

A Multimodal Analysis of AI-Generated Cover Songs and Audience Reception

Michael Oehler
Osnabrück University
Osnabrück, Germany
michael.oehler@uos.de

Jasper Oldach
Osnabrück University
Osnabrück, Germany

William Shelor
Osnabrück University
Osnabrück, Germany

Tobias Rotsch
University of Music Trossingen
Trossingen, Germany

Abstract—Artificial intelligence (AI)-generated cover songs have become a visible form of generative music culture across digital platforms, where existing songs are reinterpreted through synthetic voices, voice clones, genre transfer, and platform-specific remix practices. This paper examines how such AI covers transform their source songs and how these transformations are reflected in audience reception. We construct a corpus of 311 AI-cover/original pairs linked to 63,301 YouTube comments and combine CLAP-based audio embeddings, prompt-based measures of cover-to-original audio-semantic change, and multilingual RoBERTa-based analysis of reception frames in comments. The results show that AI covers often preserve locally recognizable similarity to their source songs while shifting along specific semantic dimensions, including genre, intensity, arousal, and synthetic-versus-organic sound. These transformations are reflected selectively in comment discourse: concrete genre- and intensity-related audio shifts are associated with corresponding reception frames, whereas broader evaluative frames such as creativity, nostalgia, authenticity, and style-change recognition are not explained by global style distance. The findings suggest that audio-semantic features capture some perceptually salient transformations but do not by themselves account for broader cultural evaluations of AI covers. The paper provides a paired multimodal workflow for studying AI covers as transformations of existing works and as culturally evaluated objects in Internet of Musical Things ecosystems.

Index Terms—AI-generated music, cover songs, audience reception, Internet of Musical Things, CLAP, RoBERTa, YouTube comments, zero-shot classification

I. INTRODUCTION

Artificial intelligence (AI)-generated music has moved from a technical demonstration to a visible form of platform culture. Recent examples of AI-assisted songs, synthetic artist personae, and AI-generated reinterpretations show that generative music systems increasingly circulate through the same infrastructures as ordinary popular music: YouTube, TikTok, Spotify, playlists, reaction videos, and comment sections. Recent data-driven work on commercial text-to-music platforms such as Suno and Udio likewise shows that AI-generated music can be studied as a large-scale cultural practice, with prompts, tags, and lyrics revealing how users organize and describe generated songs [1]. Yet the public visibility of AI music is not driven only by novelty or by the affordances of generation platforms. Many prominent cases become legible by attaching AI generation to already familiar cultural materials: recognizable songs, established genres, famous voices, nostalgic repertoires,

or meme-compatible musical templates. AI-generated cover songs are therefore a particularly revealing case. They combine the recognizability of an existing work with the technical and cultural ambiguity of generative transformation.

Cover songs have always complicated simple distinctions between copying and creativity. In musicological and popular-music research, covers are not treated as mere duplicates of an original recording. They are understood as reperformances, reinterpretations, homages, parodies, appropriations, or dialogic responses to earlier musical works [2], [3]. From the perspective of music information retrieval, this ambiguity appears as the technical problem of cover song identification: a cover may preserve version-invariant properties of a song while changing timbre, instrumentation, structure, tempo, key, language, arrangement, or vocal delivery [4], [5]. Thus, the central issue is not simply whether two recordings are similar, but which aspects of musical identity remain stable and which are transformed.

AI-generated covers intensify this problem. Unlike conventional covers, they can simulate vocal identity, artist style, genre conventions, or production aesthetics without requiring a human performer to embody the transformation. This shifts the cultural question from “Is this the same song?” to “What kind of transformation is being performed, and how is it evaluated by listeners?” Prior work on AI music, deepfake voices, and synthetic performers suggests that audience responses are not reducible to a general rejection of AI. Comments and reception studies instead point to recurring tensions around authenticity, human-likeness, authorship, consent, copyright, ethical implications, artistic integrity, nostalgia, humor, technical quality, and creative reinterpretation [6]–[13]. Especially in cases involving recognizable voices or legacy artists, listeners may respond not only to the sound of the output, but also to perceived violations of identity, ownership, or artistic legitimacy.

At the same time, the technical analysis of AI covers requires caution. A specialized cover-identification system aims to recover work identity across version changes. By contrast, audio-text embedding models such as Contrastive Language–Audio Pretraining (CLAP) are designed to place audio and natural-language descriptions in a shared semantic representation space [14]. Such models are useful for describing stylistic, affective, genre-related, and production-related proximity, but they should not be treated as definitive detectors of musical

work identity. This distinction is important for AI covers: the same generated track may be locally recognizable as a version of the source song while globally shifting toward a different genre, affective profile, or production aesthetic. We therefore treat audio similarity and audio-semantic transformation as related but analytically distinct dimensions.

A second challenge concerns audience reception. Sentiment analysis alone is insufficient for AI-cover discourse. Prior studies of AI-generated or AI-mediated music comments show that positive and neutral reactions often dominate at the aggregate level, while more specific negative or ambivalent responses emerge around voice imitation, authenticity, ethics, and artistic value [7], [8], [10]. In other words, the relevant question is not only whether audiences like or dislike an AI cover, but what they take the cover to be: a creative reinterpretation, a humorous experiment, an uncanny imitation, a disrespectful use of a voice, a technically impressive artifact, or a legitimate musical version. YouTube comments are particularly useful for this purpose because they are platform-native traces of reception and often contain genre labels, affective descriptions, judgments of authenticity, and comparisons to the source artist or song.

Despite these converging strands, existing research remains fragmented. Cover song identification focuses on recognizing versions of the same work; AI music studies often focus on generation, legal controversy, detection, or isolated cases; and comment-based reception studies usually analyze textual discourse without systematically linking it to audio transformations. What is still missing is a paired, multimodal analysis that connects AI-generated covers to their concrete source songs, measures how the audio changes, and tests whether these changes are reflected in audience comments.

This paper addresses that gap by studying AI-generated cover songs on YouTube as cases of *interpretive distance*: the distance between an AI cover and its source song as both an audio-semantic transformation and a cultural transformation of meaning. We construct a corpus of AI-cover/original pairs, compute CLAP-based global and local similarity measures, derive prompt-based audio-semantic deltas for affect, production character, and genre/style, and link these features to multilingual RoBERTa-based reception frames extracted from YouTube comments. The analysis asks:

How do audio-semantic transformations between AI-generated covers and their source songs relate to YouTube comment reception, especially comments about genre recognition, energy, creativity, nostalgia, authenticity, and AI-related critique?

