Cross-Modal Learning in Music Visualizers

Morris Blaustein [*] [1]

Abstract

Music Visualizers are designed to create immersive experiences by exploiting correlations from sound and visual structure. Cross-modal learning can map audio features to compact visual descriptors derived from video frames. Using mel-spectral and temporal audio features, we evaluate linear regression, random forests and gradient-boosting models with GroupKFold cross-validation across multiple temporal resolutions. Our results show that visual weight and motion are significantly more predictable from audio than hue-based descriptors. These findings suggest that visual dynamics provide a more learnable basis for audio-driven visualization than direct color prediction. The goal of creating a model that learns vision statistics from audio features is to develop a real time audio visualizer that can effectively create correlated visuals. For this study, the trained model was used to predict features of a song, which outputted a visual weight and motion parameter for each frame. Additionally, framerate was varied in training to optimize the models performance.

1. Introduction

As media software technology improves, the amount of control a creator has increases as well. Music and visuals have always been paired together to create immersive experiences. However, with greater possibility also comes more work. Musicians are faced with the challenge of controlling visuals, which adds another dimension of skill. A music visualizer expands a one dimensional sound signal into many more spatially represented dimensions (1). Cross-modal learning takes the role of the visual artist, and is able to appropriately match video generation to musical input. Prior methods of developing relationships between sound and light have often focused on color, however we found that visual weight and motion are the best predictors for audio features. We cross-validated our models and tested them at different frame rates to determine optimal temporal resolution. With the optimal frame frame , we tuned the models by adjusting hyperparameters and feature scaling to maximize predictive performance. Once the model was trained, we used it to predict visual weight and motion and translated the frame rate values as control signals into TouchDesigner for music visualizer generation.

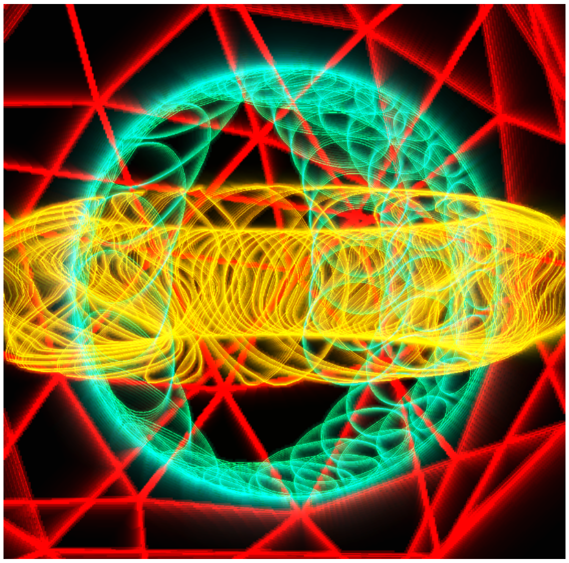


Figure 1. Music visualizer output driven by learned audio-to-visual control signals (predicted visual weight and motion) generated from extracted audio features.

2. Dataset and Feature Representation

2.1. YouTube Dataset

Videos were collected from public YouTube playlists using Pytube. FFmpeg was then used to standardize both video resolution and audio sample rate across the dataset. To study the effect of temporal resolution, video frames were extracted at five different frame rates: 1, 2, 4, 8, and 16 frames per second. Audio tracks were resampled to 22,500 Hz.

A playlist of music visualizers was selected for this dataset

[*] Equal contribution [1] Department of Statistics, Purdue University. Correspondence to: Morris Blaustein <mblauste@purdue.edu>.

because these videos are designed to respond directly to audio content. Unlike narrative or filmed video, music visualizers present a more direct and intentional mapping between sound and visual motion, making them well suited for learning and evaluating audio-visual relationships.

2.2. Audio Features

Audio features were extracted to represent both the spectral content and temporal behavior of the music used in the visualizers. A mel-spectrogram was used to summarize frequency content in a compact and perceptually meaningful way, producing low-dimensional frames that capture how energy is distributed across frequency bands.

In addition to spectral information, time-domain features were included to capture changes in amplitude and rhythm over time. Root mean square (RMS) energy was used to measure overall signal strength, while onset strength was used to indicate the presence of transient events such as note attacks or rhythmic changes.

Each audio frame consisted of a total of 66 features: 64 mel-band features derived from the mel-spectrogram, one RMS energy feature, and one onset strength feature. Together, these features provide a balanced representation of frequency content and temporal dynamics suitable for learning audio-visual relationships.

