Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 13, No. 3, 2025

Chord-Color Mapping for Audio-Visual Generation: Integrating Deep Learning and Emotion Quantification

Zimo Dong *

Xianda College of Economics and Humanities, Shanghai International Studies University, 999 Dongtan Avenue, Chongming District, Shanghai, China

*** Corresponding author Email:** dongzimo.work@gmail.com

Abstract: This study explores chord-color mapping as a means of enhancing AI-driven audio-visual art. Addressing the longstanding lack of systematic analysis of the emotional connection between chords and colors, we propose a model that integrates deep learning and emotion quantification. The model establishes a "color-emotion-chord" pathway by leveraging image color features and a chord dataset with emotional labels, trained using an LSTM architecture for emotion-driven audio-visual generation. Evaluations demonstrate the model's ability to substantially enhance emotional expression and aesthetic appeal while preserving essential image information. This research contributes new methodologies for cross-sensory art, music therapy, interactive design, and art education.

Keywords: Chord; Color; Mapping; Audiovisual Generation.

1. Introduction

The rise of digital media and artificial intelligence is reshaping artistic creation, with multisensory integration becoming a key aesthetic direction. Although connections between harmonic structures (chords) and chromatic spectra (colors) have been explored since antiquity, attempts to establish universal frameworks remain limited by philosophical debate and artistic subjectivity. [1]

This study introduces a computationally tractable model that quantifies chord–color correlations for AI implementation. It addresses three central questions: how to rigorously define the correspondence between harmonic and chromatic elements through psychophysics and cognitive science; which deep learning methods, especially sequential models, best support this mapping; and how the resulting cross-sensory artifacts can evoke emotional and aesthetic responses.

Our contributions are twofold: advancing the theoretical understanding of cross-modal perception and providing practical tools that expand creative possibilities and foster interdisciplinary innovation.

2. Theorie and Method

2.1. Artistic Foundations of Chord-Color Mapping

Chords are not merely acoustic aggregates but powerful triggers of emotional experience [2]. Consonance, dissonance, modulation, and harmonic progression shape the affective tension of music [3]. Likewise, visual parameters such as hue, saturation, and lightness strongly influence perception: high-saturation warm colors often evoke excitement and joy, while low-lightness cool colors convey serenity or melancholy.[4-5]

Thus, the chord–color relationship is less an analogy than a cross-modal emotional isomorphism [6-7]. While frequently applied in interactive art, it has traditionally relied on intuition rather than parameterized mechanisms. This study abstracts music and color into computable multidimensional parameters, providing input–output spaces for algorithmic modeling.

2.2. Emotion Quantification and Cross-Sensory Expression

Emotion functions as the mediator linking chords and colors, requiring scientific quantification for computational use. Psychological frameworks such as the Semantic Differential and Russell's circumplex model describe emotions along measurable dimensions and form the basis for this study.

Implementation involves annotating chords with tension and tonal features, and mapping color attributes (e.g., lightness, saturation, temperature) into an emotional space. This constructs a "color–emotion–chord" pathway in which emotion tags transform subjective feelings into computable data. By aligning different modalities within a unified emotional space, this approach supplies a robust foundation for subsequent deep learning.

2.3. Application of Deep Learning in Audio-Visual Generation

Deep learning enables scalable multimodal generation, with Long Short-Term Memory (LSTM) networks particularly suited for sequential tasks such as music modeling. Here, a chord dataset annotated with emotion tags is used to train an LSTM that learns correspondences between color features and chord progressions.

Image color features are first mapped into an emotional space, then fed into the LSTM to generate chords. The model balances local chord–color correspondence with global coherence and emotional consistency through sequence modeling.

This method offers two key advantages: it overcomes the rigidity of manually defined rules by adaptively learning from data, and it enhances audiovisual generation with diversity and controllability, fostering immersion and emotional resonance.