AI covers can also be situated within the Internet of Musical Things (IoMusT), which envisions networked devices and services for the production and reception of musical content that connect musicians and audiences [15], and within the broader Internet of Sounds (IoS) [16]. In this perspective, generative services act as remote production nodes, platforms as the reception layer, and comment sections as traces of audience–audience interaction. The present study contributes

to this agenda by linking content-based analysis of networked generative music to its stakeholder-level reception [17].

The paper makes three contributions. First, it proposes a paired multimodal workflow for analyzing AI covers in relation to their source songs and their platform reception. Second, it operationalizes local and global cover-to-source similarity through complementary segment-level and track-level CLAP measures. Third, it shows that audio–comment associations are selective: concrete genre- and energy-related shifts are reflected in corresponding comment frames, whereas broader evaluative frames such as creativity and nostalgia are not explained by global style distance alone. This suggests that CLAP-based audio-semantic features capture some perceptually salient transformations while leaving broader cultural evaluations only partly explained.

II. DATA AND METHODS

A. Corpus Construction

To construct the dataset, we first identified YouTube playlists that explicitly suggested a focus on AI-generated cover music. In a first retrieval stage, we used broad YouTube search queries such as “AI covers of popular songs,” “AI cover song,” and “AI cover music” and collected the first 150 playlist results. After removing playlists whose titles or contents were clearly misleading with respect to the study focus—for example, playlists labeled as AI although the audio itself was not AI-generated, or playlists that referred only to AI-generated artwork—134 playlists remained.

In a second retrieval stage, we complemented this broad search with artist-centered queries of the form “AI cover” plus artist name. We selected the five most popular artists on Spotify in 2025 and the first twenty entries from the Wikipedia list of best-selling music artists, accessed in April 2026. For each artist, we collected approximately one to two playlists, yielding an additional 37 playlists. The combined procedure produced an initial corpus of 171 YouTube playlists.

From these playlists, 14,214 individual song entries were extracted. Because many videos appeared repeatedly across playlists, all entries were deduplicated globally at the video level. Across the raw playlist pool, 3,158 entries were duplicates. In addition, 1,037 entries were already marked as private, 3,395 entries had become unavailable for reasons documented by YouTube or returned by the metadata extraction process, and 156 entries produced consistent but otherwise unspecified metadata-retrieval errors. After excluding these cases, 6,468 usable songs remained, uploaded by 1,987 distinct YouTube channels.

For each video, metadata were collected with a custom retrieval pipeline. The pipeline stored a full local JSON record for every successfully retrieved video and generated a compact spreadsheet representation containing title, description, uploader, channel, upload date, release date, view count, like count, comment count, categories, tags, and platform-provided descriptors. Error cases were retained with explicit status labels, ensuring that corpus reduction remained reproducible.

Because the analysis focused on publicly visible audience response, we filtered the dataset by comment volume. Across the 6,468 retrievable songs, the videos accumulated approximately 2.79 million comments. The average number of comments per video was 431, whereas the median was 14, indicating a strongly right-skewed engagement distribution. To retain only videos with observable reception, entries with fewer than 50 comments were removed, reducing the dataset to 1,890 songs.

We then restricted the corpus to videos uploaded on or after 1 October 2025. This cutoff was motivated by the release cycle of the two most frequently referenced music-generation systems in the corpus. Suno v5 and Udio Voices were introduced around this date and emphasized improved vocal and audio capabilities [18], [19]. The cutoff was also supported empirically: the corpus contained 231 uploads between 1 January and 30 September 2025, but 560 uploads between 1 October 2025 and 23 March 2026. These 560 videos formed the pre-classification dataset. Manual inspection and coding then retained 311 AI-cover/original pairs that referred to known existing songs and involved style-transfer or voice-plus-style-transfer transformations. The resulting sample should therefore be interpreted as a corpus of recent, visible, comment-rich AI-cover culture rather than as a representative sample of all AI-generated covers. Fig. 1 summarizes the complete corpus construction and analysis workflow.

As a descriptive source-material check, we summarized available MusicBrainz-derived style metadata for the original songs. The corpus was dominated by pop/mainstream material (40.7%), followed by rock/pop-rock/pop-punk (29.3%), metal/hard rock (17.3%), and hip-hop/rap (7.5%); soul/R&B/funk/gospel (2.3%), country/folk/singer-songwriter (2.0%), electronic/dance/synth (0.7%), and reggae/ska/island (0.3%) were rare. These broad categories collapse a more diverse multi-label tag space containing 74 distinct style terms across 705 tag assignments. The pop/mainstream category, for example, included OPM acoustic pop, boy-band pop, adult-contemporary ballads, dance-pop/funk-pop, regional pop, chanson/traditional pop, Turkish arabesk/popular ballads, and Sinhala pop/folk-pop. These categories are used as descriptive source-style proxies rather than validated genre annotations.

B. Audio Features

Audio features were designed to capture both source-song similarity and audio-semantic transformation. This distinction is important because cover-song identification aims to recover work identity across changes in timbre, instrumentation, arrangement, key, tempo, or vocal delivery [4], [5]. AI-cover reception, however, also depends on how the transformation is evaluated as imitation, reinterpretation, parody, homage, or problematic appropriation. We therefore use CLAP-based features as audio-semantic descriptors of transformation, not as a specialized cover-song identification system.

Audio was represented with LAION-CLAP embeddings [14]. LAION-CLAP was selected because it provides an

openly available audio-text embedding space suitable for prompt-based relative comparisons across heterogeneous music recordings. We do not assume that its embedding geometry is unique or model-independent. For each track, audio was converted to mono, resampled to 48 kHz, and segmented into 10-second windows with a 5-second hop. Segment embeddings were retained and then pooled to track-level representations. For every AI-cover/original pair, we computed two cosine-based similarity measures: a global similarity score using the mean-pooled track embeddings, and a local similarity score based on the mean of the five highest segment-level cosine similarities. The local score captures the possibility that an AI cover preserves recognizable passages even when the global audio-semantic profile diverges from the original.

To describe transformations, we projected track embeddings onto theory-guided prompt ensembles rather than assigning hard labels. The prompt inventory covers four feature families: similarity and recognizability, affective dimensions, production and human-likeness, and genre/style descriptors. The complete prompt ensembles are reported in Table I. Affective dimensions include valence and arousal, following dimensional models of affect [20], as well as music-specific descriptors such as darkness, aggression, and nostalgia [21]. The production dimension captures the contrast between synthetic or processed sound and organic, live, or human-sounding performance, which is particularly relevant for AI-generated cover reception.