2.3. Visual Targets

The visual targets used in this work were chosen to represent clear, perceptually meaningful aspects of video, while avoiding stylistic or aesthetic decisions such as color. Instead of trying to predict full video frames or artistic features, the targets focus on visual properties related to motion and change over time, which are more consistent across different types of content.

For each video frame, a compact representation is extracted consisting of two values: a motion term and a weight term. The motion term measures how much the image changes from one frame to the next, providing a simple estimate of visual activity that does not depend on what is shown in the scene. This allows the model to learn relationships between audio features and visual movement without relying on object recognition or scene understanding. The weight term measures how dominant the frame's primary color is, defined as the proportion of pixels belonging to the most prevalent color cluster (the "main" color) in that frame. Higher weight indicates a visually uniform frame dominated by a single color, while lower weight suggests a more mixed palette.

This target design was informed by early experiments using color-based features, such as hue, saturation, and value extracted from dominant colors in each frame. These representations consistently produced poor prediction results, suggesting that color does not have a reliable or consistent relationship with audio features. In practice, color appears to be an artistic choice rather than a statistically meaningful visual response to sound. Motion, on the other hand, provides a more general and content-independent visual response, especially in music-driven video where changes in rhythm, energy, and texture are often expressed through movement.

By limiting the prediction task to these low-dimensional and interpretable visual targets, the model is better suited to learning general audio–visual relationships instead of fitting to specific visual styles in the dataset. This approach supports the broader goal of the work, which is to test whether audio features contain useful information about visual behavior without relying on subjective aesthetic mappings.

3. Experiment Setup

3.1. Learning Models

Several regression models were evaluated to learn the relationship between audio features and the selected visual targets. All models were trained to predict the visual weight and motion values from the corresponding audio features on a frame-by-frame basis.

As a baseline, ordinary least squares linear regression with ℓ_2 regularization (ridge regression) was used to test whether a global linear mapping could capture the audio-visual relationship. This model provides a simple reference point and helps distinguish between linear and non-linear effects in the data.

To model non-linear relationships, ensemble-based methods were explored. Random forest regression was used to capture non-linear interactions. In addition, gradient boosting with decision trees was evaluated using a histogram-based gradient boosting regressor.

3.2. Training Protocol

All models were trained using supervised regression with audio features as inputs and visual targets as outputs. Training and evaluation were performed using GroupKFold cross-validation, where all frames from a given video were assigned to the same fold. This prevents temporal leakage and ensures that performance reflects generalization across unseen videos rather than memorization of individual sequences.

Model hyperparameters were selected using cross-validation on a fixed frame rate setting, and the same values were then used across experiments to ensure consistency. Performance was evaluated using the coefficient of determination (R^2) and root mean squared error (RMSE), reported as mean

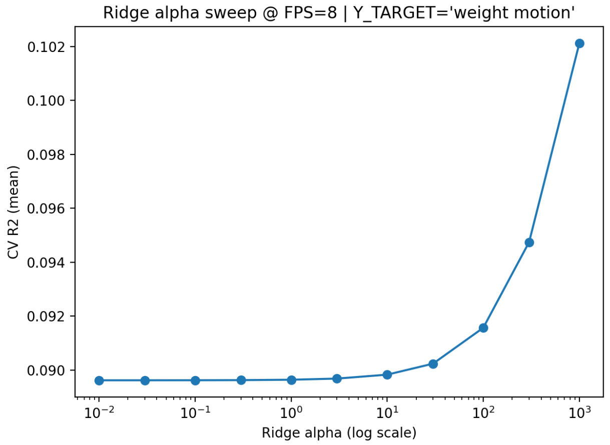


Figure 2. Ridge regression ℓ_2 regularization sweep at 8 FPS for predicting visual weight and motion. Performance is reported as mean cross-validated R^2 across GroupKFold splits by video. Stronger regularization improves generalization, consistent with the high-dimensional audio feature space.

values across folds.

3.3. Temporal Resolution Sweep

To study the effect of temporal resolution on audio-visual prediction, models were evaluated at multiple video frame rates. Video frames and corresponding audio features were aligned and sampled at 1, 2, 4, 8, and 16 frames per second. This sweep allows us to examine how performance changes as the temporal granularity of the visual targets increases.

