

Mining Multimodal Patterns: Interpretable Video-Audio Relationships for Dataset Characterization and Semantic Retrieval

Franklin Sales de Oliveira¹, Diognei de Matos¹, Caio Henrique de Miranda Onofre¹
franklin.sales@dcc.ufmg.br, diogneimatos@dcc.ufmg.br, caionofre@ufmg.br

¹Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil

ABSTRACT

The recent surge in large-scale multimodal datasets has advanced various video-audio tasks, from semantic retrieval to video-to-music generation; however, the explicit semantic rules governing how visual environments dictate musical styles remain poorly understood. This paper presents a generalizable, interpretable descriptive learning pipeline designed for multimodal dataset characterization, utilizing the generative GVMGEN benchmark as a primary case study. Because symbolic video representations suffer from inherent spatial sparsity, traditional mining algorithms succumb to sparsity bias, generating uninformative heuristics based on visual absences. To address this, our approach introduces a spatial relaxation layer that translates continuous visual embeddings into a discrete, binary Bag of Visual Words (BoVW), effectively eliminating temporal and positional constraints. We then combine global frequent pattern mining via FP-Growth with a positive-literal Constrained Beam Search to discover pure, discriminative visual signatures associated with diverse acoustic genres. The extracted semantic heuristics were validated in a downstream Semantic Video Retrieval task, where they operated as low-overhead Boolean filters, achieving a Retrieval Lift of up to $2.91\times$ over blind search. Our findings provide an interpretable characterization of multimodal topologies, offering actionable prior knowledge that paves the way for robust multimedia indexing, general video-audio retrieval, and semantically-aware generative applications.

KEYWORDS

Descriptive Learning, Multimodal Characterization, Subgroup Discovery, Semantic Retrieval, Video-to-Music

1 INTRODUCTION

Recent advances in multimodal artificial intelligence have accelerated research in video-to-music generation. The availability of large-scale datasets containing synchronized video and audio has enabled the development of generative models capable of learning complex mappings between visual and musical content [5]. However, the increasing scale and stylistic variety of these datasets create deep challenges for data understanding and structural interpretability [24]. Because the

semantic cross-modal relationships often remain unmapped and opaque, analyzing or validating recurring stylistic patterns across modalities becomes difficult [21]. Consequently, formal descriptive methodologies are required to interpret and characterize video-to-music data.

In recent years, descriptive learning and data mining have advanced significantly due to the growing demand for explainable multimedia systems [22]. Unlike predictive approaches focused strictly on optimization or classification accuracy, descriptive learning seeks to identify human-understandable patterns, local anomalies, and hidden global structures. Algorithms for frequent pattern mining and Subgroup Discovery (SD) have shown strong potential for extracting symbolic knowledge from complex, high-dimensional spaces.

However, when applied to multimodal representations, standard descriptive tools often fail due to the inherent spatial and temporal sparsity of raw video features. In a traditional frame-by-frame description, a concept appearing in different time windows is treated as distinct, creating a sparse matrix that causes traditional pattern miners to succumb to sparsity bias, often generating uninformative negative rules (e.g., basing a musical genre definition on the arbitrary absence of a visual element).

To address these challenges, this work introduces an interpretable descriptive pipeline tailored for the GVMGEN dataset [24]. We propose a framework that applies spatial relaxation to map continuous video embeddings into a binary Bag of Visual Words (BoVW) representation [19], eliminating positional constraints. We combine global co-occurrence exploration via FP-Growth with a localized, positive-literal constrained Beam Search strategy to extract pure, discriminative visual signatures for each musical genre. We also validate the utility of these discovered rules by deploying them as low-overhead Boolean filters in a downstream Semantic Video Retrieval task, demonstrating significant precision gains without the need for dense neural inference at execution time.

2 RELATED WORK

Recent architectures in video-to-music generation heavily focus on enhancing cross-modal alignment between visual backdrops and generated audio tracks. The OSSL dataset introduces curated film clips paired with soundtracks and explicit mood annotations, deploying a video-conditioned adapter to guide text-to-music models like MusicGen [5]. The GVMGEN framework addresses weak correspondence through a hierarchical attention mechanism that aligns deep

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

video and music representations, establishing a large-scale paired benchmark [24]. Similarly, VidMuse focuses on large-scale video-conditioned generation utilizing 360K video-music pairs and a framework that combines local and global visual cues [21]. While these studies push the boundaries of generation quality, they treat the dataset as an uninterpretable black box, leaving cross-modal semantic correlations unmapped.

Frequent pattern mining and SD provide the foundational mathematical tools to extract interpretable rules from complex data. FP-Growth [3] introduced an efficient tree-based structure for mining frequent itemsets without candidate generation, dramatically improving global scalability over Apriori-based approaches. In parallel, the formalization of rule evaluation metrics, such as Weighted Relative Accuracy (WRAcc) [6], provided a balanced trade-off between pattern coverage and target concentration, shaping the field of SD.

Algorithms like CN2-SD [7] leveraged heuristic beam search to discover high-quality localized rules. A comprehensive overview by Herrera et al. [4] unified these search strategies and quality metrics across nominal and numeric domains.

While foundational algorithms of SD and frequent pattern mining have been well-established, their direct application to cross-modal and multimedia contexts has been sparse but notable. Early efforts in video data mining successfully applied standard association rule mining to filter semantic concepts in raw video frames [10], and cross-modal rules have been explored for bridging vision and language in visual storytelling [9]. Within the acoustic domain, SD has been utilized to extract descriptive rules characterizing regional folk music and compositional styles [20]. While contemporary approaches in Semantic Video Retrieval frequently leverage dense, end-to-end neural embeddings [11], these models tend to act as opaque ranking systems, which can present interpretability challenges and increase inference demands. Our approach bridges this gap by reintroducing declarative SD into state-of-the-art multimodal generative topologies.