3. Image and Music Feature Extraction and Mapping Model Construction

3.1. Image Feature Extraction and Preprocessing

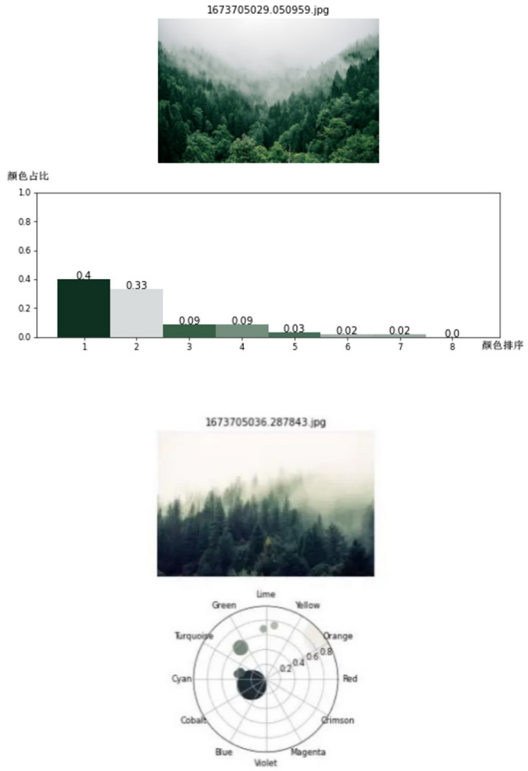

Fig 1. Histogram and Polar Plot

Feature extraction is a critical step in image analysis. This study employs OpenCV and Haishoku to derive color and shape characteristics, generating values such as dominant hue, primary color schemes, grayscale maps, and contours, which together form the data matrix for the mapping model. To ensure high-quality input, data transformation, cleaning, and classification are applied to remove irrelevant information. Key algorithms include cv2.THRESH_OTSU and cv2.CHAIN_APPROX_SIMPLE from OpenCV, as well as MinMaxScaler from scikit-learn.

Color information is first obtained through pixel-level scanning and converted from the BGR/RGB space into the perceptually uniform HSV space [8]. The dominant hue is then extracted via color histogram analysis in OpenCV. In parallel, Haishoku identifies the top eight primary color schemes by area proportion, with their ratios serving as mapping data for chord generation.

To visualize and interpret these results, dominant colors and their relationships are displayed through sorting and polar plots (Fig. 1). This process reveals the proportion, specific values, and categorical names of each hue. The color occupying the largest proportion is mapped to a corresponding musical mode, forming the foundation of vision–audition correspondence [9]. In the polar plot, colors closer to the center indicate lower lightness, which informs the selection and sequencing of chord progressions based on contrasts in hue, lightness, and scheme composition. [10]

3.2. Chord Progression and Pitch Sequence Generation

This study maps image curve structures to chord progressions, treating chord intervals as the basic unit. Using OpenCV, brightness information is extracted through grayscale conversion, binarization, denoising, and progressive scanning to obtain edge coordinates, forming a data matrix for chord progression modeling.

First, images are converted to grayscale with cv2.COLOR_BGR2GRAY, followed by weighted averaging to reduce noise. Binarization (cv2.THRESH_BINARY) segments brightness levels, appropriate for the uniform lighting of the test images. Contours are then traced with cv2.findContours; RETR_TREE captures hierarchical relationships, while CHAIN_APPROX_SIMPLE stores only key inflection points, preserving structural fidelity with minimal storage.

Contour morphology is mapped to chord sequences by polynomial fitting, linking visual linearity to acoustic linearity. Using NumPy's poly1d and non-linear least squares, a two-dimensional coordinate matrix is built, minimizing squared vertical errors to yield smooth curves. Tests showed that polynomial orders of 15–16 balance detail and stability, with fitting accuracy around 65–66% (Fig. 2).