The genre/style inventory should not be read as a set of stable, mutually exclusive acoustic classes. Genre combines sonic conventions, production practices, social labeling, market categories, and listener expectations [22]–[24], and MIR research has shown that genre classification is affected by ambiguity, subjectivity, dataset construction, and overlapping labels [25]–[27]. We therefore treat genre prompts as soft semantic coordinates rather than categorical ground truth.

For each prompt dimension, we computed an AI-minus-original delta. These deltas quantify relative movement in CLAP audio-text space and are interpreted as proxy measures of transformation, not as objective annotations of emotion, genre, or artificiality. We additionally computed a style-profile distance as the L2 norm over the style-delta vector.

The full feature inventory was used for descriptive analysis and clustering. For the main controlled models and compact results tables, we restricted attention to theoretically central and interpretable variables: global and local similarity, style-profile distance, valence, arousal, synthetic-versus-organic sound, darkness, aggression, nostalgia, and selected genre/style deltas. The selected genre/style deltas were chosen because they either showed substantial corpus-level shifts, corresponded to explicit comment-frame hypotheses, or were needed to interpret the main audio-comment associations. A four-cluster typology was then estimated from the full prompt-delta profile to summarize recurring transformation patterns.

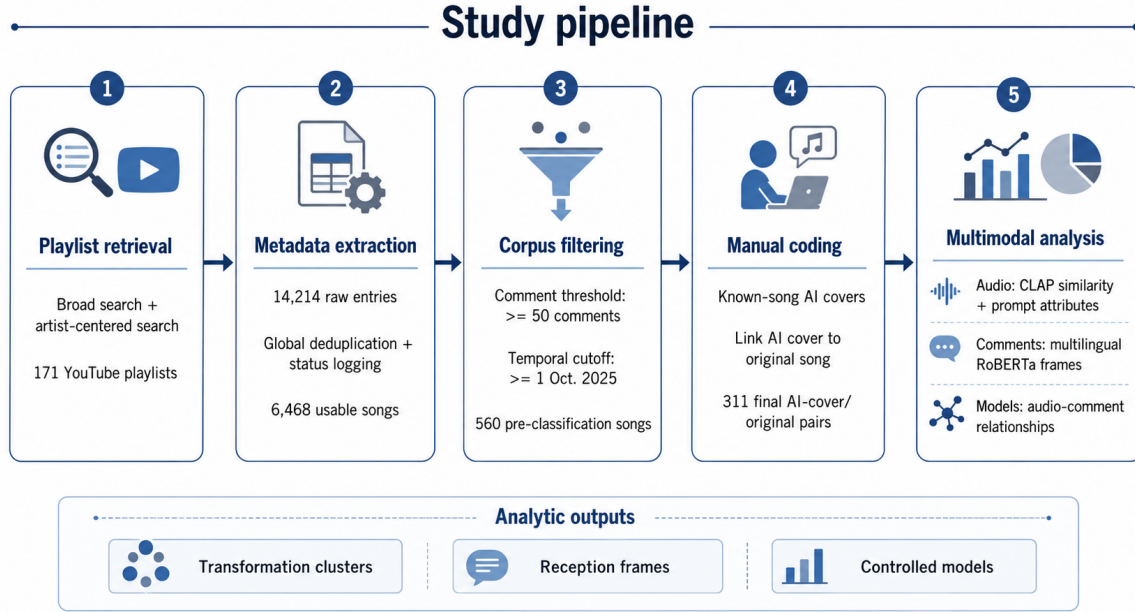


Fig. 1. Overview of the study pipeline. The workflow combines playlist retrieval, metadata extraction and status logging, corpus filtering, manual coding of known-song AI-cover/original pairs, and multimodal analysis of audio transformations and YouTube comment reception.

TABLE I
CLAP PROMPT ENSEMBLES USED FOR AUDIO-SEMANTIC ATTRIBUTE SCORING

Prompt set	Prompts
Valence positive	happy music; joyful music; uplifting song; cheerful music; positive mood music
Valence negative	sad music; melancholic music; gloomy song; depressing music; negative mood music
Arousal high	energetic music; intense music; powerful music; driving song; high energy music
Arousal low	calm music; relaxed music; soft music; peaceful song; low energy music
Dark	dark music; ominous music; moody music; dramatic dark song; sinister music
Aggressive	aggressive music; angry music; harsh music; heavy intense music; violent sounding music
Nostalgic	nostalgic music; sentimental music; retro sounding music; old fashioned song; music that feels like memories
Synthetic	synthetic music; artificial sounding music; processed vocals; digital production; AI generated music
Organic	organic music; natural vocals; live performance; acoustic sound; human sounding performance
Pop	pop music; modern pop song; commercial pop music
Rock	rock music; guitar rock song; rock band music
Metal	metal music; heavy metal song; distorted guitars and drums
Electronic	electronic music; synth electronic song; digital dance production
Acoustic	acoustic music; acoustic cover song; unplugged performance
Orchestral	orchestral music; cinematic orchestral music; symphonic arrangement
Folk/country	folk music; country music; acoustic folk song
Hip-hop/rap	hip hop music; rap song; trap beat music
Dance	dance music; club music; upbeat dance track
Lo-fi	lo-fi music; lofi chill music; bedroom pop production

C. Comment Features

The comment features were designed to test whether audio-semantic transformations are reflected in audience discourse and to capture broader interpretive reactions to AI covers. We therefore distinguish between audio-perceptual frames, which mirror the audio-feature families, and reception/evaluation frames, which capture cultural interpretation. This distinction follows prior work on AI-generated and AI-mediated music reception, where comments repeatedly revolve around authenticity, human-likeness, novelty, copyright and ethics, uncanny reactions, creativity, nostalgia, and the imitation of recognizable voices [7], [8], [10]–[12].