In this work, we build upon these classical data mining foundations but introduce a crucial methodological intervention. Rather than exposing the pattern miners to raw, sparse multi-column frame annotations, we introduce a semantic relaxation layer paired with strict algorithmic search constraints, eliminating rule redundancy and preventing the generation of uninformative negative literals in sparse multimedia representations.

3 METHODOLOGY

This section presents the methodology used to describe the dataset. First, we present details about the dataset used, how it was prepared, the extraction of visual and acoustic features, and the translation of these features into a human-interpretable format. Then, we present the descriptive algorithms used and how they were implemented.

3.1 Dataset Preparation

Figure 1 shows the methodology we used for dataset preparation. The algorithm’s input is a list containing links to the

original videos. First, the algorithm downloads the videos, divides them into 30-second segments, and separates them into multiple .mp4 and .wav files. Next, each video goes through a video feature extractor, and each audio file goes through an audio feature extractor, generating numerical features for the videos and audio files, respectively. Then, each file of numerical video and audio features goes through a feature translator that generates textual descriptions for the video and audio files, respectively. Finally, the results are concatenated into lists describing the visual and acoustic data, one list for each video-audio pair.

3.1.1 Dataset download and splitting. The GVMGEN dataset, recently introduced by Zuo et al. [24], is specifically designed for the task of automated video-to-music generation. As the dataset has not yet been fully released in public repositories, access to the raw multimedia files was secured through direct correspondence with the original authors, who provided the source links hosted on the Bilibili video-sharing platform ¹.

The collection comprises about 147 hours of footage. While the temporal distribution consists predominantly of videos under 10 minutes in length, it also includes extended outliers exceeding one hour. The visual content is highly diverse, encompassing movies, video blogs (vlogs), animated comics, and documentaries. Acoustically, the dataset is strictly curated for musicality; the audio tracks consist almost entirely of music, purposely excluding speech, dialogue, or non-musical ambient noise. Furthermore, the acoustic distribution features a substantial presence of traditional Chinese music, alongside ensembles blending regional and Western instruments.

As the authors did in their paper, we preprocessed the dataset, transforming the videos into 30-second video-audio pairs. We selected clips featuring solely music and excluded video frames that contained extensive superimposed text or captions. After manual filtering, the dataset is preprocessed by clipping for training. The statistical distribution of the preprocessed multimedia categories is summarized in Table 1.

Table 1: Statistical distribution and duration of content categories in the processed GVMGEN dataset.

Category	Duration (h)	Share (%)
Movies	89.5	60.88%
Documentaries	42.1	28.64%
Vlogs	9.9	6.73%
Other Types	5.5	3.74%
Total	147.0	100.00%

This structured distribution highlights a predominant reliance on cinematic narrative structures (Movies and Documentaries comprising over 89% of the total), which justifies the highly descriptive visual features extracted by the visual backbone.

¹<https://www.bilibili.com/>

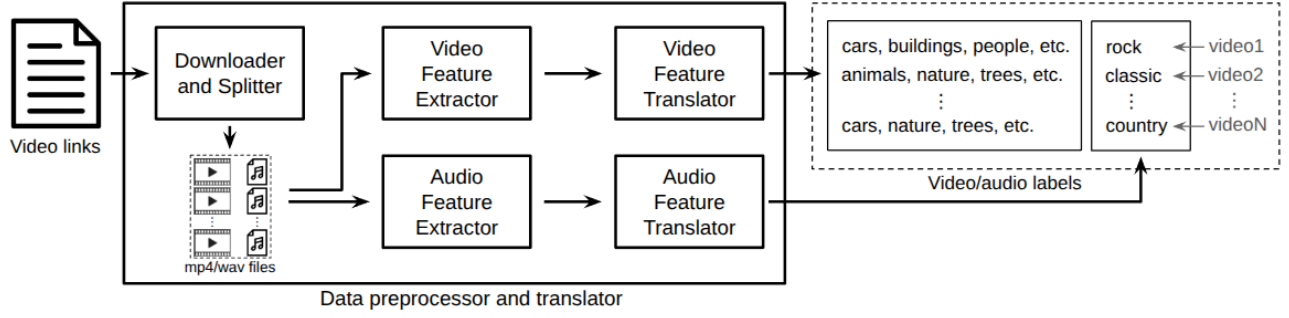


Figure 1: Methodology for dataset preparation and understanding.

3.1.2 Video Feature Extraction. Our proposed video feature extractor generates compact semantic representations from videos using the visual encoder of the Contrastive Language-Image Pretraining (CLIP) architecture [16]. Given an input video V , the method uniformly samples $N = 8$ frames across the temporal dimension, obtaining a representative set $\{f_1, f_2, \dots, f_N\}$. Each frame is decoded, converted to RGB format, and processed by the CLIP image processor before being forwarded to a pretrained Vision Transformer (ViT-B/32) [2].

For each frame f_i , the visual backbone produces a feature vector $h_i \in \mathbb{R}^d$, which is projected into the shared CLIP embedding space to obtain z_i . The resulting embedding is L2-normalized as $\hat{z}_i = z_i / |z_i|_2$, where ϵ is a small stabilization constant. A single representation for the entire video is then obtained through temporal mean pooling, $z_V = \frac{1}{N} \sum_{i=1}^N \hat{z}_i$, followed by a final normalization $\hat{z}_V = z_V / |z_V|_2$.