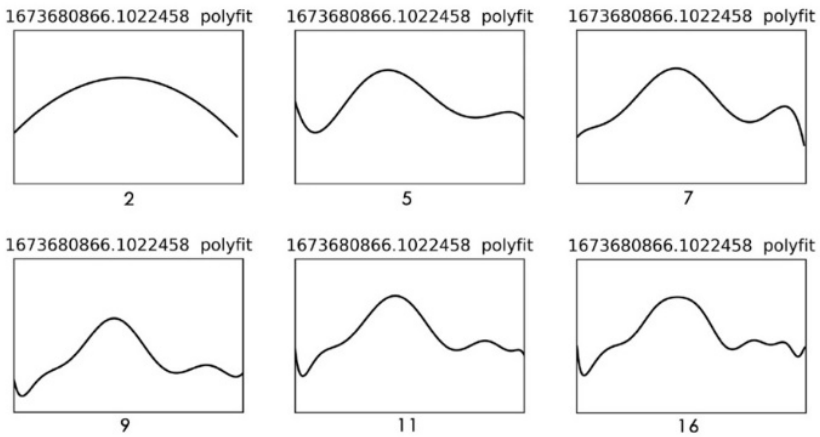

Fig 2. Test Graphs of Different Fitted Curves

These fitted curves are then translated into pitch sequences. Image aspect ratio determines the mapping strategy: horizontal images are scanned column-wise along the X-axis, while vertical images use row-wise scanning. Coordinates are mapped to the 88-key piano range (A0–C8). Normalization via MinMaxScaler scales Y-values to MIDI pitches 33–92. After duplicate removal and chronological reordering, note groups form high and low pitch arrays, which are further aligned to a musical mode to ensure harmonic coherence.

For output, the data are converted into MIDI sequences using the Mido library, with tempo and speed control. The X-axis represents musical time, with "note on/off" events determining onset and duration; absent values become rests. High and low parts are arranged in two tracks, while color-based chords derived from hue relationships are assigned to additional tracks. This process produces a MIDI audio file that mirrors the original image's color features and structural patterns (Fig. 3).

```
def choose_tone(midi, tone):                                          定义函数（调整音高）
    same_tone = 2
    for i in range(len(midi)):
        a = midi[i][1]//10
        for j in range(len(tone)):
            if tone[j] >= (a+1)*10:
                midi[i][1] = c
                break
            if tone[j] < a*10:
                continue                                              判断是否处于基本音高内
            if np.abs(midi[i][1]-tone[j]) < 4:
                c = tone[j]
                same_tone = 1
            if midi[i][1] == tone[j]:
                same_tone = 0
                break
    return midi
midi_note_bass = choose_tone(midi_note_bass, Keys_note)
midi_note_high = choose_tone(midi_note_high, Keys_note)
for i in range(len(midi_note_polyphony)):                             生成MIDI音轨
    bass_note_index = np.array(
        np.where(midi_note_bass[:, 0] == i)).flatten()
    high_note_index = np.array(
        np.where(midi_note_high[:, 0] == i)).flatten()
    if i == 0:
        if bass_note_index.size > 0:                                  选取高音部与低音部数值
            track.append(mido.Message('program_change',
                                      channel=0, program=0, time=0))
            track.append(mido.Message(
                'note_on', note=midi_note_bass[bass_note_index[0]][1], velocity=100, time=0))
            track.append(mido.Message('note_off', note=midi_note_bass[bass_note_index[0]][1],
                                      velocity=100, time=delay_time))
        if high_note_index.size > 0:
    else:
        if bass_note_index.size > 0:
            track.append(mido.Message(
                'note_on', note=midi_note_bass[bass_note_index[0]][1], velocity=100,
                time=(set_tempo*bass_i-delay_time)))
            track.append(mido.Message('note_off', note=midi_note_bass[bass_note_index[0]][1],
                                      velocity=100, time=delay_time))
            bass_i = 1
        else:
            high_i += 1
    bass_note_index = None
    high_note_index = None
    track2.append(mido.Message('program_change',                      控制音符播放与停顿
                               channel=0, program=0, time=0))
    if i==0:
        track2.append(mido.Message(
            'note_on', note=fre_index[chord_loop_count][0], velocity=100, time=0))
        track2.append(mido.Message(
            'note_on', note=fre_index[chord_loop_count][1], velocity=100, time=0))
        track2.append(mido.Message(
            'note_on', note=fre_index[chord_loop_count][2], velocity=100, time=0))
    elif (i%134)==0:
        track2.append(mido.Message(                                   控制和弦变换
            'note_off', note=fre_index[chord_loop_count][0], velocity=100,
            time=set_tempo*50-delay_time))
        track2.append(mido.Message(
            'note_off', note=fre_index[chord_loop_count][1], velocity=100, time=0))
        track2.append(mido.Message(
            'note_off', note=fre_index[chord_loop_count][2], velocity=100, time=0))
        if chord_loop_count>=7:
            chord_loop_count=0
```

