Crowdsourcing Open-License Music Tags on Closed Creative Platforms: A Negative Case Study of a Game-with-a-Purpose on TikTok

Edgar Eggert
Stanford University; Charité Berlin
edgar.eggert@charite.de

Philipp Stolberg
Independent Researcher
philipp@stolberg.ch

Abstract

Generative music systems rely on audio-caption pairs linking musical sounds to natural-language descriptions, yet openly licensed audio such as the Free Music Archive lacks these paired descriptions. This limits researchers' ability to build reproducible, openly redistributable datasets for generative music research. Games with a purpose (GWAP) could close this gap, but prior music GWAPs lacked an audience on their dedicated sites. We deploy a GWAP filter inside TikTok Effect House, whose algorithmic feed reaches 1.12 billion daily users. A ten-second filter rewards speech and keyword matches over 42 bundled FMA clips; an offline pipeline (separation, ASR, taxonomy parsing, ensemble matching) recovers transcripts and clip identities from posted videos, with full recovery on 25 author-recorded videos. There was no organic traction, and 116 outreach attempts produced zero published videos. We report this as a negative case study of platform choice: Effect House constrains game design, while algorithmic distribution cannot be reliably triggered through direct outreach, leaving the cold-start problem unresolved.

1 INTRODUCTION

Open-source generative music AI models need training pairs of audio and natural-language descriptions. The Free Music Archive (FMA) (Defferrard et al., 2017) covers tens of thousands of permissively licensed tracks researchers can freely redistribute for this purpose, but its annotations frequently stop at coarse genre tags and release-level metadata. The intersection of open-licensed music and the natural-language descriptions a modern captioning or text-to-music system consumes is therefore thin. This limits researchers' ability to build reproducible, openly redistributable datasets for training and evaluating generative music systems. More broadly, in the context of generative music, audio-caption pairs shape the language through which music can be described and created. The vocabulary in these pairs influences what users can ask models to generate and, in turn, which musical concepts models can represent. The resulting annotation problem is therefore not only how to collect labels, but how to obtain natural-language descriptions of music at sufficient scale and under conditions that preserve their usefulness for AI-music research. We test this problem through a Game with a purpose (GWAP) deployed in TikTok Effect House: the question is whether a platform that already has a large audience can turn participation into the audio-language pairs that open generative-music research requires.

GWAPs (von Ahn, 2006; von Ahn and Dabbish, 2004) have long been proposed as a cheaper alternative to expert annotation, and a string of music-specific GWAPs, including TagATune (Law et al., 2007), MajorMiner (Mandel and Ellis, 2007), the listen-game family (Turnbull et al., 2008), HerdIt (Barrington et al., 2009, 2012), MagnaTagATune (Law et al., 2009), Hooked (Burgoyne et al., 2013), and Emotify (Aljanaki et al., 2016), have demonstrated that non-expert players produce reusable music annotations when the task is wrapped inside enjoyable gameplay. None of these games has reached the scale its design assumed, however, because the deployments lived on dedicated web sites or social platforms that carried their own cold-start distribution problem: players had to be recruited to the site, and recruiting the crowd was the bottleneck.

The bet behind this work is that the mechanic is sound and the venue has been the constraint. TikTok carries roughly 1.9 billion monthly and 1.12 billion daily active users (Backlinko, 2026) and exposes a native runtime for short interactive mini-games through TikTok Effect House. A new effect can theoretically be surfaced to that audience by the recommendation system without prior follower volume. Running a GWAP music-description mechanic inside an Effect House filter is a test of whether a platform that already has the crowd can convert the GWAP idea into the dataset volume open-source music research needs, at a cost an academic project can afford.

A filter was built as a ten-second mini-game. The player taps to start, a clip drawn at random from a bundled library of 42 FMA tracks plays back, and the player describes aloud what they hear; on-device scoring rewards detected speech and matches against a curated keyword list of genres, instruments, and moods. The filter cannot log sessions directly, so annotations are recovered after the fact from the public videos players post, through an offline pipeline that separates speech from music with Demucs (Défossez et al., 2019), transcribes the speech stem through a cloud automatic speech recognition (ASR) service, parses the transcript against a music-descriptor taxonomy, and identifies which of the 42 clips played using an ensemble of MFCC and chroma features under dynamic time warping (DTW) and a Shazam-style constellation fingerprint (Wang, 2003). Section 3 describes the filter and Section 4 describes the recovery pipeline.

The pipeline handled every song that was tested. On 25 example videos recorded by 3 distinct speakers, source separation and ASR produced coherent transcripts and the matcher recovered the played clip on every video. The pro-

cessing side is ready for input. TikTok did not supply that input. Following a failure to gain organic traction, a combined outreach campaign of 116 attempts, comprising 14 TikTok One marketplace pitches and 102 direct messages to creators using comparable interactive filters, produced zero usable published videos. The filter reached no organic audience.

The paper documents the working pipeline end to end, from a published Effect House filter to per-clip structured annotations; the pipeline handles mobile short-video audio without intermediate cleanup. The 0-of-116 outreach result is reported as a negative case study: under the conditions tested, our filter did not overcome the recruitment bottleneck faced by dedicated GWAP platforms. This result does not establish that platform-based GWAPs cannot work in general, but indicates that this approach on TikTok Effect House is unlikely to provide annotation at scale. Discord, which exposes a first-class developer SDK for embedded interactive games (Discord, 2026) and centres user activity on opt-in communities rather than an algorithmic feed, is named as the next venue to test.

2 RELATED WORK

Four strands of prior work inform the design and the evaluation of this filter: games with a purpose for music and audio annotation, source separation and speech recognition for noisy mobile audio, audio fingerprinting and matching, and music tagging and captioning corpora that this kind of GWAP is meant to create.

