feature is injected directly into the system prompt (e.g., CONTEXT: BPM=128) to ground the generated description in the track's actual tempo.

Furthermore, to ensure a consistent observation space for the RL agent, we apply Vocabulary Constraints. The model is restricted to select mood tags exclusively from a predefined set of discrete categories (e.g., Happy, Epic, Dark). Crucially, this vocabulary is strictly aligned with the output space of our video segmentation module in Sec. 4.3.1, ensuring that both modalities share a unified semantic representation for calculating the alignment reward.

5 Experimental Results

5.1 Experimental Environment

All experiments were conducted on a laptop workstation equipped with an AMD Ryzen 9 8945HS CPU and an NVIDIA RTX 4060 Laptop GPU. To strictly evaluate the efficiency of the proposed CPU-based inference pipeline, the GPU was utilized solely for preliminary embedding extraction, while all Reinforcement Learning (RL) training and inference processes were executed exclusively on the CPU. The operating system was Windows Subsystem for Linux 2 (WSL2). The implementation was developed using Python 3.12, utilizing `gymnasium` for environment simulation and `stable-baselines3` along with `sb3-contrib` for the implementation of the Maskable PPO algorithm.

5.2 Dataset

5.2.1 Music Library Construction

Our background music (BGM) library consists of 123 royalty-free tracks curated from Incompetech (Kevin MacLeod), a standard source for YouTube content creators. The dataset was filtered to include diverse moods (e.g., Epic, Jazz, Horror, Upbeat) while excluding tracks longer than 5 minutes.

To simulate a professional editing workflow, we applied a Chorus Extraction preprocessing step. Raw music tracks often begin with long, low-energy intros that are unsuitable for immediate background scoring. We utilized chroma feature analysis to identify the "high-relevance" section (chorus) of each track. Furthermore, to ensure seamless concatenation during the RL process, the start point of each track was snapped to the nearest musical beat and zero-crossing point, preventing auditory artifacts (clicks) and rhythmic misalignment.

5.2.2 Video Dataset

To facilitate the training and evaluation of our RL-based recommendation system, we constructed a high-quality video dataset with 367 videos specifically focused on travel vlogs. Our collection process began with a broad search of the 12 most prominent travel vloggers from both Mandarin and English-speaking regions (24 candidates total). We then narrowed this list to 7 representative

YouTubers who consistently utilize BGM to drive their narratives: @TheDoDoMen, @bensadventure, @elephantgogo, and @LenaPatrick (Mandarin); and @KaraandNate, @drewbinsky, and @YesTheory (English). To ensure the dataset remained focused on narrative-driven travel content, we utilized Gemini to analyze video titles, automatically excluding non-travel content such as unboxings, interviews, advertisements, and challenges. Also, the top 367 videos based on view counts were selected for crawling to ensure high production value and diverse visual contexts.

Source Separation: We applied Demucs [6] to perform vocal-music separation. This separation enables high-fidelity music-video synthesis during testing. Moreover, it provides the necessary ground-truth data for our reward sanity check in Sec. 5.3.

Narrative Segmentation via KCPD: To enable the model to learn the nuanced timing of music transitions as practiced by human editors, we move beyond fixed-window observation. We implement Kernel Change-point Detection (KCPD) [7] to identify temporal boundaries where narrative shifts occur within the vlogs. Our framework utilizes these detected points as dynamic reward signals.

5.3 Reward Discriminability Analysis (Sanity Check)

Since the reward function in our framework is manually designed to capture both the cross-modal synergy between video and audio and the intra-modal consistency of the musical sequence (global reward), its performance involves an inherent degree of subjectivity . To validate the efficacy of our model, we conduct a sanity check to objectively demonstrate that our integrated reward design aligns with human preferences. The goal is to verify that the combined reward mechanism assigns higher scores to original, human-edited video-audio sequences compared to random, mismatched pairings.