Fig 3. Code for Converting Chord Progression Array to Audio Signal

3.3. Color-Emotion-Chord Mapping Model Design

The proposed "Color–Emotion–Chord" model transforms image features into chord sequences, ensuring emotional coherence between visual and musical domains. Dominant colors define the overall mode and fundamental pitch, while secondary colors map to chord roots. Saturation regulates the stability of thirds and fifths, and brightness determines register span. Relative color relationships further structure chord progressions, aligning visual and musical dynamics.

In implementation, image features are converted to HSV space. Hue determines mode category; saturation and brightness jointly specify mode attributes and base pitch. For example, bright, high-saturation hues map to major modes, while darker or desaturated hues correspond to minor modes (Fig. 4).

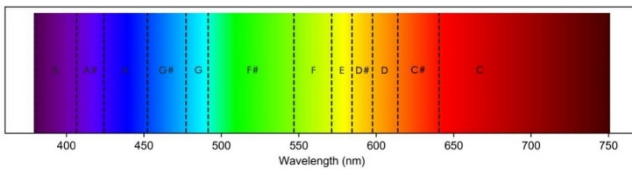

Fig 4. The Mapping of Spectra and Modes

Chord structures are then generated from color schemes: hue maps to chord roots, saturation adjusts interval qualities, and brightness sets octave range. A chord harmony table constrains intervals, correcting disharmonious combinations. The octave assignment is dynamically controlled by brightness thresholds—higher values map to upper registers, lower values to lower registers. [11]

Morphological features complement color information. Contour lines, extracted and polynomially fitted, are mapped onto the 88-key piano. Image aspect ratio determines the time axis: the x-axis for horizontal images and the y-axis for vertical ones, with fitted curves driving chord progression and rhythm.

Finally, color and morphological mappings are fused with valence–arousal coordinates in the emotional space. The system outputs a multi-track MIDI file: the main harmonic track reflects the dominant color, auxiliary colors generate chord tracks, and image contours define rhythmic layers. This integration ensures that the music not only mirrors visual structure and timbre but also achieves cross-modal emotional unity.

3.4. LSTM-based Music Sequence Generation Model

To improve the artistic quality and listenability of the generated compositions while preserving image information, this study applies a deep learning sequence model to refine the initial MIDI output. Musical rhythm and harmony depend on long-term temporal patterns, which manifest as extended sequences. While Recurrent Neural Networks (RNNs) can model temporal dependencies, they often suffer from vanishing gradients, limiting their ability to capture global features. To address this, Long Short-Term Memory (LSTM) networks are adopted for their capacity to maintain coherence across long sequences.

LSTMs extend the recurrent structure with gating mechanisms: the forget gate regulates historical information, the input gate updates current states, and the output gate generates hidden states based on the cell state. By propagating cell states over long spans, LSTMs effectively capture dependencies and enhance continuity in chord progressions.

During training, model inputs consist of image-derived color–emotion features, and outputs are chord sequences. A loss function combining cross-entropy with emotion-consistency regularization ensures both harmonic validity and alignment with emotional cues. Training uses the Adam optimizer, with hyperparameters such as learning rate, sequence length, and hidden unit size tuned for quality and generalization.

Compared to standard RNNs, LSTMs provide greater stability and accuracy in long-sequence modeling, enabling the generation of musically coherent segments with extended temporal scope and richer harmonic structures. Thus, the LSTM architecture is selected as the core engine to realize “color–emotion–chord” mapping in a manner that balances structural coherence with artistic expression.

4. Experiment and Result Analysis

4.1. Training and Results of the Generative Model

This experiment validates the effectiveness of the proposed image-to-music mapping rules and the MusicVAE framework in transforming visual information into auditory sequences. The procedure includes image data selection and annotation, preprocessing, MusicVAE training, and model fine-tuning.

The Medley2K dataset (OpenDataLab, 2020) was used, containing 2000 multi-genre tracks and 7712 genre tags [12]. To ensure consistency, all MIDI files were converted into NoteSequences, stored in TFRecord format, and split into training and testing sets at an 80/20 ratio.