For each retained video, YouTube comments were pro-

cessed with multilingual RoBERTa-based models. Sentiment was estimated with a multilingual XLM-RoBERTa sentiment model. XLM-R is appropriate for this setting because it was trained for cross-lingual transfer across many languages and evaluated on multilingual benchmarks such as XNLI [28]. Reception frames were estimated with a multilingual XLM-RoBERTa large XNLI zero-shot classifier using the candidate labels listed in Table II. This model was selected because the corpus is multilingual and the analysis required a common zero-shot entailment framework across languages. The resulting frame scores should therefore be understood as model-dependent operationalizations rather than as human-validated annotations of audience meaning. In this setup, each

TABLE II
XLM-ROBERTA ZERO-SHOT CANDIDATE LABELS FOR COMMENT FRAMES

Frame	Candidate label	Frame group
Nostalgia/memory	nostalgia, memory, or remembering the past	Audio-perceptual
Energy/power	energy, power, intensity, or hype	Audio-perceptual
Dark emotional intensity	dark mood, sadness, haunting emotion, or emotional intensity	Audio-perceptual
Artificial/synthetic	artificial, synthetic, fake, robotic, or AI sounding	Audio-perceptual
Organic/human/natural	organic, human, natural, real, or live sounding	Audio-perceptual
Rock/metal genre	rock, metal, heavy guitars, or heavy music genre	Audio-perceptual
Electronic/dance genre	electronic, dance, club, synth, or EDM genre	Audio-perceptual
Hip-hop/rap genre	hip hop, rap, trap, or beat-based genre	Audio-perceptual
Acoustic genre	acoustic, unplugged, piano, guitar, or stripped-down arrangement	Audio-perceptual
Style-change recognition	recognition of a style change, genre change, cover version, or reinterpretation	Audio-perceptual
Creativity/reinterpretation	creativity, originality, reinterpretation, or artistic transformation	Reception/evaluation
Authenticity/legitimacy	authenticity, legitimacy, realness, or whether the cover is valid	Reception/evaluation
AI-uncanny critique	AI criticism, uncanny feeling, soullessness, creepiness, or discomfort	Reception/evaluation
Ownership/rights/consent	ownership, copyright, rights, consent, permission, or ethical concerns	Reception/evaluation
Respect original artist	respect for the original artist, tribute, homage, or honoring the source	Reception/evaluation
Voice identity/likeness	voice identity, vocal likeness, imitation of a singer, or sounding like someone	Reception/evaluation
Humor/playfulness	humor, joke, meme, parody, playfulness, or funny reaction	Reception/evaluation
Technical quality	technical quality, production quality, mixing, audio quality, or flaws	Reception/evaluation
Emotional response	personal emotional response, being moved, crying, chills, or goosebumps	Reception/evaluation

frame is formulated as a natural-language hypothesis, and each comment receives a continuous entailment-style score for that frame. This procedure is useful for a multilingual YouTube corpus because it avoids training a new supervised classifier for each frame while still allowing theory-guided categories to be operationalized through natural-language descriptions [29], [30].

The audio-perceptual frames capture whether commenters explicitly refer to dimensions that correspond to the CLAP audio features, including energy, affective intensity, synthetic or organic sound, genre recognition, and style-change recognition (Table II).

The reception/evaluation frames capture broader interpretations of AI covers as cultural artifacts, including creativity, authenticity, AI-related critique, rights and consent, voice identity, humor, technical quality, and emotional response (Table II). These frames are not direct acoustic descriptors. Rather, they operationalize how audiences evaluate the status of AI covers: as creative transformations, playful artifacts, technically impressive outputs, uncanny imitations, ethically problematic uses of voice identity, or legitimate musical versions. This distinction is motivated by prior work showing that sentiment alone is insufficient for AI-music reception, where responses may become critical around recognizable voices, authenticity, authorship, or artistic integrity [8], [10], [12].

As with the audio features, the comment-frame inventory was intentionally broader than the final model set. The broader inventory allowed us to inspect which reception categories were empirically usable and which suffered from ceiling or floor effects. For the main statistical models, we therefore use continuous AI-cover-upload-level frame means rather than only binary hit rates. Thresholded frame shares, count outcomes, and binomial-rate models are retained as robustness checks for selected categories, but the primary interpretation is based on continuous frame means because they preserve more information and are less sensitive to arbitrary threshold choices.

The transition from the broader candidate inventory to the reported results followed a hypothesis-directed screening logic. All frames listed in Table II were scored for individual comments and then aggregated for each AI-cover upload, but the main results focus on frame–audio combinations with an explicit theoretical mapping to the CLAP features and sufficient empirical variation. Frames with severe ceiling or floor effects were retained for descriptive diagnostics and robustness checks, but were not foregrounded as primary outcomes. Thus, the reported effects are not a post-hoc subset of arbitrary significant associations; they are selected from the pre-specified audio–comment mapping and interpreted together with theoretically relevant null results.

D. Statistical Models

The main models are ordinary least squares regressions with HC3 robust standard errors. Audio predictors were standardized, so coefficients indicate the expected change in the outcome for a one-standard-deviation increase in the audio feature. Models controlled for comment volume, transformation type, and the exploratory transformation cluster. Because the cluster variable is derived from the same prompt-delta space as several audio predictors, it is used only as a descriptive control and the corresponding estimates are interpreted cautiously. For thresholded count outcomes, we estimated Poisson and negative-binomial models with an offset for $\log(n_{\text{comments}})$, as well as binomial-rate models for successes out of total comments. P-values were corrected using the Benjamini–Hochberg false discovery rate procedure within each model family [31].

III. RESULTS

A. Corpus and Analytic Coverage

The final analytic sample consisted of 311 AI-cover/original pairs. These videos contained 63,301 comments in total, with a median of 99 comments per video and a mean of 204 comments per video. Thus, the sample retained sufficient

TABLE III
SELECTED AUDIO TRANSFORMATION MEASURES FOR 311
AI-COVER/ORIGINAL PAIRS

Measure	Mean	Median	P25	P75
Global CLAP similarity	0.701	0.708	0.618	0.786
Top-5 segment similarity	0.808	0.831	0.761	0.868
Style-profile distance	0.319	0.321	0.224	0.406
Δ metal	0.070	0.053	-0.048	0.232
Δ hip-hop/rap	0.047	0.061	-0.017	0.110
Δ aggressive	0.041	0.027	-0.056	0.176
Δ rock	0.033	0.040	-0.061	0.131
Δ dark	0.029	0.024	-0.026	0.077
Δ arousal	-0.007	-0.003	-0.057	0.043
Δ synthetic-organic	-0.040	-0.038	-0.083	0.006

comment volume for reception analysis while still reflecting a skewed engagement distribution.

Table III reports a compact subset of audio variables used to characterize the corpus and interpret the main models. It includes the three main similarity/transformation-summary measures, the largest or substantively relevant mean prompt-delta shifts, and variables central to the subsequent audio-comment analyses. Other prompt attributes were computed and retained in the full feature inventory and clustering.