2.1 Games With A Purpose for Music and Audio Annotation

The games-with-a-purpose (GWAP) paradigm was introduced by von Ahn and colleagues (von Ahn and Dabbish, 2004; von Ahn, 2006) as a design pattern in which human computation is embedded inside enjoyable gameplay so that players generate labelled data as a byproduct of a game they would play anyway. The ESP Game (von Ahn and Dabbish, 2004) demonstrated the design on image labelling at web scale and articulated the output-agreement template that subsequent music GWAPs adopted.

A line of music GWAPs followed: TagATune, MagnaTagATune, MajorMiner, the listen-game family, HerdIt, Hooked, and Emotify (Law et al., 2007, 2009; Mandel and Ellis, 2007; Turnbull et al., 2008; Barrington et al., 2009, 2012; Burgoyne et al., 2013; Aljanaki et al., 2016). They shared a basic shape (short clips with free-form or constrained tags, rewards tied to inter-rater agreement) and a basic limitation: each had to recruit its own audience on a dedicated web page or social plugin. Player recruitment was the rate-limiting step.

A more recent line of work has begun moving the GWAP onto a major social platform. Polyphonic (Martelaro et al., 2021) ran a Twitch-based audio GWAP whose game consumed sound contributions inside a streamer's existing community, and reported 212 validated sounds across 41 categories from three of four field-tested streamers; the platform was the venue for sustained interactive gameplay, while the streamer brought the audience. The experiment did not test what happens when the streamer-host is removed.

The TikTok filter in Section 3 extends the GWAP idea with a single-player, asynchronous, spoken-description mechanic, deployed inside a short-video platform that already has an audience but no streamer-host. The motivation is the same gap these earlier games set out to close: openly licensed music such as FMA is available in volume but is not paired with the natural-language descriptions open-source captioning or text-to-music models need.

A complementary line of work studies crowdsourced audio annotation outside the GWAP tradition. FSD50K (Fonseca et al., 2020) is an openly licensed corpus of more than fifty thousand Freesound clips labelled across two hundred sound-event classes through iterative validation rounds with per-class agreement metrics. FSD50K is the operating point a mature crowdsourced pipeline aims for; this case study, with its small validation set, is far from that scale and is cited as the eventual target rather than a present comparator.

2.2 Source Separation and Speech Recognition for Noisy Audio

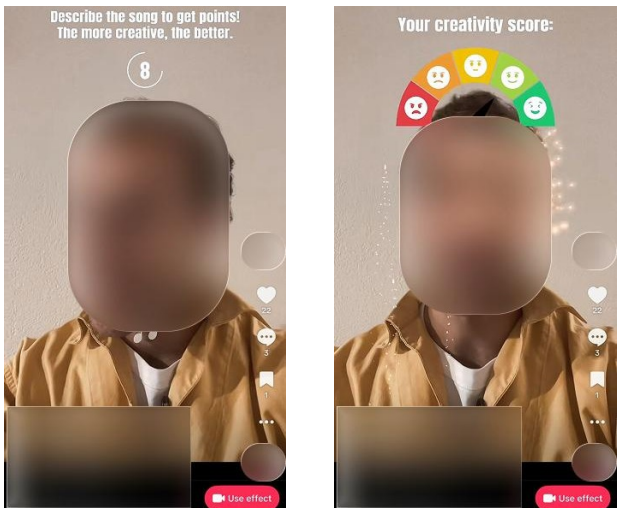
Recovering a player's spoken description requires separating speech from the filter's music playback, since both are captured through the same device microphone. The pipeline in Section 4 uses the Hybrid Transformer Demucs variant from the Demucs toolkit lineage (Défossez et al., 2019); the cited paper is the originating waveform-domain architecture rather than the specific weights deployed here. Speech transcription on the separated voice stem is performed by Deepgram Nova-2, a commercial cloud automatic speech recognition (ASR) service. Deepgram is one of several large-scale weakly supervised cloud recognisers in the same general class as Whisper (Radford et al., 2022); the analogy is approximate, since Deepgram's training corpus is not public, and the analogy is named here only to locate the technology, not to claim a controlled comparison. Whisper is included as a fallback path inside the pipeline.

2.3 Audio Fingerprinting and Matching

Identifying which of the 42 bundled clips a player heard is a closed-set retrieval problem against short, compressed speaker-microphone queries. The reference design is the constellation fingerprint of Wang (2003), in which each track is represented by combinatorially hashed pairs of spectrogram peaks and matching proceeds through constant-time hash lookups followed by a time-offset histogram consistency check. The approach tolerates additive noise and lossy compression, but it assumes the query is a near-copy of the reference, which breaks under pitch shifts and under overlaid speech. The pipeline therefore pairs the fingerprint matcher with a feature-based branch that compares Mel-frequency cepstral coefficients (MFCC) and chroma features under dynamic time warping (DTW). DTW absorbs small local timing deviations the fingerprint cannot.

2.4 Music Tagging and Captioning Models

The clip library bundled into the filter is drawn from the Free Music Archive (FMA) (Defferrard et al., 2017), whose permissive terms support academic and derivative use. FMA is a common reference corpus for music-analysis work that needs to avoid commercial licensing constraints, which is why it is also a natural training resource for open-source captioning and text-to-music systems. Its annotation layer stops at coarse genre tags and release-level metadata, and that gap is what filter-collected descriptions are meant to fill.



(a) Countdown (b) Score reveal

Figure 1: Two states of a filter session. **(a)** During the ten-second countdown, the player is prompted to "Describe the song to get points! The more creative, the better." while a randomly drawn FMA clip plays through the phone speaker. The on-device scoring rule awards $+3$ points for any detected speech utterance and $+10$ points each time the player's speech matches a term in a curated keyword taxonomy of genres, instruments, and moods. **(b)** At countdown expiry, an animated meter reveals a final creativity score on a 0 to 100 scale, mapped onto five expressive icons. Faces and caption text are blurred for anonymity.