Training employed MusicVAE (LSTM architecture, *cat-mel_2bar_big* configuration) with the following parameters: batch_size = 512, max_seq_len = 32, z_size = 256, enc_rnn_size = [256,256], free_bits = 0, max_beta = 0.2, beta_rate = 0.99999, sampling_schedule = 'inverse_sigmoid', and sampling_rate = 1000. The model ran in a computing environment equipped with an Intel processor and 2TB of memory, with a total duration of approximately 209 hours.

During the training process, monitored metrics included Loss, Global Norm, Global Step, Learning Rate, and Sampling Probability. The trend of the Loss function indicated that kl_beta gradually increased with iterations, kl_bits continuously decreased, and kl_loss rapidly dropped before stabilizing. The low variance suggested that the model learned the data distribution well; r_loss showed some fluctuation in the later stages, indicating a potential risk of overfitting. Global Norm fluctuated within the range of 0.5–1.3, with occasional spikes, showing that gradients were generally reasonable but local iterations had instability. Global Step showed a steady increase, confirming the continuity of the training process. Learning Rate gradually decreased from higher values to a stable range, consistent with the expected learning rate adjustment strategy. Sampling Probability quickly rose to close to 1 in the initial stage and remained stable, indicating that the model gradually relied on its own predictions for training.

To visually present the experimental results, this study summarizes the trends of all performance indicators, like figure 5.

These results demonstrate that the model can maintain stable convergence during prolonged training and exhibits good generalization ability in the color-chord mapping task, providing a solid foundation for subsequent optimization and application. [13]

4.2. Model Performance and Algorithm Optimization

This section explores advancements in audio-visual generation through optimizing generative model performance and algorithms. Using the MusicVAE framework as a baseline, targeted optimizations were applied. The strategy focused on two paths: model structure/training strategy, and code/data preprocessing improvements. Both enhance the "color-emotion-chord" mapping, improving quality and efficiency.

The hierdec-mel_16bar model was chosen, with label conditional vectors introduced. Latent space and network depth were tuned to mitigate overfitting and strengthen representation. The MTG-Jamendo dataset, containing fine-grained emotion tags, provided data. Consistent with color-emotion mapping, 4220 songs across Happy, Sad, Calm, and Energetic categories were selected. For training efficiency and balance, 200 songs per category (800 total) formed the main training set; remaining samples were for validation. Training utilized two NVIDIA 2080Ti-11G GPUs; total training time reduced to 54 hours.