The prompt-based deltas show that AI covers were not simply shifted toward a generic “AI-like” sound. The strongest mean positive shifts were toward metal, hip-hop/rap, aggression, rock, and darkness, while the synthetic-versus-organic contrast decreased. In CLAP audio-text space, AI covers were therefore on average closer to organic or human-sounding prompts relative to the originals, rather than more synthetic. Because hip-hop/rap recognition showed floor effects in the comment features, the corresponding audio-side shift is treated descriptively rather than foregrounded in the audio-comment models.

B. Transformation Typology

The four-cluster solution produced an interpretable but exploratory typology (silhouette = 0.191). The largest cluster was a metal/rock intensification profile ($n = 104$), characterized by large positive deltas for metal, aggression, rock, and arousal. The second cluster ($n = 94$) consisted of conservative or low-shift covers with deltas close to zero. A third cluster ($n = 60$) captured darker, more nostalgic, and more acoustic reinterpretations. The fourth cluster ($n = 53$) captured pop, electronic, hip-hop, dance, and lo-fi reformatting, typically away from rock/metal and aggression.

As shown in Fig. 2, the clusters occupy partly distinct regions of the prompt-delta space, although the typology remains exploratory rather than a hard classification. The main empirical variation lies in the type of transformation: some covers intensify genre and energy, some retain the source profile, and others reformat the song into a different stylistic register.

C. Comment Reception Features

The comment corpus contained predominantly positive or neutral sentiment. The mean continuous sentiment score was 0.355, with an average positive-label share of 0.459 and negative-label share of 0.050. Several broad zero-shot reception frames showed high activation across much of the corpus. In particular, voice identity, humor/playfulness, respect for the original artist, and nostalgia produced ceiling or low-variance patterns. Conversely, AI-uncanny critique, ownership/rights concerns, hip-hop/rap recognition, and negative sentiment showed floor effects.

These diagnostics motivated the use of continuous frame means as the primary outcomes in the controlled models. Thresholded shares and count variables are informative for robustness, but they are more sensitive to the choice of zero-shot threshold and to ceiling/floor artifacts.

D. Audio Transformations Selectively Predict Comment Reception

The controlled models show that comment reception is linked selectively to specific audio transformations rather than to overall transformation magnitude. Fig. 3 summarizes the main OLS-HC3 effects used for interpretation. Positive coefficients indicate that stronger audio shifts are associated with higher activation of the corresponding comment frame; negative coefficients indicate lower activation. Table IV reports the corresponding OLS-HC3 estimates. The selected effects should be read as theoretically targeted associations rather than as a complete model-selection exercise; the full model grid is available from the corresponding author upon request. Because the outcomes are continuous frame means, coefficient magnitudes should be interpreted relative to the empirical dispersion of each outcome. For the OLS models shown in Table IV, outcome standard deviations range from 0.037 to 0.065 and model R^2 values range from 0.038 to 0.133, indicating modest but directionally specific associations.

The clearest positive association concerns a correlated genre/intensity bundle. A one-standard-deviation increase in the metal delta was associated with a 0.016 increase in the rock/metal comment-frame mean ($p_{\text{FDR}} = .004$). The same outcome was also associated with aggressive shift ($\beta = 0.013$, $p_{\text{FDR}} = .012$) and rock shift ($\beta = 0.012$, $p_{\text{FDR}} = .022$). Because metal, rock, and aggressive deltas are strongly correlated, these estimates should be interpreted as one genre/intensity-related transformation pattern rather than as three independent confirmations. The result is best understood as evidence that concrete audio-side style changes can be linguistically registered in corresponding comment frames.

A second pattern concerns arousal and energy. Variation in the arousal delta was associated with both the energy/power frame ($\beta = 0.017$, $p_{\text{FDR}} = .038$) and the emotional-response frame ($\beta = 0.012$, $p_{\text{FDR}} = .011$). Because the mean arousal delta was close to zero, this result should be interpreted as a cross-pair association rather than as evidence for a systematic arousal increase across the corpus.

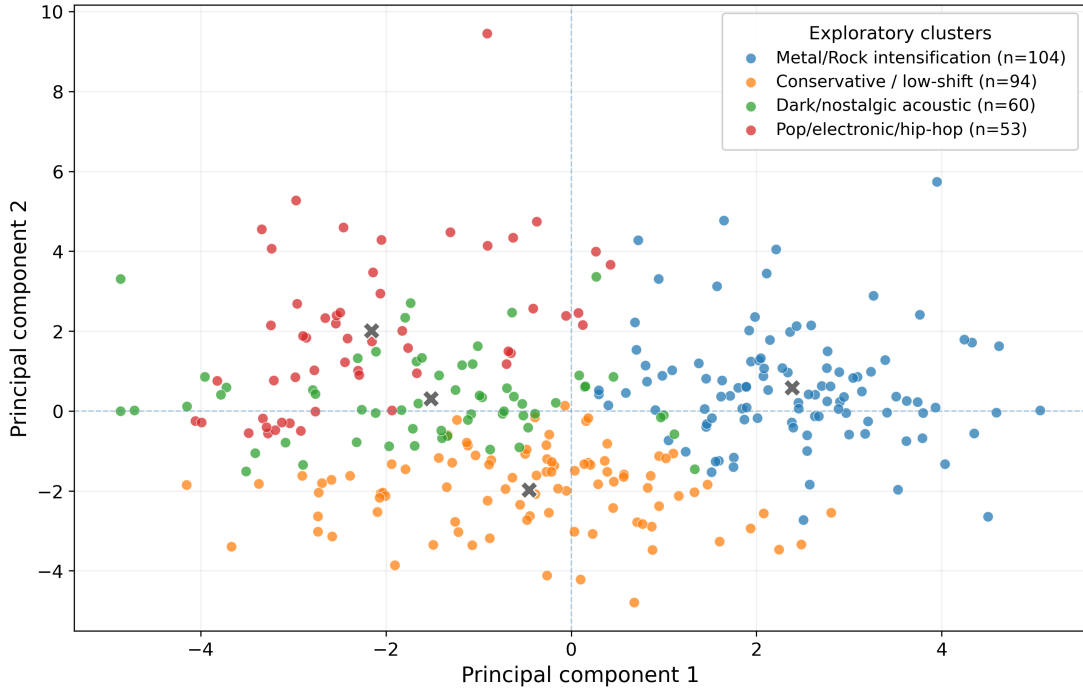


Fig. 2. Exploratory transformation clusters in CLAP prompt-delta space. Points represent AI-cover/original pairs projected onto the first two principal components of the prompt-attribute delta features. Colors indicate four k-means clusters; crosses mark cluster centroids.