The extracted embedding captures high-level semantic information, including objects, scenes, and global visual context, while remaining computationally efficient. The final descriptor is stored together with metadata such as frame rate, resolution, and extraction parameters. Although temporal dynamics are not explicitly modeled, the representation provides a strong baseline for multimodal retrieval tasks.

3.1.3 Audio Feature Extraction. Complementing the visual representations, acoustic features are independently extracted using the Librosa framework [13]. Audio signals are decoded at a sampling rate of 22,050 Hz, optionally converted to mono, and truncated to a maximum duration of $T = 30.0$ seconds.

Temporal descriptors include root mean square (RMS) energy, $\text{RMS} = \sqrt{\frac{1}{N} \sum_{n=1}^N x_n^2}$, and zero-crossing rate (ZCR), which characterize signal intensity and temporal activity. Spectral descriptors such as spectral centroid, bandwidth, and rolloff are extracted from short-time analysis windows. The spectral centroid, computed as $C = \frac{\sum_k f_k X_k}{\sum_k X_k}$, estimates the center of mass of the spectrum. To capture harmonic and timbral properties, the method also computes chroma features and Mel-frequency cepstral coefficients (MFCCs), while global tempo is estimated in beats per minute (BPM) using Librosa’s beat tracking algorithm.

Instead of storing frame-level features, each descriptor stream is summarized through statistics including mean,

standard deviation, minimum, and maximum values. The resulting descriptors are saved as JSON files together with metadata such as sample rate, duration, and extraction parameters, providing an efficient representation for music analysis and multimodal learning tasks.

3.1.4 Video Feature Translation. Our proposed video feature translator converts pre-computed video embeddings into human-readable semantic descriptors using the shared multimodal space of CLIP. Given a normalized video embedding $\mathbf{v} \in \mathbb{R}^D$, the method compares it against text embeddings generated from a predefined vocabulary, providing an efficient approach for automatic video annotation and semantic indexing.

The predefined vocabulary is structured as a task-driven ontology covering key visual domains, including human presence, urban infrastructure, natural environments, and weather conditions. This complete vocabulary is made publicly available in our repository to ensure full reproducibility.

To ensure semantic robustness during text encoding and stabilize the zero-shot baseline mapping, each vocabulary label ℓ_i is embedded using a set of multiple composition prompt templates $\mathcal{T} = \{t_1, \dots, t_M\}$ (e.g., descriptions explicitly emphasizing scene presence). Each prompt is encoded by the CLIP text encoder as $e_i^m = f_{t_m} \ell_i$ and normalized according to $\hat{e}_i^m = e_i^m / \sqrt{e_{i,j}^{m,2}}$. The embeddings obtained from different templates are then averaged as $e_i = \frac{1}{M} \sum_{m=1}^M \hat{e}_i^m$, producing a single robust representation for each vocabulary label.

The final output consists of the top- K labels (evaluated at $K = 100$) ranked by similarity score and their corresponding confidence values. By leveraging CLIP’s [16] shared embedding space, the method efficiently associates video representations with natural language semantics without requiring additional supervised training.

3.1.5 Audio Feature Translation. Our proposed audio translation module converts low-level acoustic descriptors into human-readable semantic tags and approximate music style annotations. Operating on pre-computed audio features, the method generates ranked labels for perceptual properties such as tempo, energy, texture, tonality, and genre using lightweight heuristic rules instead of supervised learning.

The system uses descriptors extracted via Librosa [13] in a previous stage, including BPM, RMS energy, ZCR, spectral centroid, bandwidth, rolloff, and chroma distributions. These features are converted into semantic confidence scores using fuzzy membership functions. Given a feature value x , the triangular, ascending, and descending functions are defined as $\mu_{\text{tri}}x = \max(0, 1 - |x - c|/w)$, $\mu_{\text{up}}x = \text{clip}(x - \tau w, 0, 1)$, and $\mu_{\text{down}}x = \text{clip}(\tau - xw, 0, 1)$, where c , w , and τ are parameters. To guarantee deterministic reproducibility, these parameters are calibrated following standard fuzzy membership function formulations [18], grounded on physical and musical thresholds (such as structural boundaries for low/high acoustic energy and localized rhythmic BPM ranges) hardcoded within our open-source translation module. They are used to infer tempo, energy, brightness, and texture from the corresponding audio descriptors. Tonality is obtained from chroma distributions, with tonal confidence computed as $\gamma = \max_{i=1}^{12} c_i$ for $\mathbf{c} \in \mathbb{R}^{12}$.

The framework also estimates musical styles using weighted rule combinations, where each score is computed as $S = \frac{\sum_{i=1}^N w_i \mu_i x_i}{\sum_{i=1}^N w_i}$. Finally, labels and style predictions are ranked and exported as textual annotations, providing an interpretable and efficient baseline for audio characterization.

3.1.6 Semantic Grouping and Spatial Relaxation. To mitigate the spatial sparsity inherent to the raw visual frame descriptions, where the same semantic concept appearing in different feature positions (e.g., *feature_1* versus *feature_8*) is treated as entirely distinct by the mining algorithms, a feature optimization stage was introduced. This module applies a dual transformation to the categorical dataset.