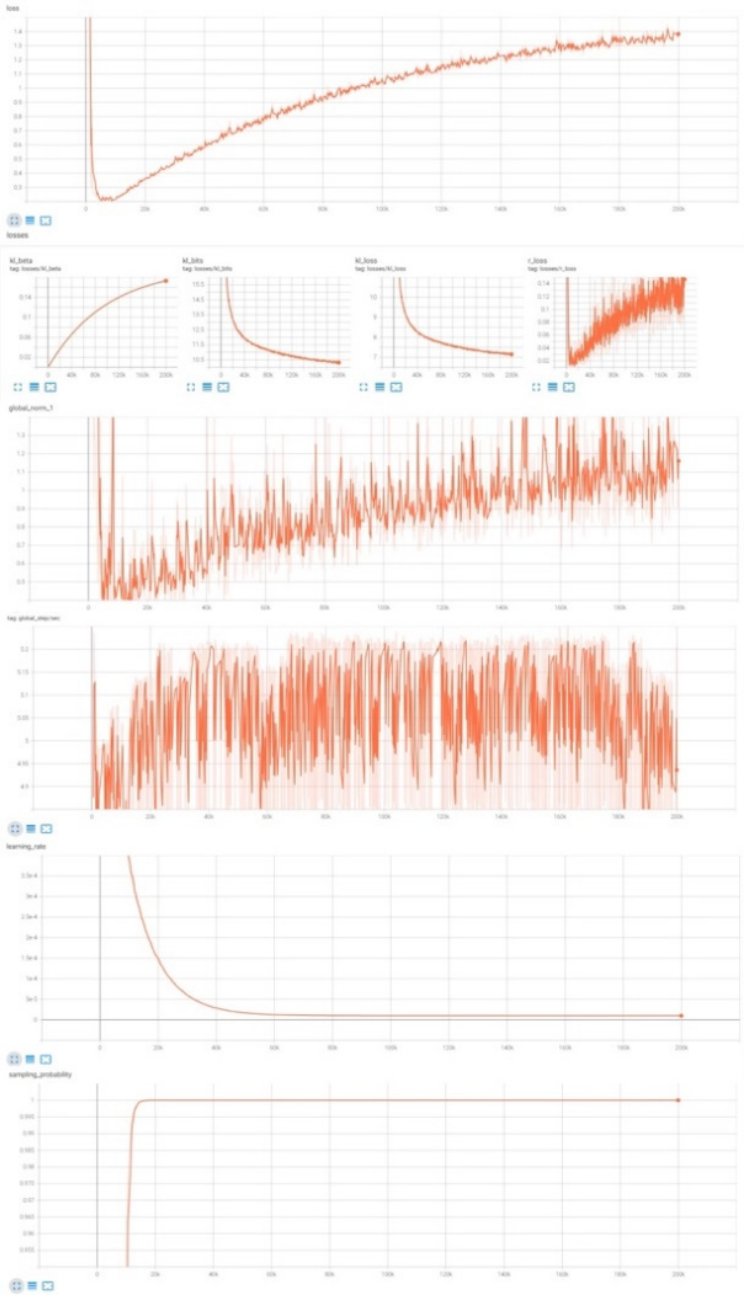

Fig 5. Training Monitoring Indicators

To address overfitting and latent space redundancy, key parameters were adjusted: z_size reduced from 256 to 32 to lower model capacity; free_bits increased to 128 for better early reconstruction; max_beta set to 0.5 to strengthen KL constraint for a more regular latent space; encoder/decoder recurrent layer sizes increased (Enc_rnn_size=[1024,1024], Dec_rnn_size=[512,512]) for enhanced temporal modeling. This balance yielded a more compact latent space, coherent chord sequences, and better emotional consistency. [14]

Training dynamics showed metric improvements. Global_norm decreased and stabilized, indicating effective gradient control and smoother convergence. Global_step progression slightly slowed mid-to-late stage, showing the model dedicated more computation to refined weight updates, with improved robustness. Linear decay for Learning_rate prevented premature overfitting, ensuring initial speed and later refinement. Sampling Probability evolved more stably, transitioning to higher auto-regressive input, enhancing self-generation and reducing overfitting risk.

Algorithmically, image preprocessing was optimized to improve contour quality and mapping stability. Binarization used Otsu's adaptive threshold with THRESH_BINARY for automatic thresholding and adaptability to various lighting. A 5×5 Gaussian blur (σ=0) then reduced noise while preserving edges, aiding cv2.findContours and CHAIN_ APPROX_ SIMPLE for sparse representation. Figure 7 illustrates binarization and Gaussian blur improvements: contour closure, smoothness, and fitability significantly enhanced, leading to more stable time series mapping to 88 keys and reduced post-processing for "skipped/broken notes."