Aggressive transformation was also inversely associated with the dark-emotional frame ($\beta = -0.013$, $p_{\text{FDR}} = .022$). This suggests that, in this corpus, genre-coded aggression or heaviness was not generally discussed as darker affective expression; instead, it appears to align more closely with genre recognition than with dark-emotional language.

Aggressive transformation additionally showed a negative association with the AI-uncanny critique frame ($\beta = -0.015$, $p_{\text{FDR}} < .001$). The corresponding negative-binomial robustness model estimated a lower rate of AI-uncanny comments for aggressive transformations (incidence-rate ratio = 0.797, 95% CI [0.722, 0.879], $p_{\text{FDR}} = .002$). This count-model result is treated as a robustness check rather than as primary evidence. The pattern is consistent with the interpretation that genre-coded aggressive or metal-like covers may be discussed less as uncanny AI imitation and more as recognizable stylistic reinterpretations; however, it remains observational and may also reflect contextual differences between videos.

The controlled models therefore do not support a simple “more transformation equals more creativity” account. The global style-profile distance did not predict creativity/reinterpretation comments and did not predict recognition of style change. Likewise, the CLAP nostalgia delta did not predict nostalgia comments. These null findings are important because they delimit what the audio features explain: concrete genre- and energy-specific transformations are reflected in reception, whereas broader interpretive frames are not straightforwardly reducible to audio distance.

E. Robustness and Collinearity

The model grid produced 289 OLS-HC3 models, 238 Poisson offset models, 238 negative-binomial offset models, and 238 binomial-rate models. After FDR correction within model families, 23 OLS effects, 35 Poisson effects, 12 negative-binomial effects, and 29 binomial-rate effects remained significant. Table IV reports the subset of theoretically targeted effects used for interpretation.

Predictor diagnostics showed substantial collinearity among several audio features. For example, aggressive and metal deltas were strongly correlated ($\rho = .952$), global similarity and top-5 segment similarity were also highly correlated ($\rho = .894$), and global similarity was negatively correlated with style-profile distance ($\rho = -.786$). For this reason, the models are interpreted as controlled, hypothesis-directed screening models rather than as a single fully saturated multivariate explanation. The collinearity is also substantively meaningful: in this corpus, metal, rock, and aggressive transformations often occur together. The exploratory cluster variable should likewise be interpreted cautiously because it is derived from the same prompt-delta profile as several predictors and because the cluster structure is modest (silhouette = 0.191).

IV. DISCUSSION

This study examined AI-generated covers as cases of interpretive distance between an existing source song and a generated reinterpretation. The local and global CLAP measures provide complementary descriptive views of cover-to-source similarity, although they were strongly correlated in this corpus

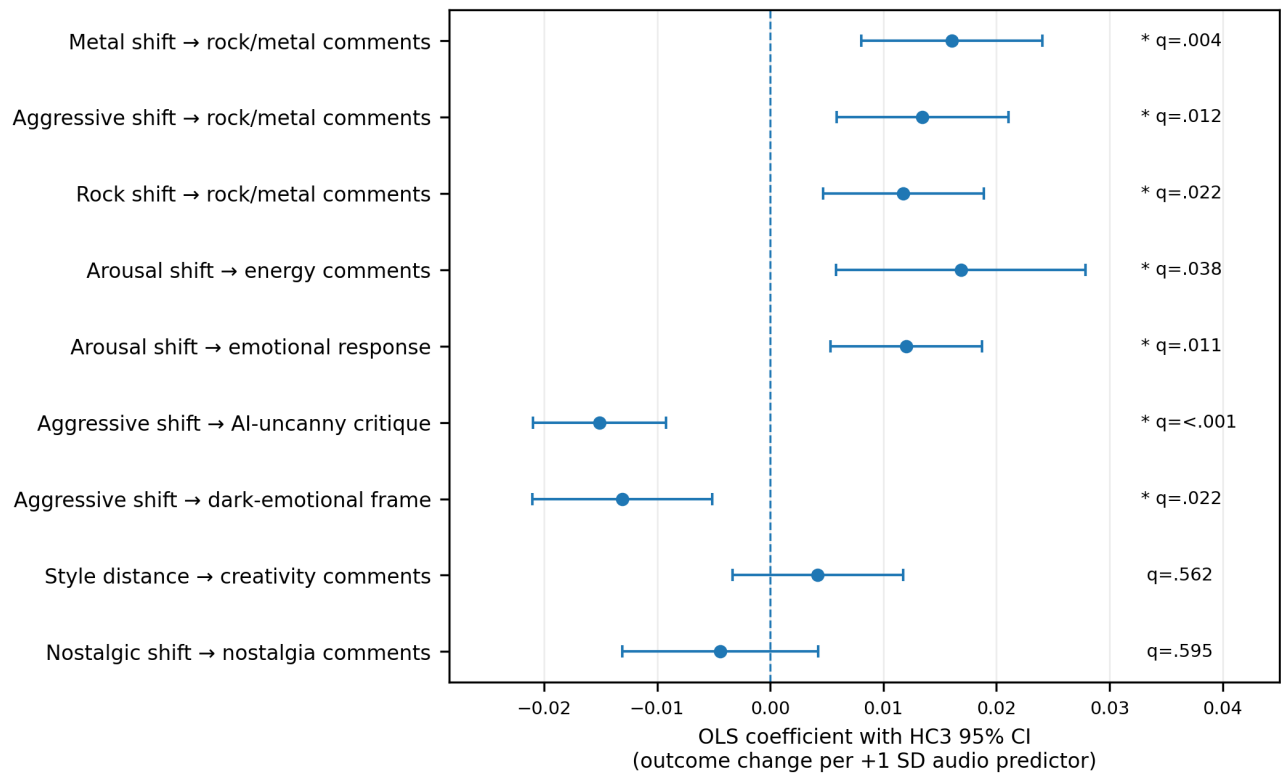


Fig. 3. Main controlled audio-comment model effects. Points show OLS-HC3 coefficients for continuous comment-frame outcomes; horizontal lines indicate 95% confidence intervals. Audio predictors were standardized, so estimates represent the expected change in the comment-frame mean for a one-standard-deviation increase in the audio predictor. Asterisks mark effects significant after Benjamini-Hochberg FDR correction.