First, a semantic mapping dictionary consolidates synonymous or highly correlated visual concepts into high-level taxonomic categories (e.g., mapping *lake*, *river*, and *sea* into a unified *water* token). Second, the positional constraints of the ten visual feature columns are relaxed by converting the dataset into a binary BoVW representation. Each unique semantic concept becomes a binary attribute ($\text{has_concept} \in \{0, 1\}$), indicating its presence or absence anywhere within the 30-second video segment, regardless of its original frame index. This spatial relaxation effectively increases the statistical relevance, robustness, and coverage of the discovered patterns.

3.2 Dataset Description

After extracting and translating visual and audio features into semantic labels, a descriptive analysis was performed to characterize the dataset and identify relevant patterns. The goal was to uncover recurring elements and associations within and across modalities.

The semantic labels extracted from the visual, audio, and combined representations are used as input for two complementary descriptive mining approaches: FP-Growth and Constrained Beam Search. FP-Growth is applied to identify frequent visual, audio, and multimodal patterns, providing a global characterization of the dataset. In contrast, the Constrained Beam Search is applied to discover discriminative

video-audio associations, enabling the identification of localized visual signatures that exclusively characterize specific musical styles without succumbing to sparsity bias.

3.2.1 Descriptive Mining Algorithms.

FP-Growth (Global Exploration). The FP-Growth algorithm [3] was employed to identify frequent combinations of semantic labels within the dataset. After the feature extraction and semantic translation stages, each instance was represented as a transaction $T_i = \{l_1, l_2, \dots, l_k\}$, where each label l_j corresponds to an interpretable concept describing the instance. The complete dataset is therefore represented as a collection of transactions $D = \{T_1, T_2, \dots, T_n\}$. The objective is to discover frequent itemsets $X \subseteq L$, where L denotes the set of all semantic labels present in the dataset. The support of an itemset X is defined as $\text{support}X = |\{T_i \in D \mid X \subseteq T_i\}|$. In this work, a minimum support threshold of 10 occurrences was adopted. The resulting frequent itemsets provide a compact description of the dataset by revealing recurring co-occurrences of semantic concepts, establishing a foundational baseline of the collection's overarching aesthetic choices.

Constrained Beam Search (Localized Discrimination). To isolate semantic patterns associated with specific target categories (musical genres) and overcome the limitations of classic SD in highly sparse multimedia matrices, we deployed a constrained Beam Search strategy, a standard heuristic approach in SD [7]. In a standard setup, pattern miners tend to rely heavily on the absence of elements ($\text{has_concept} == 0$) to form rules, generating uninformative negative heuristics (e.g., defining Rock by the absence of a skyscraper).

To prevent this sparsity bias, the search space was strictly restricted by deploying an `EqualitySelector` locked to the value 1.0. This methodological intervention forces the algorithm to navigate the search space evaluating only structural presences ($\text{has_concept} == 1$). Evaluated by the WRAcc metric, this constrained approach successfully identifies localized, purely positive visual signatures that distinguish specific subsets of the data, providing explicit Boolean filters that are highly valuable for semantic indexing.

3.2.2 Combined Data Description. The combined data description aims to characterize the relationships between the visual and audio modalities. Let $C = \{c_1, c_2, \dots, c_n\}$ denote the set of multimodal instances, where each instance c_i is represented by a set of semantic labels $Z_i = V_i \cup A_i$. By combining labels from both modalities into a single representation, it becomes possible to investigate recurring associations between visual concepts and audio characteristics.

To obtain a global characterization, FP-Growth was applied to the collection of multimodal label sets, revealing the most common generic contexts present in the dataset (e.g., the overarching dominance of natural environments across all genres).

Subsequently, the Constrained Beam Search was employed to extract highly discriminative visual signatures associated with specific acoustic categories. In this setting, the binary visual labels were used as descriptive attributes, while

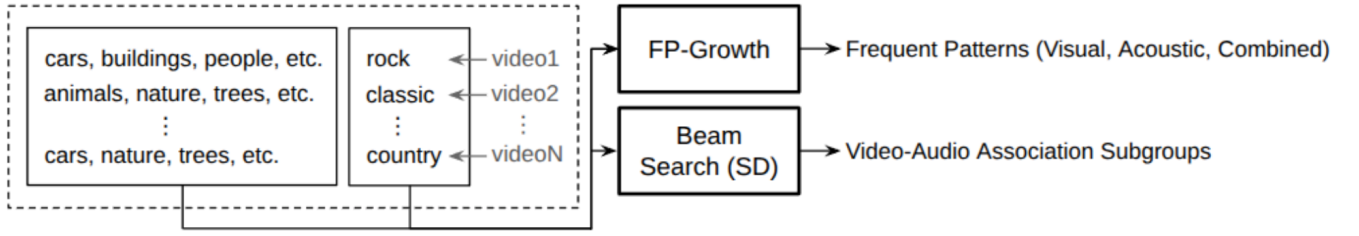


Figure 2: Methodology for translated dataset description.

the dominant audio genre was adopted as the target variable. Consequently, this targeted search complements the frequency-based analysis by bypassing the generic dense areas of the dataset, revealing isolated multimodal rules (e.g., acoustic/folk directly linked to pastoral visual features) that serve as highly precise query filters for semantic retrieval tasks.

4 EXPERIMENTS

This section presents the experiments conducted using the proposed methodology to process and describe the GVMGEN dataset. We first introduce the experimental setup and implementation details adopted throughout the study. Then, we analyze the descriptive patterns obtained from video and audio data individually, as well as from their multimodal combination, highlighting the knowledge extracted by the descriptive learning algorithms.