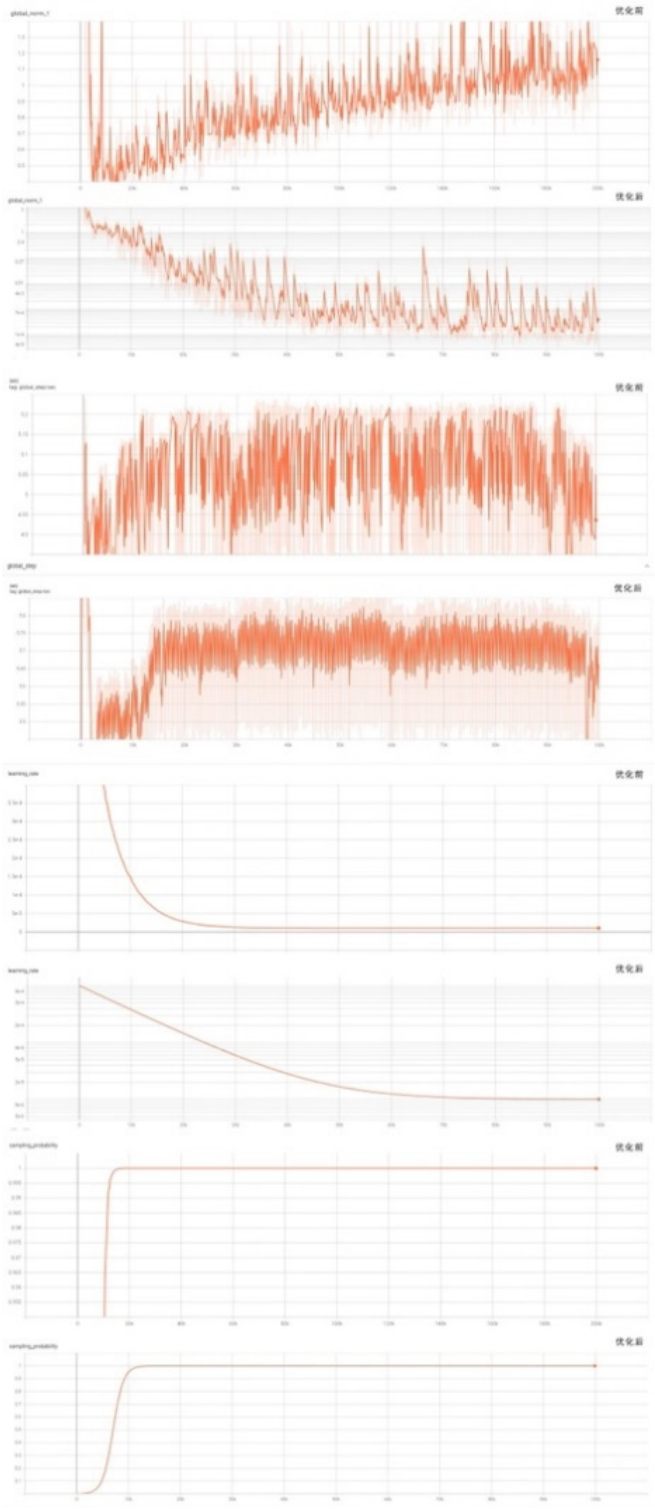

Fig 6. Training Dynamics Comparison

In summary, these improvements resulted in faster convergence and stronger generalization for the model under the same computational budget. The latent space structure was more regular, emotion-to-chord alignment more robust, and output MIDI continuity/audibility enhanced. Figure 6 summarizes training dynamics before and after optimization. Subsequent chapters will validate these optimizations' benefits in audio-visual consistency and artistic expression.

4.3. Evaluation and Analysis of Audio-Visual Effects

This study assessed the cross-modal consistency of audio-visual generation models by evaluating the harmony, color–mode coherence, and emotional correlation of generated works. A survey of 937 participants, 18% of whom had music or art backgrounds, provided diverse data.

Findings on auditory harmony indicated varied responses to chord types. C major and minor triads were rated as

harmonious (49.73% and 54.85%, respectively), while augmented and diminished triads scored about 10% lower, aligning with waveform analysis. Spearman's rank correlation showed a positive trend ($r = 0.6$) but was not statistically significant ($p = 0.285$), possibly due to the limited musical background of respondents. Harmony ratings for consonant, neutral, and dissonant chords averaged 45.08%, 30.95%, and 23.97%, respectively, following a normal distribution.

Regarding color–mode coherence, matching varied significantly. Orange–D major works achieved 44.82% audiovisual consistency, while blue–A minor works were lower at 39.68%. This suggests that individual preferences influence color–music mapping, yet the overall trend supports the model's cross-modal design.

For emotion perception, works with red hues most frequently evoked "happiness" (39.12%), while blue–violet tones prompted "sadness" or "calmness." Green and orange–yellow showed more complex, mixed emotional responses, consistent with color psychology.

Further analysis using K-Means clustering grouped samples into two categories: A (green, orange–yellow, blue–violet), characterized by higher "calmness" and lower "happiness," exhibiting complex mixed emotions; and B (red), characterized by strong, direct "happiness." Covariance analysis revealed significant differences in "happiness" and "sadness" across hues, with variances of 94.34 and 44.00, respectively, reflecting variability in individual evaluations.