Recent work shows the demand side of this loop. Pons et al. (2018) reported that end-to-end music-tagging accuracy scales with labelled corpus size on a 1.2-million-track proprietary corpus, motivating open-data investment for reproducible work. MusCaps (Manco et al., 2021) trained an encoder-decoder for free-form music captioning and showed that natural-language descriptions recover instrumentation and mood cues that closed tag vocabularies miss. MuLan (Huang et al., 2022) pushed further with a contrastive two-tower model on roughly 44 million weakly aligned music-text pairs and demonstrated zero-shot tagging without per-task fine-tuning, but its weights and corpus are not publicly released. Reproducible alternatives in this space need open-licensed audio paired with natural-language descriptions, which is precisely what a TikTok-bundled FMA library run through the filter pipeline would produce. The 25-song validation in Section 5 demonstrates that the pipeline can manufacture such pairs end to end; the missing piece is volume, which Section 6 argues TikTok cannot supply for this kind of filter.

3 FILTER DESIGN

The data-collection surface is a TikTok Effect House filter that runs as a short mini-game inside the TikTok camera. The filter was authored in Effect House version 5.1.1, published roughly one week after final edits once it cleared TikTok's standard review pipeline, and has remained live in that form since. The compiled project contains 64 logic layers and ten audio-related assets, with additional song clips delivered through a referenced asset library. Figure 1 shows two states from a session.

3.1 Why the Filter Exists

The filter is a games-with-a-purpose (GWAP) instrument for expanding the natural-language annotation coverage of openly licensed music (von Ahn, 2006). Open-source captioning and text-to-music models train on audio the developers can redistribute, and the Free Music Archive (FMA) (Defferrard et al., 2017) is one of the larger pools that meets that constraint; FMA ships with genre tags and release metadata, which leaves the descriptive layer modern audio-language models consume unfilled. The filter is the GWAP idea re-cast as a short-video effect, on the bet that TikTok's recommendation system can supply the audience that the dedicated-site GWAPs surveyed in Section 2 struggled to attract.

3.2 Game Mechanics

Gameplay is a ten-second creativity challenge. On tap-to-start a countdown begins and one clip is drawn at random from the bundled library of 42 short songs. The clip plays back through the phone speaker while the player speaks into the device microphone. Two scoring rules run live on-device: any detected spoken input scores $+3$ points, and each match against a curated keyword list scores $+10$ points (Figure 1a). When the countdown expires, an animated meter reveals a creativity score on a 0 to 100 scale (Figure 1b).

The spoken output is what the downstream pipeline harvests. The score is a proxy for engagement and is not a proxy for annotation quality. No ground-truth label is available on-device, so the keyword-hit rule is the only mechanism that nudges the player toward music vocabulary during the session itself.

3.3 Keyword Taxonomy

The on-device keyword list is small by design, because Effect House runs it as a set of string-match nodes in the visual graph and offers no learned-classifier surface that could absorb a wider vocabulary. After plumbing strings are stripped, 17 distinct keyword terms remain, organised into three categories: ten genre terms, four instrument terms, and three mood terms. These counts are gameplay parameters; the dataset annotations come from the cloud-ASR transcript parsed against the richer downstream taxonomy in Section 4, not from the in-game keyword list. The two remaining slots in the seed ontology, geography-and-culture and an *other* catch-all, hold no terms in this release of the filter.

Each keyword is paired with a phonetic alternate that is close in sound but different in spelling, such as *Ambient/ambien*, *Electronic/Electro*, *Pop/pope*, *Hip-hop/hiphop*, *Rock/Rack*, *Bass/Base*, *Beat/bit*, *Funky/Fun*, and a standalone *Emotional*. The alternates exist because on-device automatic speech recognition (ASR) is lower in fidelity than cloud ASR: many players say *rock* but the on-device recogniser transcribes *rack*, so both strings trigger the $+10$ reward. This avoids punishing players for recogniser errors outside their control. The richer taxonomy used for downstream transcript analysis is discussed in Section 4.

The narrow vocabulary is a deliberate trade-off given the platform constraints below: a broader list would inflate false positives on common English words inside long utterances, and Effect House's visual-graph string-matching surface does not support a learned classifier that could

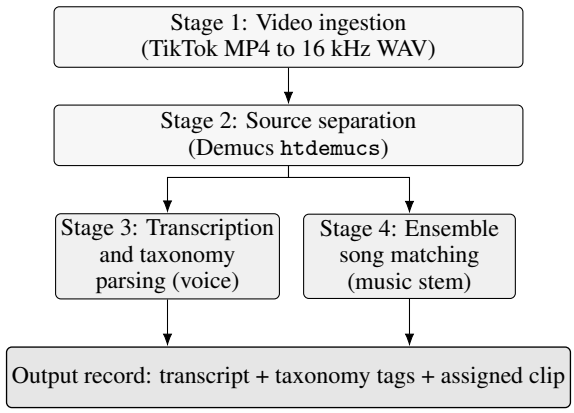


Figure 2: Four-stage pipeline that recovers a spoken description and an assigned bundled clip from each TikTok video, with the shared ingestion and separation stages feeding two parallel branches of equal depth that converge into a single output record.

resolve the ambiguity. The transcript itself, parsed downstream against a richer taxonomy, is what feeds the dataset.

3.4 Platform Constraints

Effect House imposes constraints that shape the filter, each documented in TikTok's developer guidance (TikTok Effect House, 2026). The platform caps the total bundled-asset size of a single filter at 8 MB and offers no live connection to an external audio database, so the clip library was restricted to 42 songs, each heavily downsampled and trimmed to fit within the cap. The on-device ASR surface is limited and non-configurable, with no language model or vocabulary biasing, which forces the phonetic-alternate design above. The filter cannot call out to a logging endpoint, so the only record of a session is the TikTok video the player chooses to post publicly, recovered later through TikTok's built-in share-and-download path. Finally, Effect House's scripting surface is a visual node graph with no general-purpose programming substrate, which rules out richer interactions such as multi-turn description or per-clip difficulty curves. An implementation walkthrough of the patterns used to live within these constraints is given in Appendix C.