4.1 Experimental Setup

To comprehensively analyze the GVMGEN dataset, the descriptive learning process was divided into three distinct groups: visual data description, acoustic data description, and combined visual-acoustic data description. This separation allows the characteristics of each modality to be investigated independently, providing insights into the visual patterns present in the generated videos and the acoustic properties of the associated music tracks.

In addition to the unimodal analyses, a multimodal description was performed by combining visual and acoustic features into a single representation. The goal of this analysis is to identify patterns that emerge from the interaction between the two modalities, enabling the discovery of potential relationships between video content and musical characteristics. This multimodal perspective is particularly relevant for understanding how generative systems align visual and auditory elements when producing music videos.

4.2 Implementation Details

To facilitate reproducibility and compatibility, all experiments were executed using Docker containers. The source code repository provides a Dockerfile and a shell script that automatically install and configure all required libraries and dependencies.

The entire framework was implemented in Python. For data preprocessing, feature extraction and translation, we employed widely used libraries including *Pillow*, *Librosa*, *OpenCV*, *Transformers* [23], *PyTorch* [15], and *TorchVision*, which provide efficient tools for image processing, audio analysis, deep learning, and multimodal representation extraction. The descriptive learning stage was implemented using the *PyFPGrowth*² and *PySubgroup* [8] libraries. *PyFPGrowth* was employed to efficiently discover frequent patterns from the discretized feature representations, enabling the identification of recurring attribute combinations within the dataset. In parallel, *PySubgroup* was used to perform SD, searching for statistically interesting subsets of instances that exhibit distinctive characteristics with respect to the target classes.

To establish a comprehensive comparative analysis to perform SD, the descriptive mining process was divided into two main algorithmic paradigms:

- **Constrained Beam Search:** SD tasks were implemented using the *PySubgroup* framework. The search space was explored using a *Beam Search* strategy with a maximum search depth of three attributes ($depth = 3$), optimized via the *Standard Quality Function* ($\alpha = 1.0$), which aligns with the *WRAcc* metric [6]. Formally, the quality of a discovered subgroup S with respect to a target class is evaluated as:

$$WRAccS = PS \cdot (P_{target} | S - P_{target}) \quad (1)$$

where PS represents the fractional coverage of the subgroup, $P_{target} | S$ is the precision within the subgroup, and P_{target} is the global base rate of the target genre. Crucially, the search space was restricted using an *EqualitySelector* locked to positive occurrences (1.0) to avoid sparsity bias.

The source code and the extracted semantic features used in this study are publicly available on GitHub³. The raw multimedia files of the GVMGEN dataset [24] are the intellectual property of their original creators and must be requested directly from them.

²<https://pypi.org/project/pyfpgrowth/>

³<https://github.com/Frankl1sales/mining-multimodal-patterns>

4.3 Results and Discussion

This section presents the descriptive learning results obtained from the GVMGEN dataset. The results are organized into three categories: visual data description, acoustic data description, and combined visual-acoustic data description.

4.3.1 Data Distribution and Imbalance. The processed dataset comprises 20,100 synchronized 30-second segments distributed across eight distinct acoustic styles. Preliminary data exploration, as summarized in Table 2, reveals a severe class imbalance. Mainstream commercial genres heavily dominate the collection, with Rock comprising 38.77% (7,793 videos) and Electronic/Dance making up 20.35% (4,090 videos). Conversely, niche structural styles like Ambient/Relax (2.21%) and Classical/Orchestral (0.19%) represent a minute fraction of the dataset, establishing a highly skewed environment that complicates vanilla descriptive exploration.

Table 2: Video Distribution across Musical Genres in the GVMGEN Dataset

Musical Genre	Number of Videos	Percentage (%)
Rock	7,793	38.77
Electronic/Dance	4,090	20.35
Jazz	3,860	19.20
Hip-Hop	1,938	9.64
Acoustic/Folk	981	4.88
Pop	955	4.75
Ambient/Relax	444	2.21
Classical/Orchestral	39	0.19
Total	20,100	100.00

Regarding the semantic visual content, the dataset features a consistent vocabulary of up to 88 unique concepts per genre. A global frequency analysis reveals a strong prevalence of natural and external elements across the entire collection. The most frequent individual concepts include *outdoor* (14,619 occurrences), *park* (11,429), *river* (9,215), *day* (9,081), and *lake* (8,534). While these absolute frequencies provide a baseline understanding of the most common visual archetypes, they only capture individual, isolated occurrences. To fully understand how these features interact and form complex visual narratives, it is necessary to analyze their combinations, which motivates the use of frequent pattern mining algorithms.

4.3.2 Global Dataset Topology. To visualize the semantic layout of the dataset prior to rule mining, a two-dimensional projection of the multimodal feature space was generated using Uniform Manifold Approximation and Projection (UMAP) [14], as illustrated in Figure 3. The projection was computed directly over the binary BoVW feature matrix using the following hyperparameters: $n_neighbors = 15$, $min_dist = 0.1$, a random seed of 42, and Euclidean distance as the metric. While Jaccard or Dice distances are conventionally tailored for binary vector spaces, Euclidean distance was deliberately

selected here as it yielded significantly tighter and more stable cluster separations during preliminary baseline explorations, effectively preventing the hyper-sparse peripheral tokens from artificially fragmenting the primary structural components.