In conclusion, these findings confirm the audio-visual generation model's ability to maintain visual–auditory emotional consistency. Despite some non-significant statistical results, overall trends align with perceptual psychology, demonstrating the model's potential in cross-sensory artistic generation.

5. Conclusion and Discussion

This study developed the *Harmony and Color* device, demonstrating the application of chord–color mapping in cross-sensory art. The device achieved a high degree of integration between music and visuals, hardware and software, and immersive spatial design. It successfully enabled interactive experiences where audiences received instant, matched audio-visual feedback through emotional tags. However, limitations remain in universality and scalability due to individual differences in emotional mapping, constraints of RFID trigger methods, potential delays from cloud dependency, and restrictions imposed by projection materials and hardware resolution.

The research not only provides a new practical case for interdisciplinary art creation but also highlights art's potential in emotional healing and social applications. The device's interactivity enhanced audience engagement, while the synergy of technology and materials showcased the possibilities of art–technology integration. Based on the "color–emotion–chord" mapping model, this study further combined mathematical modeling with an LSTM framework to achieve image-to-chord conversion. Experimental results indicated strong performance in harmonic coherence, emotional consistency, and artistic expressiveness. Questionnaire surveys and statistical analyses showed that the generated works closely matched visual characteristics in harmony and emotional induction; red tones more easily evoked pleasure, while blue–purple tones tended toward calmness or sadness, demonstrating the model's ability to resonate emotionally with audiences.

Overall, this research validated the theoretical feasibility and practical value of "color–emotion–chord" mapping, showing broad application prospects in multimedia art, music therapy, education, and entertainment. Future work will focus on expanding dataset scale, introducing more advanced cross-modal modeling methods, and optimizing real-time systems to enhance robustness and adaptability across scenarios, thereby providing new avenues and insights for audio-visual generation and cross-sensory art exploration.

References

[1] Shen, L. (2022). *Audiovisual Dynamics: Intellectual Construction and Artistic Evolution of Music–Image Association* (Master's thesis). China Academy of Art.

[2] Zheng, R. (2011). Harmonic Functions and Their Acoustic Principles. *Huangzhong (Journal of Wuhan Conservatory of Music)*, (04), 105–113.

[3] Li, R. (2018). *A Study on the Scientific Basis of Musical Harmony* (Master's thesis). Shanxi University.

[4] Jin, Y., & Zhou, F. (2022). Synesthetic Expression Elements of Audiovisual Interaction Design in Musical Performance. *Art Research*, (02), 95–98.

[5] Han, D., & Zhao, H. (2009). Color Synesthesia and Emotional Recognition in Conceptual Integration. *Foreign Language Research*, (1), 40–43.

[6] Eaton, J. (1999). *The Art of Color*. Beijing: World Book Publishing Company.

[7] Hua, C. (2012). *Color Harmony*. Beijing: Central Conservatory of Music Press.

[8] Guo, C., Huang, M., & Xi, Y. (2020). Effects of Primary Color Spectrum and Visual Field on Color Perception. *Spectroscopy and Spectral Analysis*, 40(12), 3765–3771.

[9] Wu, B., & Liang, X. (2023). Research on the Relationship Between Music and Visual Expression in Audiovisual Interactive Design. *Packaging and Design*, (5), 142–143.

[10] Qu, T., Huang, D., & Tong, K. (2007). A Review of Music Visualization Research. *Computer Science*, 34(9), 16–22.

[11] Wang, X. (2022). *Research on Emotion-Based Music Visualization Technology Using EEG* (Master's thesis). Changchun University of Science and Technology.

[12] Wang, L., Du, L., & Wang, J. (2007). Music Emotion Classification Based on AdaBoost. *Journal of Electronics & Information Technology*, 29(9), 2067–2072.

[13] Roberts, A., Engel, J., Raffel, C., Hawthorne, C., et al. (2018). A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. *arXiv preprint arXiv:1803.05428*.

[14] Su, J. (2021). *Research on AI Composition Practice and Theory Based on Artificial Neural Networks: A Case Study of the Keras Framework* (Master's thesis). Shanghai Conservatory of Music.