The filter's audio clips are drawn from the Free Music Archive (FMA) (Defferrard et al., 2017). FMA was selected because its tracks are openly licensed and because the motivation of this work is to expand the annotated surface of open-source, royalty-free music for open-model developers. No part of the pipeline targets a proprietary training corpus.

4 PIPELINE

The filter cannot log sessions directly (Section 3), so the labelling signal is recovered after the fact from public TikTok videos that players post. The recovery pipeline has four sequential stages, shown in Figure 2. The pipeline is supporting evidence that the processing side handles what arrives; the contribution of this paper sits in Section 6.

Video ingestion downloads each public TikTok video through the platform's built-in share path and extracts a 16 kHz mono audio track. The sample rate matches the downstream automatic speech recognition (ASR) service and retains enough bandwidth in the music stem for feature-based matching.

Source separation splits the extracted track into a player-voice stem and a residual music stem; the phone speaker plays the clip while the microphone captures the player's description, so the raw audio mixes the two. We use the Demucs `htdemucs` variant (Défossez et al., 2019). When the player records in a noisy environment or speaks over percussive sections of the clip, the separator leaks speech energy into the music stem and vice versa, and that leakage causes downstream transcription and matching errors.

Transcription converts the voice stem to text using the Deepgram Nova-2 prerecorded ASR service, with Whisper retained as an offline fallback. A parser then scans the transcript for tokens from the reference music taxonomy, which covers genre, instrumentation, mood or texture, and cultural or geographic descriptors.

Song matching binds the residual music stem to one of the 42 bundled clips. This is a closed-set retrieval problem with strong audio degradation from the phone speaker-microphone chain, platform compression, and source separation. The pipeline runs an ensemble of two complementary matchers. A feature branch uses Mel-frequency cepstral coefficients (MFCC) for timbral colouring and chroma for harmonic content, both under dynamic time warping (DTW). A fingerprint branch follows the spectrogram-landmark construction of Wang (2003): pairs of nearby time-frequency peaks are hashed, and consistent matches are counted against a precomputed reference index. A guarded weighted vote combines the two branches. For each input video the pipeline writes one record per clip containing the assigned song identifier with its combined matcher confidence, the transcript, and the parsed taxonomy tags.

4.1 Parameter Choices and Scope of Validation

The matcher weights and thresholds were chosen heuristically from pilot listening tests during development; they were not tuned against a held-out set. In the feature branch, the combined DTW score weights MFCC at 0.6 and chroma at 0.4, because chroma is more fragile on compressed speaker-microphone audio. When the primary branch's top-match confidence falls below a guard threshold of 0.85, the fingerprint branch is consulted and the two branches enter a 0.5 and 0.5 weighted vote. A full statistical ablation that attempts to partition the accuracy figure into per-branch contributions is not run at the current sample size; 25 songs are below the count at which such a partition can be distinguished from noise.

5 EVALUATION

The evaluation has two parts. The pipeline is validated end to end on 25 example videos to confirm that the technology side is ready for input, and the distribution effort is reported as a 116-attempt outreach campaign that produced zero published videos.

5.1 Pipeline Validation

The pipeline was run end to end on 25 example videos recorded by 3 distinct speakers using the published filter. Each video was a single ten-second session. For every video, the authors compared the recovered transcript against the original recording and confirmed that the tran-

script reflected what was actually said, and confirmed that the matcher's top-1 song identification corresponded to the FMA clip the filter had played.

Transcription. Automatic speech recognition (ASR) produced a transcript for each of the 25 inputs with zero pipeline failures. All 25 outputs were classified as coherent, meaning the recognised tokens formed interpretable music descriptors suitable for downstream natural-language processing. Transcript length ranged from 3 to 16 words, consistent with the ten-second open-ended description mechanic. Representative outputs include "Chinese strings, melancholic.", "Emotional guitar, sweets, high school movie.", and "This is giving, like, radio head slash the Smith vibes kind of cute.". Some are terse genre-plus-mood pairs; others are longer, reference-laden lines.

Audio matching. The audio matcher was tasked with identifying, for each video, which of the 42 clips bundled into the filter had been played. Top-1 accuracy was 25 of 25 on author-recorded, author-verified videos, with the played clip recovered correctly on every video despite the fact that the matcher's confidence values were uniformly low in absolute terms, capping at 0.44. The conservative scoring is consistent with accurate top-1 selection over a small closed set of 42 clips. The figure cannot be generalised to the full 42-clip library or to conditions outside the filter environment.

5.2 Outreach and Distribution

After initial publication of the filter did not reach organic traction, the authors began with direct distribution strategies. The distribution effort ran as a single case-study campaign with no controlled comparator. Fourteen creators with relation to music content were pitched through the TikTok One creator marketplace and none accepted. A subsequent direct-messaging campaign sent 102 messages and received 12 responses, a response rate of 11.76%. Three of the twelve responders requested payments above $100, which was outside the available budget; the remaining responses did not convert. The conversion rate across both channels was 0%, across 0 of 116 total outreach attempts.

The evaluation's headline is therefore distributional: TikTok did not deliver a usable audience for this GWAP. Section 6 unpacks the structural factors behind the 0-of-116 figure.

5.3 Threats to Internal Validity

The 25-of-25 top-1 accuracy figure should be read against three threats that the small validation regime cannot fully rule out. The ground-truth protocol is not blinded: the authors verified each transcript and each top-1 match against the original recordings, so a reviewer should treat the figure as author-verified, with blinded third-party re-verification outside this study's scope. The matcher weights and the guarded-vote threshold were selected during development with author listening feedback; the reported accuracy is therefore a joint property of the full ensemble rather than a per-branch estimate. The 25-video sample is also too small to support any population-level claim about pipeline accuracy, so the figure indicates that transcription and matching are not the failure mode without supporting a stable accuracy estimate.