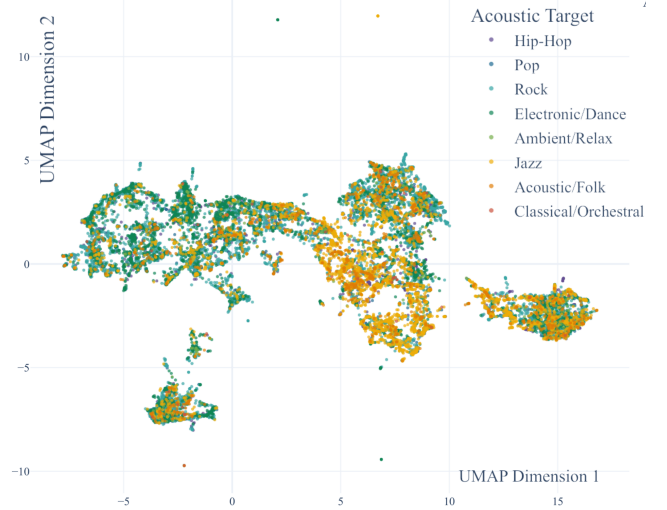


Figure 3: Global Semantic Topology of the GVMGEN dataset via UMAP projection ($n_neighbors = 15$, $min_dist = 0.1$, Euclidean metric, $random_state = 42$) applied over binary BoVW representations. Colors indicate the dominant acoustic genre label per segment.

The continuous topology reveals that while specific semantic niches form isolated clusters, the vast majority of commercial videos form a dense, heavily intertwined central mass. This high spatial intersection indicates that different musical genres frequently share identical visual backdrops and production designs, providing a strong methodological justification for utilizing discriminative mining algorithms to extract localized signatures from the overlapping noise.

4.3.3 Global Frequent Patterns (FP-Growth). To establish a foundational understanding of the semantic structures within the GVMGEN dataset, the FP-Growth algorithm was applied to extract the most recurrent itemsets. Table 3 presents the top 15 most frequent visual-only, audio-only, and combined combinations discovered across the entire collection.

The purely visual patterns reveal a marked dominance of natural and daylight environments. The combination of *{outdoor, park}* is the most prevalent purely visual context, appearing in 11,607 video segments. Similarly, the frequent co-occurrence of *{lake, river}* (8,190 occurrences) and *{forest, outdoor}* (8,593 occurrences) confirms that a substantial portion of the generated videos relies on nature-oriented aesthetics.

Interestingly, the frequent itemsets extracted from *Audio Features Only* often include multi-genre combinations rather than isolated, mutually exclusive labels. For instance, the itemset *{electronic/dance, rock}* ranks third with a significant support of 9,096 segments, followed by combinations

like $\{jazz, rock\}$ (6,062 occurrences) and $\{acoustic/folk, jazz\}$ (5,029 occurrences). This phenomenon stems naturally from the multi-label design of our audio translation module. From a dataset understanding perspective, these co-occurrences reveal a substantial semantic overlap and cross-genre hybridization inherent to modern commercial soundtracks, where electronic production frequently blends with traditional rock or jazz structures.

When evaluating the combined multimodal patterns, the results strongly reflect both the cross-genre audio blending and the overarching class imbalance. Because *rock* and *electronic/dance* are the dominant musical genres and *outdoor* is the most frequent visual marker, they naturally coalesce into the most frequent cross-modal itemsets. The pattern $\{outdoor, rock\}$ is the most recurrent cross-modal structure, found in 9,510 segments, followed closely by other combinations of high-frequency genres mapped to generic outdoor scenarios (e.g., $\{park, rock\}$ and $\{day, rock\}$).

While these FP-Growth results successfully capture the global topology and the most common archetypes of the dataset, they also highlight a crucial limitation: high frequency does not necessarily imply high discriminative power. Because tags like *outdoor* and *rock* are overwhelmingly common, they co-occur across substantial portions of the dataset, providing little to no specificity. This behavior justifies the necessity of applying SD to transition from simple frequent occurrences to statistically significant, discriminative relationships.

4.3.4 Discriminative Semantic Discovery. By applying the positive-literal constrained Beam Search, the pipeline extracted clean, explicit visual signatures for individual target styles, eliminating negative clutter. As shown in Table 4, the algorithm discovered highly cohesive structural rules. For instance, the Acoustic/Folk genre is mapped directly to pastoral environments ($has_road \wedge has_park \wedge has_village$). Mainstream genres like Pop isolated stable signatures tied to summer aesthetics ($has_sea \wedge has_outdoor \wedge has_beach$). By restricting the selectors to positive occurrences, the search space was completely protected against noise fragmentation, yielding highly interpretable, compact conjunctions.

4.3.5 Downstream Task: Semantic Video Retrieval. To validate the practical utility of the discovered rules, we evaluated them in a downstream Semantic Video Retrieval task. The extracted Boolean rules were deployed as low-overhead database indexing filters. To quantify the real-world advantage of using these discovered patterns over an uninformed search, we adopt the Retrieval Lift metric [1]. For a given discovered subgroup S , the Lift is mathematically defined as:

$$Lift_S = \frac{P_{target | S}}{P_{target}} \quad (2)$$

where $P_{target | S}$ represents the precision of the rule (the conditional probability of observing the target musical genre within the subgroup S), and P_{target} represents the global baseline probability of that genre across the entire GVMGEN dataset. A Lift value greater than 1.0 indicates that the visual

signature successfully concentrates the target class, providing a directional advantage for search engines.

While state-of-the-art zero-shot retrieval utilizing dense embedding similarity (e.g., standard CLIP embeddings) typically establishes the upper bound for precision, it requires expensive computational resources. Our objective is not to surpass dense neural retrieval in raw accuracy, but to demonstrate that the extracted Boolean rules provide a clear directional advantage acting as extremely lightweight, first-stage symbolic filters.