TABLE IV
SELECTED OLS-HC3 AUDIO-COMMENT MODELS FOR CONTINUOUS FRAME MEANS ($n = 311$)

Outcome	Audio predictor	Est.	SE	p_{FDR}	R^2	SD
Rock/metal frame mean	Δ metal	0.016	0.004	.004	.073	.047
Rock/metal frame mean	Δ aggressive	0.013	0.004	.012	.062	.047
Rock/metal frame mean	Δ rock	0.012	0.004	.022	.056	.047
Energy/power frame mean	Δ arousal	0.017	0.006	.038	.079	.065
Emotional-response frame mean	Δ arousal	0.012	0.003	.011	.060	.052
Dark-emotional frame mean	Δ aggressive	-0.013	0.004	.022	.075	.049
Creativity/reinterpretation mean	Style-profile distance	0.004	0.004	.562	.064	.060
Nostalgia frame mean	Δ nostalgic	-0.004	0.004	.595	.038	.050
AI-uncanny critique mean	Δ aggressive	-0.015	0.003	< .001	.133	.037

($p = .894$). The results therefore do not support treating them as independent dimensions, but the segment-level measure remains useful descriptively for identifying locally similar passages within broader audio-semantic transformations. This pattern is consistent with cover-song and version-identification research, which distinguishes work-preserving properties from version-specific transformations [4], [5].

The clearest reception results concern concrete genre- and intensity-related transformations. Covers that moved toward the correlated metal/rock/aggression bundle in CLAP prompt space were more likely to elicit rock/metal-related comment frames. Likewise, variation in arousal shift was associated with energy/power and emotional-response frames. These find-

ings should not be overread as independent effects of metal, rock, and aggression; rather, they indicate that some concrete perceptual transformations captured by the audio embeddings are also linguistically registered in platform-native comment discourse. The explanatory power of the corresponding models was modest ($R^2 = .038-.133$), indicating that audio-semantic transformation accounts for only a limited share of variation in comment reception. The practical significance of the positive associations should therefore be interpreted cautiously. Part of the association between audio-side genre shifts and genre-related comments may reflect platform framing rather than auditory perception alone, as video titles or descriptions may explicitly label a transformation as, for example, a “metal

version.”

The absence of associations for style-profile distance and nostalgia further delimits what audio-semantic features can explain. Broad evaluative frames such as creativity, nostalgia, legitimacy, or reinterpretation appear not to follow directly from acoustic or style distance alone. They likely depend on additional contextual cues, including the identity of the source artist, the recognizability of the voice, video titles and descriptions, audience expectations, fan cultures, platform circulation, and whether the AI status of the cover is known or salient. This is consistent with prior work on AI music reception, where authenticity, authorship, and ethical concerns become especially salient in relation to recognizable human voices or legacy artists [8], [10], [12].

A further finding concerns the synthetic-versus-organic dimension. On average, AI covers were not shifted toward synthetic prompt regions relative to their originals. Instead, the synthetic-versus-organic contrast decreased, suggesting that many covers became closer to organic, human, or live-sounding prompt regions in CLAP space. This result complicates the assumption that AI-generated music is primarily characterized by audible artificiality. It may indicate that current systems often produce stylistically coherent genre versions rather than a uniformly synthetic sound, but it may also partly reflect how CLAP maps dense, polished, or human-like production cues onto organic descriptors. The negative association between aggressive transformations and AI-uncanny critique should therefore be read cautiously: genre-coded aggressive covers may be discussed less as uncanny AI imitation, but contextual and platform factors may also contribute.

Methodologically, the study demonstrates the value of treating CLAP prompt scores as relative, interpretable proxy measures rather than as hard genre, affect, or authenticity labels. This is particularly important for genre. Music genres are socially and historically unstable categories, and MIR research has repeatedly shown that genre classification is affected by ambiguity, dataset construction, and overlapping labels [22]–[26]. The prompt-delta approach used here therefore does not classify a cover as “metal” or “acoustic” in an absolute sense. It measures whether the AI cover moves closer to or farther away from such descriptors relative to its source song. This relative design is well suited to AI covers, where the research object is not the generated track in isolation, but the transformation between a source version and a new version.

The comment-analysis pipeline follows the same logic. Multilingual RoBERTa-based frame scores are not treated as definitive annotations of audience meaning. They are scalable indicators of whether comments express theoretically motivated reception frames. The ceiling and floor effects observed for several frames show why this caution is necessary. Some broad frames, such as voice identity, humor, respect for the original artist, and nostalgia, were activated across much of the corpus, while others, such as ownership/rights concerns or AI-uncanny critique, were comparatively rare. This supports the decision to use continuous frame means as the primary

outcomes and thresholded shares or count models only as robustness checks.

Several limitations remain. First, the corpus is platform-specific and based on YouTube playlist retrieval. It captures visible, recent, and comment-rich AI-cover culture rather than the full population of AI-generated covers. Second, the pairing of AI covers and originals relies on manual coding of known-song references and transformation types; this improves construct validity but limits scale and should be validated with independent coders in future work. Third, CLAP is not a specialized cover-identification system. Its strength in this study lies in modeling audio-semantic transformation, not in proving musical work identity, and the prompt-based deltas should be treated as proxy measures rather than validated perceptual annotations. The audio-side findings are therefore also conditional on the representation geometry of the selected LAION-CLAP model and should be examined with alternative audio-text embedding models in future work. Fourth, the zero-shot comment frames were not validated against human annotations and depend on prompt design, model choice, and model calibration. Their interpretation is therefore conditional on the selected XLM-R-based zero-shot framework, and multilingual comments may contain irony, code-switching, slang, or culturally specific references that are only imperfectly captured. Fifth, the statistical models are observational and should not be interpreted causally. Audio transformations are associated with comment frames, but the comments may also be shaped by source-song genre and era, metadata, uploader communities, video popularity, recommendation dynamics, and prior knowledge about the original artist.

Future work should combine stricter work-identity measures from cover song identification with audio-text embeddings, incorporate platform framing such as titles, descriptions, tags, thumbnails, and channel metadata, and validate CLAP prompt deltas and comment-frame scores against manually annotated subsets. It should also incorporate source-song genre, era, language distribution, and permutation baselines for audio-comment pairings. These extensions would help separate what listeners hear, what they know from platform context, and how they culturally evaluate the legitimacy of the AI cover.

V. CONCLUSION

AI-generated covers are neither simple copies of existing songs nor merely generic examples of AI music. They are transformations of recognizable works, and their reception depends on which aspects of the source are preserved, which are changed, and how these changes become meaningful to listeners. This paper proposed a paired multimodal approach for studying this problem by linking AI-cover/original audio features to YouTube comment reception.