6 DISCUSSION

This case study addresses a practical problem in open-source music research: producing audio paired with natural-language descriptions at scale. Open-source captioning and text-to-music models need such pairs, and the Free Music Archive (FMA)(Defferrard et al., 2017) provides a large pool of openly licensed audio but not the descriptive layer these models consume. The games-with-a-purpose paradigm (von Ahn, 2006; von Ahn and Dabbish, 2004; Law et al., 2007; Mandel and Ellis, 2007; Barrington et al., 2009, 2012; Burgoyne et al., 2013; Aljanaki et al., 2016) is the mechanism this paper tests for filling that layer. The narrow question is whether TikTok is the right venue for a GWAP of this shape, and the answer reported here is that it is not.

6.1 Why TikTok Did Not Work for This GWAP

The 0-of-116 outreach result is the headline. The more useful contribution is the structural reason behind it. Three factors interact, and each points at a property of TikTok or its Effect House runtime, not the pipeline.

The first is the set of design constraints imposed by Effect House (TikTok Effect House, 2026). The 8 MB asset cap forces the clip library to roughly 42 short downsampled songs per published filter, which limits per-player variety as the surprise of the random draw decays. The visual-graph scripting surface offers no learned-classifier substrate, so the in-game reward is restricted to keyword string matching against 17 terms; richer mechanics like multi-turn description or per-clip difficulty are not implementable. The on-device ASR surface is non-configurable, which is why the keyword list ships with phonetic alternates instead of vocabulary biasing. These constraints together produce a thin game loop, and a thin loop is a hard sell on a platform whose median session is built around face effects rather than spoken description.

The second is the social ask. Speaking aloud on a public recording is heavier than being transformed by a silent face effect, and it inherits the social friction of recorded speech on a public feed. The interaction also rewards verbal creativity, which is a posture few users will perform on camera in front of an algorithmic audience.

The third is virality. Short-form recommendation systems are stochastic for new effects without prior follower mass or trend attachment, and the TikTok One marketplace is gated on creator economics that this project's per-post budget could not match: three of twelve responders quoted asks above $100, outside the available budget. The project posted from accounts with no prior follower base or trend signal, removing the largest lever short-form video offers a new effect; TikTok's recommender weights existing engagement and trend attachment when deciding what to surface.

No controlled comparison was run, and the evaluation does not attribute the 0-of-116 figure to any one factor in isolation. Together, these conditions constrained what could be tested within the deployed Effect House configuration: the asset cap and scripting surface constrained the game, the social ask was a property of the chosen mechanic, and virality depended on conditions outside the study's control. The TikTok One ask prices encountered in outreach also provide a small but relevant signal about the labor involved in collecting music descriptions: three creators quoted more than the project could afford for a single post.

Although these quotes cannot establish a market rate, they illustrate that annotation is not a free byproduct of a game when participation must compete for attention on a commercial platform. This matters for open generative-music research, where acquiring natural-language descriptions at scale entails a labor cost that is often obscured by the dataset itself (Martelaro et al., 2021). The case study cannot rule out a hypothetical filter that succeeded under a different design or with a paid-creator partnership inside the recommender's trust envelope.

6.2 Why the Pipeline Worked

The pipeline side handled what it received. On every input video, Demucs-family source separation (Défossez et al., 2019) split player speech from the bundled clip, cloud ASR produced a coherent transcript, and the ensemble matcher recovered the played clip under author verification. Confidence values capped at 0.44, but over a closed set of 42 clips top-1 ranking holds well below the absolute confidence levels the matcher emits. These numbers place the failure outside the pipeline; they are supporting evidence for the distributional finding.

6.3 Discord as the Remaining Alternative

Polyphonic's success on Twitch was conditional on a recruited streamer-host whose community supplied the audience and whose voice prompted contributions live. The TikTok experiment reported here substituted an algorithmic feed for that host, and the result was negative. We therefore do not claim that algorithmic-platform GWAPs fail generally; rather, this case suggests that algorithmic reach alone is insufficient to overcome the recruitment bottleneck under the conditions tested here. The next venue we propose to test is one that retains the host-supplies-community property without requiring per-streamer recruitment. Discord fits, and Polyphonic itself names it: in their discussion of small-streamer dynamics, the authors observe that a Discord or another dedicated out-of-stream space is where ongoing community interaction lives when the live stream alone cannot support it (Martelaro et al., 2021).

The mechanic deployed in this paper translates onto Discord without the constraints that bound it on TikTok. Discord's Embedded App SDK exposes a programmable iframe surface inside voice channels, text channels, and direct messages; Activities are full HTML5 web applications, with the platform handling authentication and presence (Discord, 2026). Clip libraries can sit on a developer-controlled origin and need not fit inside an 8 MB asset bundle. Scoring can run server-side against a wider vocabulary than the 17-term keyword list a visual-graph node tree can express. The recruitment surface is opt-in music-oriented servers whose members have already declared an interest in the topic. These properties make Discord a plausible next test, not a demonstrated solution. The open question is whether community-mediated recruitment can generate useful music descriptions at scale, and whether the resulting annotations are sufficiently consistent and descriptive for use in open generative-music research. Twitch is not named as a future recommendation: Polyphonic already covered that platform under recruited-streamer conditions.

6.4 Alternative Designs Considered

Three alternative designs were weighed during development. A constrained multiple-choice mechanic (tap to pick from four genre terms) would have lowered cognitive cost and removed the speak-aloud social ask, but would have collapsed each session to a single categorical tag and lost the multi-attribute description that recovered transcripts already carry; example transcripts in Section 5 mention genre, mood, era, and instrument inside a single ten-second utterance.

A browser-based standalone GWAP would avoid TikTok's runtime constraints, at the cost of reintroducing the cold-start distribution problem TikTok was supposed to solve. A backbone-level swap inside the recovery pipeline, such as a domain-adapted music-vocals separator or a far-field music-aware recogniser that skips separation, is worth a future pass and is out of scope at the current 25-video sample. Appendix E gives the longer reasoning on each option.

7 LIMITATIONS

The case study has limitations on both sides of the result, recorded here so a reviewer can weigh the negative finding against the conditions under which it was produced.