Hypothesis A well-designed reward function assigns higher scores to correctly matched video–music pairs, while mismatched pairs receive lower rewards

Experimental Setup Given a set of ground-truth (GT) video-audio pairs extracted from human-edited travel vlogs. To isolate the BGM, we utilized source separation to obtain no-vocal audio tracks from the original footage. We then constructed a Reward Confusion Matrix by calculating the total reward for all possible permutations of video segments and audio tracks within the sample set.

Results and Analysis The results, visualized via rank-based coloring (row-wise), reveal strong discriminative capabilities for our complex reward structure, as can be seen from Figure 2:

- **Diagonal Dominance**: For a given video stream, the highest reward score was assigned to the original ground-truth audio pair in 6 out of 7 cases (ranking 1st).

- **Near-Perfect Alignment**: In the single case where the ground-truth was not the top-ranked choice, it still achieved the 2nd highest reward among all possible pairings, demonstrating that even "mismatched" high-scoring tracks likely share similar thematic properties with the original .

Conclusion These findings provide empirical evidence that our reward design effectively codifies the latent aesthetic and narrative choices made by human editors. By consistently prioritizing ground-truth pairs over mismatched ones, the reward function proves to be an objective and reliable metric that successfully captures the professional "flow" of human-produced content.

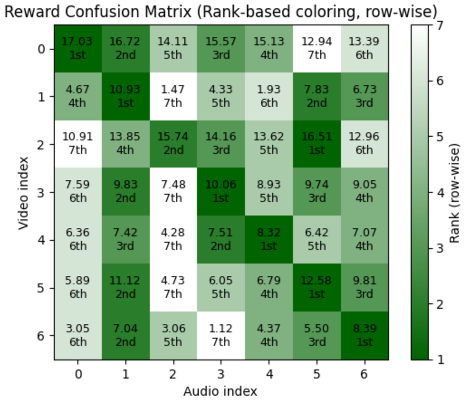

Figure 2: Reward Confusion Matrix for Video-Music Alignment.

5.4 Comparative Evaluation

We compared the proposed RL-Soundtrack method against three baselines to evaluate its effectiveness in long-horizon decision-making:

- **Random:** An agent that selects valid actions uniformly at random, serving as a lower bound for performance.
- **Greedy VM-aligned:** A greedy heuristic that selects the music track with the highest immediate video-music alignment reward r_t^{align} at each step, ignoring temporal smoothness and global consistency.
- **Greedy All:** A stronger heuristic that optimizes the full step-wise reward r_t (alignment + smoothness) greedily. This baseline accounts for local transitions but lacks the foresight to optimize global constraints due to the natural limitation of greedy algorithms.

5.5 Performance Analysis

We evaluated the cumulative returns of all methods across both the training and testing sets. Fig. 3 and Fig. 4 present the Cumulative Distribution Function (CDF) of the total rewards per episode.

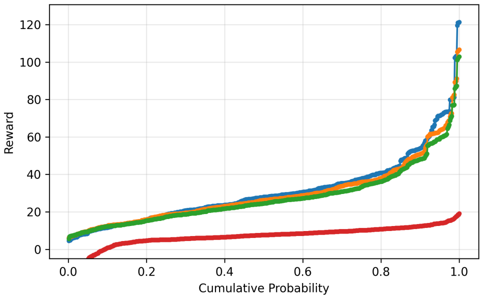

Figure 3: CDF of total reward on training set.

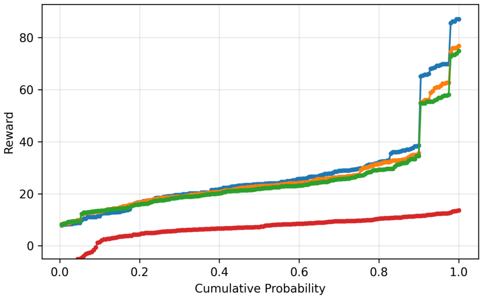

Figure 4: CDF of total reward on testing set.

Training Performance The RL agent demonstrated superior performance on the training set, achieving a 15% increase in total reward compared to the strongest baseline (Greedy All). The Random policy performed significantly worse, confirming the non-triviality of the task.