The quantitative results of this extrinsic evaluation are summarized in Table 4. The application of the symbolic rules as search queries yielded a substantial precision increase across all musical styles, achieving a Retrieval Lift of up to $2.91\times$ over the random baseline. Similarly, the Pop filter ($has_sea==1.0 \wedge has_outdoor==1.0 \wedge has_beach==1.0$) elevated retrieval precision to 12.99%, providing a $2.73\times$ lift.

Crucially, while evaluating the Boolean query over a database of N instances requires a linear scan of ON or a sub-linear lookup of $O(\log N)$ when leveraging standard inverted indexes [12], the operational cost per record is reduced to a lightweight $O(1)$ bitwise comparison on the CPU. This effectively replaces the computationally expensive $ON \times d$ floating-point matrix multiplications required by dense cross-modal embeddings (where $d = 512$ for the utilized CLIP architecture) at inference time, substantially lowering both the constant execution factors and hardware dependencies for real-time multimedia indexing.

4.4 Threats to Validity

While our descriptive pipeline successfully extracts interpretable multimodal rules, we acknowledge certain methodological boundaries inherent to this study. First, operating strictly under a descriptive learning paradigm aimed at global data profiling, our algorithms ingested the entire dataset to map its complete topology, rather than utilizing a predictive train/test split. Consequently, the reported Retrieval Lift demonstrates a statistical concentration over the base rate, acting as a lightweight first-stage indexing mechanism rather than competing directly with the generalized predictive accuracy of dense neural baselines (e.g., end-to-end CLIP retrieval). Furthermore, we note that the extreme Lift observed for the Classical/Orchestral genre ($2.91\times$) is heavily influenced by its severe rarity (0.19% of the dataset), indicating that minority class signatures should be interpreted with statistical caution.

Additionally, certain design choices impact the generalizability of our findings. The spatial relaxation into a BoVW representation, while crucial for eliminating spatial sparsity, intentionally discards temporal narrative sequences, potentially missing sequential video-audio correlations. Finally, the audio translation module relies on hardcoded heuristic thresholds and fuzzy membership functions. While grounded in acoustic theory, these subjective parameters and the reliance on CLIP’s fixed zero-shot vocabulary may limit the robustness of the extracted rules when applied to multimedia

Table 3: Top 15 most frequent visual, audio, and combined patterns discovered by the FP-Growth algorithm. Support indicates the absolute number of segments containing the itemset.

Visual Features Only			Audio Features Only			Combined (Multimodal)		
Rank	Itemset	Support	Rank	Itemset	Support	Rank	Itemset	Support
1	{outdoor}	16,379	1	{rock}	13,570	1	{outdoor}	14,619
2	{park}	13,309	2	{jazz}	11,803	2	{rock}	12,614
3	{outdoor, park}	11,607	3	{electronic/dance, rock}	9,096	3	{outdoor, rock}	9,510
4	{day, outdoor}	9,984	4	{jazz, rock}	6,062	4	{outdoor, park}	9,127
5	{forest, outdoor}	8,593	5	{acoustic/folk, jazz}	5,029	5	{forest, outdoor}	7,619
6	{outdoor, river}	8,369	6	{pop, rock}	4,838	6	{electronic/dance, rock}	7,588
7	{lake, river}	8,190	7	{jazz, pop}	4,716	7	{lake, river}	7,056
8	{lake, outdoor}	7,932	8	{electronic/dance, jazz}	4,122	8	{day, outdoor}	7,024
9	{park, river}	7,790	9	{hip-hop, rock}	3,309	9	{park, rock}	6,692
10	{day, park}	7,377	10	{electronic/dance, jazz, rock}	3,214	10	{day, rock}	6,592
11	{mountain, outdoor}	7,206	11	{electronic/dance, pop, rock}	2,925	11	{garden, outdoor}	6,420
12	{garden, outdoor}	7,171	12	{acoustic/folk, rock}	2,609	12	{electronic/dance, outdoor}	6,414
13	{forest, park}	7,144	13	{acoustic/folk, hip-hop}	2,331	13	{outdoor, river}	6,312
14	{rain, outdoor}	6,948	14	{electronic/dance, hip-hop}	2,212	14	{park, river}	6,154
15	{forest, outdoor, park}	6,943	15	{acoustic/folk, jazz, pop}	2,088	15	{lake, outdoor}	6,050

Table 4: Discriminative Visual Signatures Discovered for Each Musical Genre Ranked by Retrieval Lift

Musical Genre	Discriminant Boolean Rule	Coverage	Precision	Lift ($\times$)
Acoustic/Folk	$\text{has_road}==1.0 \wedge \text{has_park}==1.0 \wedge \text{has_village}==1.0$	1,255	12.03%	2.47 $\times$
Pop	$\text{has_sea}==1.0 \wedge \text{has_outdoor}==1.0 \wedge \text{has_beach}==1.0$	962	12.99%	2.73 $\times$
Rock	$\text{has_animal}==1.0 \wedge \text{has_outdoor}==1.0$	3,155	53.47%	1.38 $\times$
Electronic/Dance	$\text{has_rain}==1.0 \wedge \text{has_garden}==1.0 \wedge \text{has_outdoor}==1.0$	3,532	34.97%	1.72 $\times$
Ambient/Relax	$\text{has_mountain}==1.0 \wedge \text{has_park}==1.0 \wedge \text{has_village}==1.0$	2,557	5.16%	2.34 $\times$
Hip-Hop	$\text{has_rain}==1.0 \wedge \text{has_garden}==1.0 \wedge \text{has_park}==1.0$	2,159	14.91%	1.55 $\times$
Jazz	$\text{has_small_town}==1.0 \wedge \text{has_mountain}==1.0 \wedge \text{has_village}==1.0$	2,911	47.10%	2.45 $\times$
Classical/Orchestral	$\text{has_people}==1.0$	1,773	0.56%	2.91 $\times$

datasets with different cultural contexts or musical distributions. In addition, our empirical validation was conducted exclusively on the GVMGEN dataset; although the pipeline is designed to be dataset-agnostic, its generalization to other video-audio corpora remains to be verified in future work.