Single filter, single cold-start campaign. Only one Effect House filter was built and only one outreach campaign was run, distributed from accounts with no prior follower base, no trend attachment, and no creator partnerships. Existing follower volume and recent trend signal are the two largest levers TikTok's recommender weights when surfacing a new effect, so the case study tests the hardest setting in which a filter has to earn organic reach from scratch. The 0-of-116 result provides evidence against the simpler hypothesis that any working GWAP filter on TikTok will collect data; it does not rule out the possibility that a differently designed filter, a more professionally produced game, or coordination with an ongoing music trend could perform differently. After the initial filter failed to gain traction, the authors considered a second iteration based on the alternative designs described in Appendix E. However, the fixed asset cap, limitations of the scripting surface, absence of session logging, and absence of an agreement mechanism would have remained unchanged. A second campaign would therefore have tested a different design under essentially the same platform constraints, without resolving whether those constraints were responsible for the lack of participation. We leave iterative redesign and comparative testing to future work.

Game-design expertise. The authors are computational researchers and independent musicians, not professional short-form game designers. A more virality-engineered design within Effect House's bounds (different scoring curves, sound design, prompt phrasing, panel pacing) could plausibly have improved retention and shareability. The case study cannot separate the contribution of design skill from the contribution of platform constraints.

Annotation validity. The filter prompts players to "Describe the song to get points! The more creative, the better." Creative descriptions are not necessarily accurate or useful as music captions. Because the campaign produced no participant data at scale, we could not evaluate the quality, consistency, or usefulness of the resulting descriptions. Future deployments should compare prompting strategies and assess caption quality and agreement across independent participants before treating collected descriptions as training data.

Effect House technical bottleneck. Section 6 argues the Effect House runtime is the structural bottleneck, but that is an inference from the observed pattern rather than a controlled finding. A direct test would build the same mechanic on a platform with a programmable runtime, such as Discord Activities, and compare retention and contribution volume side by side. However, Effect House caps the clip library at about 42 short songs per published filter through its 8 MB asset cap (TikTok Effect House, 2026); scaling to hundreds of tracks requires a platform architecture change.

Sample size and ground truth. The pipeline was validated on 25 example videos recorded by 3 distinct speakers. The 25-of-25 transcript-coherence and top-1 matching figures are reported as indicators that the pipeline is not the bottleneck; they do not support a population-level claim about pipeline accuracy on arbitrary speakers, environments, or recording conditions. A reviewer should treat the figures as author-verified rather than held-out-verified, since the authors assessed each transcript against the original recording and each top-1 match against the played clip. A true word error rate cannot be computed because ground-truth transcripts were not collected at recording time; transcript coherence is reported instead.

Licensing. The Free Music Archive licensing (Defferrard et al., 2017) has not been independently audited for the specific case of redistribution inside a TikTok filter. A licence review is required before any dataset constructed from filter-collected annotations is released publicly under an open licence.

These limitations should be read jointly. A more engaging filter design (Appendix E lists the alternatives weighed during development) might collect richer tags than the version reported here, but the cold-start exposure problem on the algorithmic feed, the creator-marketplace economics that bounded the paid channel and especially the platform's hard limits on game-design surface and clip volume would likely keep even a well-engineered filter from producing a curated music corpus on TikTok at scale.

8 CONCLUSION

This paper reports a negative case study that questions the feasibility of TikTok as a games-with-a-purpose (GWAP) distribution channel for crowdsourced open-license music annotation on the Free Music Archive (FMA) (Defferrard et al., 2017). A ten-second Effect House mini-game asked players to describe a bundled clip aloud and rewarded detected speech and on-device keyword matches against a curated music vocabulary. A four-stage pipeline of source separation, automatic speech recognition (ASR), taxonomy parsing, and ensemble matching turned each posted video into a structured per-clip annotation. The pipeline was validated end to end on 25 videos recorded by 3 distinct speakers, with all transcripts coherent and the played clip recovered correctly on every video, so the processing side is ready for input. The distribution side produced no usable videos to derive tags from: 116 outreach attempts converted to zero published videos.

The structural cause behind the negative result lies in TikTok's Effect House runtime. The platform caps a filter at 8 MB of bundled assets and roughly 42 short clips, exposes only a visual-graph scripting surface for in-game logic, and offers no learned-classifier surface for richer scoring (TikTok Effect House, 2026). Those limits constrain the GWAP design available inside Effect House, and the resulting game did not overcome the organic-reach problem under cold-start conditions. The case study cannot separate platform constraints from imperfect game design or unsponsored launch dynamics with a single filter and a single campaign; however, the Effect House constraints were fixed conditions of the deployment rather than variables the study could manipulate.

The compiled Effect House project and recovery pipeline are available from the authors upon request. The project includes the visual graph and bundled clip metadata; neither artifact redistributes the FMA audio. The FMA licensing terms have not been independently audited for redistribution within a TikTok filter, so any derived annotation dataset would require a separate licensing review before public release.

Discord is the next venue to test. It exposes a first-class developer SDK for embedded interactive games (Discord, 2026) and operates around opt-in communities rather than algorithmic feeds. A Discord Activity removes the constraints binding a TikTok filter (the 8 MB asset cap, the visual-graph scripting surface, the cold-start exposure problem) and recruits from music-oriented servers whose members have already declared interest in the topic. Polyphonic (Martelaro et al., 2021) demonstrated the streaming-platform GWAP direction on Twitch under recruited-streamer conditions; the open question is whether community-mediated recruitment on Discord generalises that result to a music-description mechanic without per-streamer compensation. Two further follow-ups are needed before a derived dataset can be released publicly: a ground-truth transcript collection exercise to replace the coherence-based proxy used in Section 5 with a true word error rate, and a licensing audit of the bundled FMA clips against the FMA terms as those terms apply to a TikTok-filter redistribution case, so it is clear what can be shared under an open licence for open-model research.