Generalization On the unseen testing set, our method maintained its advantage, outperforming the baselines by 12%. This indicates that the RL agent successfully generalized its policy to new video narratives rather than memorizing specific track-video combinations.

Horizon Analysis Because video with more segments naturally have higher total rewards, longer videos will be located on the right side of the CDF. As shown in Fig. 3 and Fig. 4, the performance gap between RL and Greedy approaches is most pronounced in videos with longer horizons. In videos with few segments, the search space for the optimal solution $|\mathcal{M}|^T$ is relatively small, allowing greedy choices to occasionally approximate the global optimum. As the number of segments T increases, the sequential search space explodes exponentially ($|\mathcal{M}|^T$). Greedy algorithms, limited to a linear exploration scope of $|\mathcal{M}| \cdot T$, fail to capture long-term dependencies such as thematic coherence and switching budgets. By optimizing expected cumulative returns, our RL agent effectively navigates this expanded search space to maximize global objectives.

5.6 Inference Efficiency

We analyzed the computational efficiency of the proposed method versus the greedy baseline. Let $|\mathcal{M}|$ be the size of the BGM library.

- Greedy Policy: Requires iterating through all available tracks in the library at every decision step to calculate alignment scores. The inference complexity is $\mathcal{O}(|\mathcal{M}|)$, making it linearly dependent on library size.
- RL Policy: The inference cost is determined by the forward pass of the policy network, which is independent of the library size (once trained). The complexity is $\mathcal{O}(1)$ relative to $|\mathcal{M}|$.

Empirical results confirm this theoretical advantage. As shown in Table. 1, our RL policy achieves a 13× speedup than the Greedy approach. This constant-time inference capability ensures that our system remains scalable and suitable for real-time applications even as the BGM library grows significantly.

Table 1: Inference time of our method and baselines

Policy	Description	Inference Time (ms/step)
Greedy	Step Reward	5.1725 ± 0.2141
Ours RL	Maskable PPO	**0.3845 ± 0.0974**

6 Conclusion and Discussion

In this work, we presented RL-Soundtrack, a novel reinforcement learning framework designed for autonomous, video-aware sequential background music (BGM) recommendation. By reformulating the BGM curation process as a Markov Decision Process (MDP) and employing a Maskable PPO agent, we successfully addressed the inherent "myopic" limitations of traditional greedy retrieval methods.

Our experimental results demonstrate that RL-Soundtrack achieves a 12% performance gain in total rewards compared to greedy baselines, particularly in long-horizon video scenarios where the search space grows exponentially. By effectively balancing local multi-modal alignment (via ImageBind and signal-constrained captioning) with global thematic coherence, our model captures the narrative *flow* essential for high-quality vlog production.

Furthermore, we highlighted a significant breakthrough in inference efficiency. Our RL approach operates with constant-time $O(1)$ complexity, resulting in a 13x speedup (reducing latency from 5.17 ms to 0.38 ms per step) . This scalability ensures that the system remains highly responsive even as BGM libraries expand, making it an ideal solution for real-time video editing and large-scale content creation platforms.

In conclusion, RL-Soundtrack validates the superiority of sequential decision-making over simple retrieval for complex multi-modal. Future work will focus on integrating this framework directly into Non-linear Editing (NLE) software. Leveraging its constant-time $O(1)$ inference efficiency, the system can provide creators with real-time, automated soundtracking that dynamically adapts to the video as they edit.

References

[1] Shengyi Huang and Santiago Ontañón. A closer look at invalid action masking in policy gradient algorithms. *arXiv preprint arXiv:2006.14171*, 2020.

[2] Gemini Team and Google. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[3] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[4] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

[5] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

[7] Wei-Cheng Chang, Chun-Liang Li, Yiming Yang, and Barnabás Póczos. Kernel change-point detection with auxiliary deep generative models. *arXiv preprint arXiv:1901.06077*, 2019.