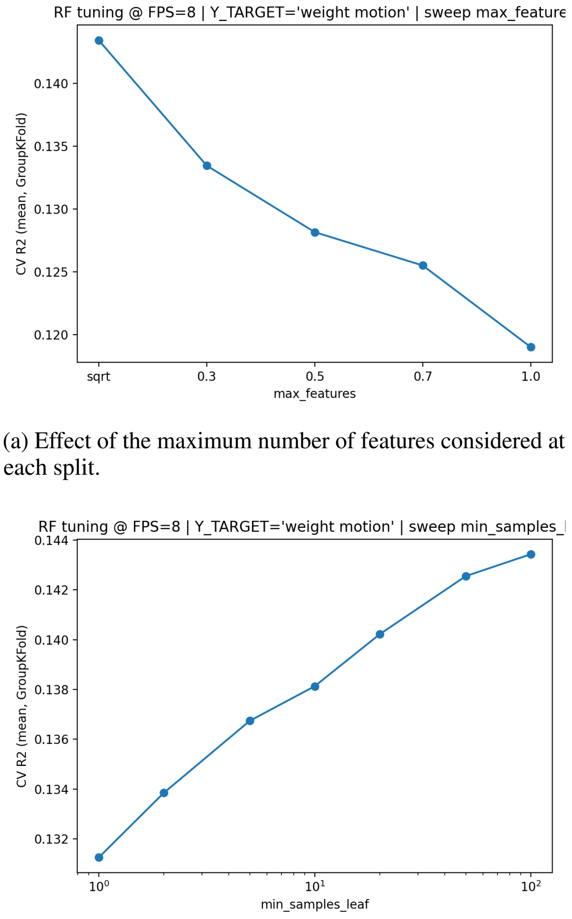
Lower frame rates emphasize slower, more global changes in visual behavior, while higher frame rates capture finer temporal detail but introduce increased noise and redundancy between adjacent frames. By evaluating models across this range, we aim to identify a temporal scale at which audio features most reliably predict visual motion and weight.

All other aspects of the training procedure and model configuration were held constant across frame rates, ensuring that observed differences in performance are attributable to temporal resolution rather than changes in model capacity or data processing.

4. Results

4.1. Frame Rate Sweep and Hyperparameter Tuning

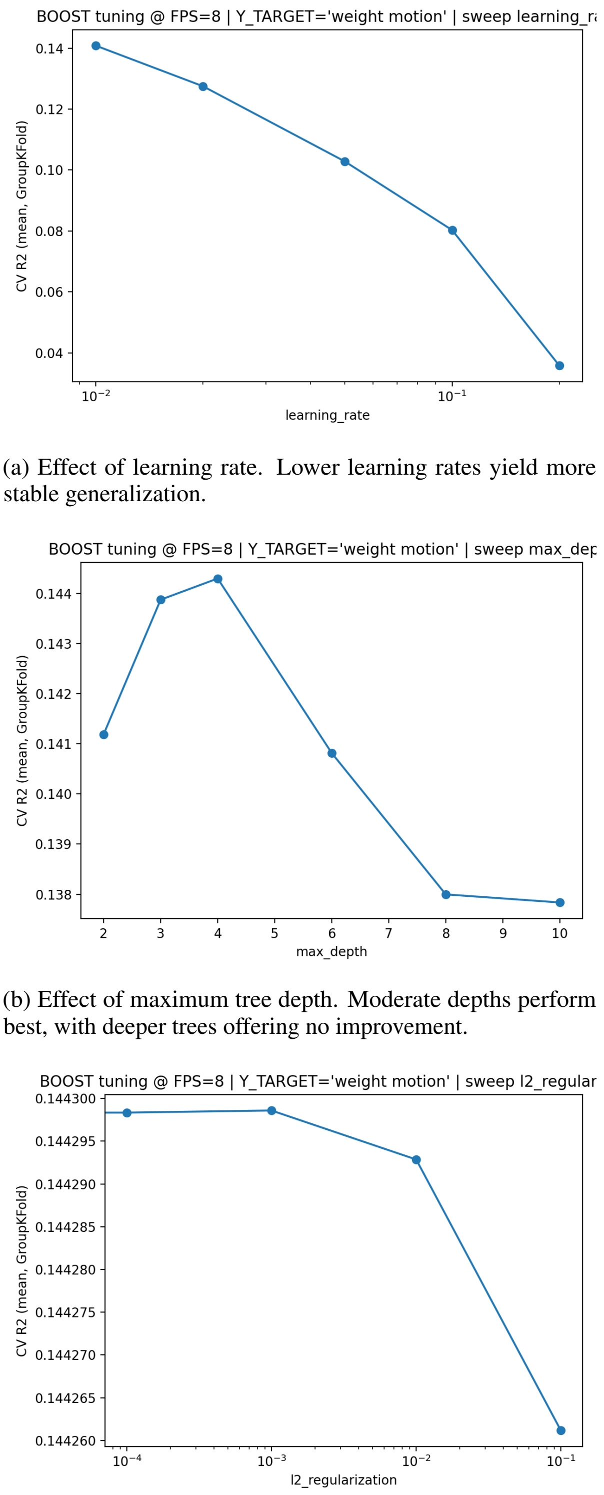
We first evaluated model performance as a function of temporal resolution by sweeping the video frame rate while holding the target definition and training protocol fixed. Ridge regression was used for this initial sweep to provide a stable baseline with minimal model complexity. As shown