AUTHOR DECLARATIONS AND ETHICS STATEMENT

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

Ethical issues. The 25 validation videos were recorded by the authors themselves across three internal speakers; no third-party consent applies. The 116 outreach contacts (14 TikTok One creator-marketplace pitches and 102 direct messages to creators behind comparable filters) were standard creator-outreach contacts; the contact log was not collected as a human-subjects research protocol. The Free Music Archive licensing has not been independently audited for the redistribution-inside-a-TikTok-filter case; a licence review is required before any annotation dataset built from filter-collected data is released publicly under an open licence. The recovery pipeline is designed to process publicly posted TikTok videos that use the filter, extracting speech to derive per-clip annotations. Videos were obtained through TikTok's built-in download button, whose use is covered by the platform's Terms of Service at the time of filter publication; no scraping or automated retrieval was performed. No third-party videos were collected for this study because no external creator published the filter. A future deployment yielding user-generated videos would require a renewed ToS review at that time and a participant-facing notice inside the filter informing

players that posted videos may be processed for research purposes.

Use of AI. The AI coding assistant Claude Code was used for developing the source separation code, ChatGPT was used to cut words, and to proofread spelling and grammar under author direction; all numerical results, claims, and figures were author-verified.

REFERENCES

Aljanaki, A., Wiering, F., and Veltkamp, R. C. (2016). Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management*, 52(1):115–128.

Backlinko (2026). TikTok statistics: Daily and monthly active users.

Barrington, L., O'Malley, D., Turnbull, D., and Lanckriet, G. (2009). User-centered design of a social game to tag music. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*, pages 7–10, Paris, France.

Barrington, L., Turnbull, D., and Lanckriet, G. (2012). Game-powered machine learning. *Proceedings of the National Academy of Sciences (PNAS)*, 109(17):6411–6416.

Burgoyne, J. A., Bountouridis, D., van Balen, J., and Honing, H. (2013). Hooked: A game for discovering what makes music catchy. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 245–250, Curitiba, Brazil.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. (2017). FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China.

Défossez, A., Usunier, N., Bottou, L., and Bach, F. (2019). Music source separation in the waveform domain.

Discord (2026). Discord embedded app SDK (discord activities).

Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. (2020). FSD50K: an open dataset of human-labeled sound events.

Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., and Ellis, D. P. W. (2022). MuLan: A joint embedding of music audio and natural language. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, pages 559–566, Bengaluru, India.

Law, E., von Ahn, L., Dannenberg, R., and Crawford, M. (2007). TagATune: A game for music and sound annotation. In *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, Vienna, Austria.

Law, E., West, K., Mandel, M. I., Bay, M., and Downie, J. S. (2009). Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 387–392, Kobe, Japan.

Manco, I., Benetos, E., Quinton, E., and Fazekas, G. (2021). MusCaps: Generating captions for music audio. Also presented at the International Joint Conference on Neural Networks (IJCNN).

Mandel, M. I. and Ellis, D. P. W. (2007). A web-based game for collecting music metadata. In *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, pages 365–366, Vienna, Austria.

Martelaro, N., Lakdawala, T., Chen, J., and Hammer, J. (2021). Leveraging the Twitch platform and gamification to generate home audio datasets. In *Designing Interactive Systems Conference 2021 (DIS '21)*, pages 1765–1782, New York, NY, USA. Association for Computing Machinery.

Pons, J., Nieto, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., and Serra, X. (2018). End-to-end learning for music audio tagging at scale. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.

TikTok Effect House (2026). TikTok effect house technical optimization guide.

Turnbull, D., Barrington, L., and Lanckriet, G. (2008). Five approaches to collecting tags for music. In *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR)*, pages 225–230, Philadelphia, USA.

von Ahn, L. (2006). Games with a purpose. *IEEE Computer*, 39(6):92–94.

von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 319–326, Vienna, Austria.

Wang, A. L.-C. (2003). An industrial-strength audio search algorithm. In *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR)*, Baltimore, USA.

APPENDIX

A BUNDLED SONG LIST

The 42 clips bundled into the filter are drawn from the Free Music Archive (FMA) (Defferrard et al., 2017). Titles are reproduced below in the order in which they are indexed inside the compiled filter. Four titles appear twice in the list; these are not duplicate entries but two distinct ten-second excerpts drawn from the same parent track, included to give the player a second, non-identical slice of material that the downstream matcher resolves as a separate clip identifier.

B KEYWORD TAXONOMY REFERENCE

The on-device keyword list reduces to 17 distinct terms after plumbing strings are stripped from the graph. The distribution across the seed ontology is shown in Table 2.

#	Title	#	Title	#	Title
1	A Weird Mechanism	15	Insideoutworld	29	Une Paquerette Pour Toi
2	Beach	16	Skysand	30	You Yourself and the Main Character
3	Castle in the Cloud	17	Soundtrack	31	Christmas Lights
4	Childhood Scene	18	Standing	32	Creation Is Like That
5	Christmas Lights	19	Stay Quiet	33	Free To Use 13
6	Dans la messe sans le sang du christ	20	Stephen Katz	34	Frog in the Well
7	You Yourself and the Main Character	21	The Army of You	35	Ice Cream with You
8	Disco Cat	22	The Call of the Polar Star	36	Inversions Part 2
9	Free To Use 9	23	The Evil Plan	37	Land of Comestible People
10	Frog in the Well	24	The Place That Never Get Old	38	Listening
11	Frozen Jungle	25	The Sickness	39	Lost in the City
12	Heimito	26	The Star of Bethlehem	40	Poupi, Great Adventures
13	I Need a Hero Figure, Main Character Theme	27	This Is Happening	41	Saint Wave
14	Ice Cream with You	28	Tu n'as pas de non	42	Shut Up the Mermaid Is Talking

Table 1: Bundled clip titles, indexed by their position in the compiled filter.

Category	Count	Examples
Genre	10	Ambient, Electronic, Pop, Hip-hop, Rock
Instrument	4	Bass, Beat
Mood	3	Funky, Emotional
Geography, culture	0	(none in this release)
Other	0	(none in this release)

Table 2: Keyword terms extracted from the Effect House graph, grouped by seed category.