The results show that AI covers retain substantial cover-to-source similarity while also exhibiting specific shifts in audio-semantic profile. The most interpretable positive associations involved concrete genre and intensity frames: a correlated metal/rock/aggression bundle and arousal variation were reflected in corresponding comments about rock/metal, energy,

and emotional response. By contrast, broader evaluative frames such as creativity and nostalgia were not associated with global style distance or with the corresponding CLAP prompt deltas. This indicates that audience reception is not simply a function of how far an AI cover moves away from the original. Rather, reception depends on which transformation dimensions become salient and how they interact with genre conventions, voice identity, and platform context.

The study therefore contributes to AI music research by connecting three levels that are often studied separately: the source song, the generated audio transformation, and the audience discourse surrounding the resulting cover. It also shows that CLAP-based prompt deltas and multilingual RoBERTa frame analysis can be used as scalable proxy measures for studying AI-cover culture, provided that they are interpreted cautiously and supported by transparent prompt inventories and model outputs. The central finding is that AI-cover reception is selective: specific audio-semantic transformations are recognized in comments, while broader cultural evaluations remain only partly explained by audio features alone. Understanding AI covers therefore requires moving beyond overall similarity toward a multidimensional account of interpretive distance.

DATA AND CODE AVAILABILITY

The analysis scripts, prompt inventories, derived feature tables, model outputs, and figure-generation code are available from the corresponding author upon request. Because the corpus is based on YouTube material, shared data are limited to derived metadata, feature tables, model outputs, and references to public video identifiers where permissible. We analyzed publicly available YouTube comments, did not report usernames or quote individual comments, and do not redistribute raw comment text.

REFERENCES

- [1] L. Casini, L. Cros Vila, D. Dalmazzo, A.-K. Kaila, and B. L. Sturm, "Data-driven analysis of text-conditioning in ai-generated music: A case study with suno and udio," *Transactions of the International Society for Music Information Retrieval*, vol. 9, no. 1, pp. 194–209, 2026.
- [2] K. Mosser, "Cover songs: Ambiguity, multivalence, polysemy," *Popular Musicology Online*, vol. 2, 2008.
- [3] G. Plasketes, Ed., *Play It Again: Cover Songs in Popular Music*. Farnham: Ashgate, 2010.
- [4] J. Serra, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," in *Proceedings of the Audio Engineering Society Convention*, 2008.
- [5] M. Li and N. Chen, "A robust cover song identification system with two-level similarity fusion and post-processing," *Applied Sciences*, vol. 8, no. 8, p. 1383, 2018.
- [6] P. Gałuszka, "The influence of generative ai on popular music: Fan productions and the reimagining of iconic voices," *Media, Culture & Society*, vol. 47, no. 3, 2025.
- [7] V. Lac and F. T. Moura, "Sentiment analysis of ai generated music using latent dirichlet allocation (lda)," in *Proceedings of the International Conference on AI Music Creativity (AIMC)*, 2024.
- [8] F. T. Moura and V. Lac, "Deepfake music and listener sentiments: A large-scale analysis of youtube comments," in *Proceedings of AIMEDIA 2025*, 2025.
- [9] M. Oehler, J. Oldach, and F. Zwißler, "Perceiving ai in music: Human evaluation of ai-generated melodies and ai detection sensitivity," in *Proceedings of the 1st AES International Conference on Machine Learning and Artificial Intelligence for Audio (AES AIMLA 2025)*, 2025, pp. 1–9.
- [10] R. Ünal and A. Taylan, "Ai covers: Listener perspectives on a controversial production model between technology, copyright and ethical violations," *Ankara Üniversitesi İlefl Dergisi*, vol. 12, no. 2, pp. 216–254, 2025.
- [11] D. Ching, J. Twomey, C. Linehan, and G. Murphy, "Encountering deepfakes: A thematic analysis of comments on a music video featuring deepfake content," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 7, pp. 1–23, 2025.
- [12] A. Pfeiffer and N. Krishna, "Ai in music reconstruction: Sentiment in pro versus amateur contexts," *Journal of Cultural Management and Cultural Policy*, vol. 11, no. 2, pp. 256–268, 2025.
- [13] M. Oehler, J. Schepmann, and B. Saurbier, "Assessing expectancy bias in listener evaluations of ai and human music compositions," in *Proceedings of the 6th Conference on AI Music Creativity (AIMC 2025)*, 2025, pp. 1–11.
- [14] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
- [15] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of musical things: Vision and challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.
- [16] L. Turchet, M. Lagrange, C. Rottondi, G. Fazekas, N. Peters, J. Østergaard, F. Font, T. Bäckström, and C. Fischione, "The internet of sounds: Convergent trends, insights, and future directions," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 264–11 292, 2023.
- [17] L. Turchet, "Entangled internet of musical things and people: A more-than-human design framework for networked musical ecosystems," *IEEE Transactions on Technology and Society*, vol. 5, no. 4, pp. 355–367, 2024.
- [18] Suno, "Introducing v5," Online, Oct. 2025, accessed: 2026-05. [Online]. Available: <https://help.suno.com/en/categories/1453377-introducing-v-5>
- [19] Udio, "Udio changelog: Voices," Online, Sep. 2025, accessed: 2026-05. [Online]. Available: <https://help.udio.com/en/articles/12055317-voices>
- [20] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [21] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.
- [22] F. Fabbri, "A theory of musical genres: Two applications," in *Popular Music Perspectives*, D. Horn and P. Tagg, Eds. Gothenburg and Exeter: International Assoc. for the Study of Popular Music, 1982, pp. 52–81.
- [23] S. Frith, *Performing Rites: On the Value of Popular Music*. Cambridge, MA: Harvard University Press, 1996.
- [24] F. Holt, *Genre in Popular Music*. Chicago: University of Chicago Press, 2007.
- [25] J.-J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.
- [26] B. L. T. Sturm, "The "horse" inside: Seeking causes behind the behaviours of music content analysis systems," *Computers in Entertainment*, vol. 14, no. 2, pp. 1–32, 2017.
- [27] J. Ramírez and M. J. Flores, "Machine learning for music genre: multifaceted review and experimentation with audioset," *Journal of Intelligent Information Systems*, vol. 55, no. 3, pp. 469–499, 2020.
- [28] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451.
- [29] A. Conneau, R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "Xnli: Evaluating cross-lingual sentence representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2475–2485.
- [30] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 3914–3923.
- [31] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.