(a) Effect of the maximum number of features considered at each split.

(b) Effect of the minimum number of samples required per leaf.

Figure 3. Random forest hyperparameter tuning at 8 FPS for predicting visual weight and motion. Mean cross-validated R^2 is reported across GroupKFold splits by video. Restricting feature usage per split and increasing the minimum leaf size improves generalization, highlighting sensitivity to overfitting in high-dimensional audio feature spaces.



(a) Effect of learning rate. Lower learning rates yield more stable generalization.

(b) Effect of maximum tree depth. Moderate depths perform best, with deeper trees offering no improvement.

(c) Effect of ℓ_2 regularization. Performance is relatively insensitive within the tested range.

Figure 4. Gradient boosting hyperparameter tuning at 8 FPS for predicting visual weight and motion. Mean cross-validated R^2 is reported across GroupKFold splits by video. Overall performance is most sensitive to learning rate and tree depth, while additional regularization provides diminishing returns.

in the frame rate sweep, performance improves substantially when moving from very low temporal resolutions (1 FPS) to intermediate rates, with the highest cross-validated R^2 observed at 8 FPS. Increasing the frame rate beyond this point leads to a slight decline in performance, suggesting diminishing returns as adjacent frames become increasingly redundant. Based on this result, 8 FPS was selected as the operating point for subsequent model comparisons and hyperparameter tuning.

At the selected frame rate of 8 FPS, ridge regression hyperparameters were tuned by sweeping the ℓ_2 regularization strength. Performance was relatively flat for small regularization values and increased steadily for larger values, peaking at the highest regularization setting used. This behavior indicates that strong regularization is beneficial for this task, consistent with the high dimensionality of the audio feature space and the relatively low signal-to-noise ratio of the prediction targets.

Nonlinear models were then explored using gradient boosting with decision trees. Hyperparameter sweeps were performed over learning rate, tree depth, number of leaf nodes, and ℓ_2 regularization. Lower learning rates consistently produced better performance, while higher learning rates led to rapid degradation, indicating overfitting and unstable optimization. Moderate tree depths achieved the best results, with deeper trees offering no additional benefit and in some cases reducing generalization. Performance saturated quickly with respect to the number of leaf nodes, suggesting that relatively simple tree structures were sufficient. The effect of ℓ_2 regularization was comparatively minor within the tested range, with only small variations in performance observed.

Across all tuning experiments, performance gains from hyperparameter optimization were modest but consistent, reinforcing the conclusion that the predictive relationship between audio features and visual motion is present but limited in strength. Importantly, the observed trends were stable across folds, indicating that improvements were not driven by overfitting to specific videos but reflect genuine differences in model behavior and temporal resolution.

4.2. Predicting Visual Weight and Motion

Using the selected temporal resolution and tuned hyperparameters, models were trained to predict the visual weight and motion targets from audio features. Performance was evaluated using GroupKFold cross-validation at the video level to measure generalization to unseen visual sequences.

Across models, prediction performance was consistently stronger for the motion target than for the weight target. Motion exhibited positive R^2 values across most folds, indicating that audio features contain meaningful information about changes in visual activity over time. In contrast, prediction of visual weight was less reliable, with R^2 values often near zero or negative. This suggests that while motion aligns naturally with audio dynamics such as rhythm and energy, frame-level prominence is influenced by additional factors not captured by the audio features alone.