Each keyword is paired inside the graph with a phonetic alternate that sounds close to the base term but differs in spelling, for example *Ambient* paired with *ambien*, *Pop* paired with *pope*, and *Rock* paired with *Rack*. The purpose of the paired strings is to absorb systematic on-device automatic speech recognition (ASR) substitutions so that a correct player utterance still triggers the in-game reward. The two empty categories are retained in the ontology as placeholders for future releases. See Section 3 for the in-game reward mechanic that consumes this list.

C FILTER IMPLEMENTATION NOTES

The compiled filter project uses Effect House version 5.1.1. The exported graph contains 64 logic layers, and ten assets in the project tree are audio-related (bundled clips and associated metadata). The filter went live roughly one week after final edits once it cleared TikTok's standard review pipeline, and has remained live in that form since. The project source is available from the authors upon request (see Section 8). The notes below cover the patterns the filter uses to live within Effect House's runtime constraints.

C.1 Asset budget under the 8 MB ceiling

Effect House caps the total bundled-asset size of a single filter at 8 MB. The compiled project sits underneath that cap with 42 ten-second clips drawn from FMA, downsampled to a low-bitrate compressed audio format that the platform accepts. Once the audio assets are in place, the remaining budget covers textures, the project scene, language packs, and the visual-graph definition. Adding more clips inside a single published filter is not feasible without a platform-level architecture change, since redistribution from a developer-controlled origin is not available inside the filter runtime.

C.2 Random clip selection in the visual node graph

Effect House's visual-graph scripting surface does not expose a general-purpose pseudo-random number generator. The filter assembles a draw from system-clock state at session start and indexes that value against the 42-clip table. The graph is structured to avoid replaying the same clip inside a single session, so a player who returns to the camera within a session sees a fresh draw; cross-session memory is not available, so the same clip can recur across separate sessions on the same device.

C.3 Phonetic-alternate keyword design

On-device ASR systematically substitutes phonetically similar tokens for the intended keywords (*rack* for *rock*, *ambien* for *ambient*, *pope* for *pop*). The keyword list ships with the most common substitutions paired alongside the canonical tokens, so a correct player utterance still triggers the $+10$ reward. The alternates were harvested empirically during pilot runs; any future GWAP that scores on-device against a vocabulary needs the same alias-engineering pass.

C.4 Publication-as-logging

The filter cannot call out to a logging endpoint. The only record of a session is the public TikTok video the player chooses to post, recovered after the fact through TikTok's built-in share-and-download path. This shapes both the dataset throughput (sessions are bounded by the rate at which players post) and the ethics protocol (every annotation is also a public artefact, with the user's own consent under TikTok's terms).

C.5 64-layer logic graph anatomy

The compiled visual graph contains 64 logic layers, organised around five functional groups: a countdown timer that drives the ten-second session, audio playback against the ten audio-related assets, an ASR listener with the keyword-match string nodes, a score-animation block (the 0-to-100 meter at session end), and an end-screen layer that surfaces the score. The project tree under `Assets/` carries the audio, textures, and `main.scene`; the `Graph/` folder carries the `graph.json` that defines the logic layers; `Library/`, `Record/`, and `Translations/` carry auxiliary assets.

D OUTREACH CAMPAIGN SUMMARY

Table 3 summarises the two outreach channels that were attempted during the distribution phase of the project.

Channel	Pitch.	Resp.	Conv.	Notes
TikTok One	14	0	0	0 accepted
Direct messages	102	12	0	3 asks >$100

Table 3: Outreach by channel. "TikTok One" refers to the TikTok One creator marketplace; "Direct messages" refers to direct messaging creators behind comparable filters.

These numbers are reproduced from the outreach log and are the same figures that support the outreach subsection of Section 5. They are included here as a compact reference for readers who want the per-channel breakdown in one place, and they introduce no claims beyond those in the main text.

E ALTERNATIVE DESIGNS CONSIDERED (EXTENDED)

This appendix gives the longer reasoning behind the three alternative designs named in Section 6.

E.1 Constrained multiple-choice mechanic (rejected)

A constrained mechanic was an early candidate: tap to pick the best-fitting term from a small fixed set, for example one of four genre terms, instead of speaking a free-form description. Lower cognitive cost per session and removal of the speak-aloud social ask are real advantages; the constrained variant might plausibly improve virality relative to the speech variant. The cost is descriptive resolution. Recovered transcripts in this work include "Chinese strings, melancholic.", "Emotional guitar, sweets, high school movie.", and "This is giving, like, radio head slash the Smith vibes kind of cute.". Those utterances carry genre, mood, era, instrument, and cultural-reference cues inside a single ten-second session. A four-option button would have collapsed each session to a single categorical tag; the open-source captioning and text-to-music systems this work targets would lose most of the descriptive signal a transcript carries.

E.2 Browser-based standalone GWAP (rejected)

Migrating the mechanic off TikTok into a browser-based standalone GWAP was a second alternative. The target there is a stand-alone page or citizen-science platform where visitors describe short clips with no TikTok dependency. The 0-of-116 outreach result is evidence TikTok's algorithmic recommender did not supply the audience the project hoped for, so migrating off the platform is a plausible response. A browser-based GWAP reintroduces the cold-start distribution problem TikTok was supposed to solve, however, since a stand-alone site lacks any organic discovery mechanism. Discord, covered in Section 6, sits between these two extremes: a developer-grade game surface plus an audience anchored by community membership.

E.3 Backbone-level swap inside the recovery pipeline (deferred)

A backbone-level swap inside the recovery pipeline is worth a future pass. Replacing the Hybrid Transformer Demucs variant with a domain-adapted music-vocals separator, or skipping separation in favour of a far-field music-aware recogniser, are both plausible variants. Both experiments are out of scope for the current data; the pipeline ran a single backbone configuration on the 25-video validation set and a backbone comparison would need a larger sample to support per-variant inference.