5 CONCLUSION

Rather than treating modern multimodal generative datasets as opaque neural black boxes, this work demonstrated that explicit, interpretable semantic rules can be successfully extracted to characterize their underlying video-audio relationships. By introducing a spatial relaxation layer that transforms continuous visual embeddings into a binary BoVW representation, we bypassed the temporal fragmentation typical of frame-level data. Additionally, enforcing positive-literal constraints during Beam Search eliminated sparsity bias, preventing the generation of uninformative negative rules.

Experimentally, FP-Growth mapped the dataset’s global structure, revealing a dominance of outdoor backdrops, while Constrained Beam Search isolated specific visual signatures

for individual musical genres. We validated these symbolic rules in a Semantic Video Retrieval task. Operating as declarative queries, they achieved a Retrieval Lift of up to 2.91 $\times$ over the random baseline using lightweight *O1* CPU Boolean filtering, effectively bypassing expensive dense matrix multiplications at inference time.

Future work will explore injecting these semantic rules as prior knowledge into generative architectures, acting as attention reward functions for video summarization and region-of-interest indicators for neural video compression. Finally, we aim to develop an information-theoretic approach based on the Minimum Description Length principle [17] to natively handle hyper-sparse transactional databases without manual selector constraints.

6 ETHICAL CONSIDERATIONS

This study did not involve the direct participation of human subjects, and no interviews, surveys, experiments, or personal data were collected directly from individuals by the authors. The experiments used the GVMGEN dataset

[24], previously collected by the authors of the corresponding reference work from publicly available audiovisual content on an online video-sharing platform and accessed for research purposes upon request. The authors of this study did not participate in the original data collection and therefore do not make additional claims regarding the consent, anonymization, or ethical procedures adopted during its construction beyond the information provided by the creators. The dataset was used exclusively for academic purposes, without identifying or analyzing individuals depicted in the source content.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. Generative AI (Gemini, Claude) was used solely for textual and code refinement under the authors' full accountability. All core methodology, algorithmic design, and data analysis were developed and executed by the human authors

REFERENCES

- [1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. 207–216.
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [3] Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (Dallas, Texas, USA) (*SIGMOD '00*). Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/342009.335372
- [4] Francisco Herrera, Cristóbal José Carmona, Pedro González, and María José del Jesus. 2011. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems* 29, 3 (2011), 495–525. doi:10.1007/s10115-010-0356-2
- [5] Haven Kim, Zachary Novack, Weihao Xu, Julian McAuley, and Hao-Wen Dong. 2025. Video-Guided Text-to-Music Generation Using Public Domain Movie Collections. *arXiv:2506.12573 [cs.SD]* <https://arxiv.org/abs/2506.12573>
- [6] Nada Lavrač, Peter Flach, and Blaz Zupan. 1999. Rule Evaluation Measures: A Unifying View. In *Inductive Logic Programming*, Sašo Džeroski and Peter Flach (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 174–185.
- [7] Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. 2004. Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research* 5 (2004), 153–188.
- [8] Florian Lemmerich and Martin Becker. 2018. pysubgroup: Easy-to-use subgroup discovery in python. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 658–662.
- [9] Jiacheng Li, Haizhou Shi, Siliang Tang, Fei Wu, and Yueting Zhuang. 2019. Informative visual storytelling with cross-modal rules. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2314–2322.
- [10] Ken-Hao Liu, Ming-Fang Weng, Chi-Yao Tseng, Yung-Yu Chuang, and Ming-Syan Chen. 2008. Association and temporal rule mining for post-filtering of semantic concept detection in video. *IEEE Transactions on Multimedia* 10, 2 (2008), 240–251.
- [11] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.
- [12] Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.
- [13] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matthew McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*.
- [14] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
- [15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).
- [16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*.
- [17] Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica* 14, 5 (1978), 465–471.
- [18] Timothy J Ross. 2005. *Fuzzy logic with engineering applications*. John Wiley & Sons.
- [19] Sivic and Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings ninth IEEE international conference on computer vision*. IEEE, 1470–1477.
- [20] Jonatan Taminau, Ruben Hillewaere, Stijn Meganck, Darrell Conklin, Ann Nowé, and Bernard Manderick. 2009. Descriptive subgroup mining of folk music. In *2nd International Workshop on Machine Learning and Music (MML 2009)*, Bled, Slovenia.
- [21] Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2025. VidMuse: A Simple Video-to-Music Generation Framework with Long-Short-Term Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18782–18793.
- [22] V. Vijayakumar and R. Nedunchezian. 2012. A study on video data mining. *International Journal of Multimedia Information Retrieval* 1, 3 (01 Oct 2012), 153–172. doi:10.1007/s13735-012-0016-2
- [23] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.
- [24] Heda Zuo, Weitao You, Junxian Wu, Shihong Ren, Pei Chen, Mingxu Zhou, Yujia Lu, and Lingyun Sun. 2025. GVMGen: a general video-to-music generation model with hierarchical attentions. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Article 2576, 9 pages. doi:10.1609/aaai.v39i21.34474