Overall, these results support the conclusion that audio features are predictive of visual motion at appropriate temporal scales, but are insufficient to fully explain visual weighting behavior. This reinforces the broader finding of the study: audio–visual relationships in music-driven video are strongest when grounded in temporal dynamics rather than stylistic or compositional choices.

Table 1. Best cross-validated R^2 and RMSE for each model (results shown for the best frame rate, 8FPS).

Model	R^2 mean	R^2 std	RMSE mean
Ridge	0.141	0.085	0.114
Random Forest	0.143	0.089	0.1139
Boosting	0.083	0.097	0.117

4.3. Reinterpreting the Role of Color

Prior surveys of music visualizers emphasize color as a primary visual encoding (1). However, results from our experiments indicate that hue is not an effective output feature for learning audio-visual mappings in this setting. Across all tested models, prediction of color-based targets yielded weak or non-significant performance, suggesting that color palette cannot be reliably inferred from audio features alone.

These findings imply that color in music visualizers functions primarily as an artistic or stylistic choice rather than a direct, learnable response to sound. While prior studies and surveys of music visualizers have often treated color as a primary visual encoder, our results suggest that such mappings do not generalize well across content when evaluated statistically. In contrast to motion-based targets, color does not exhibit a stable global relationship with audio features at the frame level.

It is important to note that this result does not imply that color is perceptually unimportant or irrelevant to audiovisual experiences. Rather, it suggests that color relationships may depend on higher-level semantic, contextual, or narrative factors not captured by the audio features or modeling approaches used in this study. Table 1 reports the R^2 values obtained when color was included as an output target, highlighting its consistently weaker performance relative to motion and visual weight.

Table 2. Performance of regression models for predicting dominant color representations from audio features.

Model	**Overall R^2**	**RMSE**
Linear Regression (OLS)	-0.98	0.61
Random Forest Regressor	-0.11	0.28

5. Discussion

The results of this study highlight a clear distinction between different types of visual attributes and their relationship to audio features. Across all models and temporal resolutions, visual motion emerged as the most reliably predictable target, while visual weight showed a weaker but still measurable relationship with audio. In contrast, color-based targets consistently failed to produce meaningful predictive performance. Together, these findings suggest that audio-visual coupling in music-driven video is strongest when grounded in temporal dynamics rather than aesthetic or stylistic choices.

The temporal resolution sweep revealed that intermediate frame rates provide the best balance between temporal context and noise. Very low frame rates under-sample visual behavior, while higher frame rates introduce redundancy and reduce generalization. The consistent performance peak around 8 FPS suggests that audio features encode visual dynamics most effectively at a timescale corresponding to perceptual musical structure rather than instantaneous events. This aligns with the observation that sustained motion and section-level changes are more predictable than frame-level detail.

Model comparisons further support this interpretation. Linear ridge regression was sufficient to capture a baseline relationship between audio features and visual motion, indicating that a global linear mapping explains a small but non-negligible portion of the variance. Random forest models achieved comparable performance, suggesting that limited non-linear structure is present but does not substantially improve generalization. In contrast, gradient boosting consistently underperformed relative to both ridge regression and random forests, likely due to overfitting or sensitivity to noise in the high-dimensional feature space. Together, these results indicate that increasing model complexity does not necessarily improve performance for this task and that the audio-visual relationship being modeled is relatively weak and close to linear.

The poor performance of color-based targets provides an important counterpoint to prior work. While color is often emphasized in surveys and qualitative analyses of music visualizers, the results here indicate that color does not exhibit a stable, global statistical relationship with audio features when evaluated across diverse content. This does not imply that color is perceptually unimportant, but rather that its use is likely governed by higher-level semantic, contextual, or artistic factors not captured by frame-level audio descriptors.

Beyond quantitative evaluation,the predicted motion signals offer practical value as control parameters for real-time visual systems. Rather than acting as direct responses to transients, these signals function as estimates of visual activity or momentum, enabling smoother and more intentional visual behavior. Overall, this work contributes to a clearer understanding of which aspects of visual behavior are meaningfully grounded in audio features and which are better understood as artistic or editorial choices layered on top of sound.

5.1. Citations and References

References

[1] Hugo B. Lima, Carlos G. R. dos Santos, and Bianchi S. Meiguins. A survey of music visualization techniques. *ACM Computing Surveys*, 54(7):143:1–143:29, 2021.