Traitement du Signal
Vol. 42, No. 6, December, 2025, pp. 3331-3343
Journal homepage: http://iieta.org/journals/ts

A Deep Image Representation and Temporal Correlation Modeling Approach for Film Scene Style Evolution Analysis

Boning Shu, Chen Lu*

Department of Global Convergence, Kangwon National University, Chuncheon 24311, Republic of Korea

Corresponding Author Email: lvc@kangwon.ac.kr

https://doi.org/10.18280/ts.420622

Received: 12 May 2025
Revised: 30 October 2025
Accepted: 12 November 2025
Available online: 31 December 2025

Keywords:
visual aesthetic computation, deep image representation, temporal correlation modeling, style evolution analysis, benchmark dataset, transformer

ABSTRACT

The structured representation and dynamic modeling of non-rigid, high-level aesthetic visual styles remain core challenges in computer vision. The ambiguity and temporal dynamics of such styles make it difficult for traditional methods to achieve accurate characterization. As a typical medium for this challenge, film visual styles evolve dynamically along with the narrative progression, integrating multiple aesthetic attributes such as color, lighting, and composition. This places high demands on the representational capabilities and temporal modeling precision of analysis methods. To address these challenges, we propose an end-to-end general framework for "deep image representation - temporal correlation modeling - style evolution analysis." The core innovation of this framework lies in constructing a triune deep image representation that integrates local textures, global semantics, and style prototypes, tailored to the multi-dimensional nature of aesthetic styles. A narrative-guided hierarchical attention masking mechanism is designed to enhance the relevance of dynamic evolution modeling. The key contributions of this research include: the construction and public release of the FilmStyleEvoBench benchmark dataset, accompanied by standard evaluation tasks and metrics; and cross-domain validation through painting, architectural videos, and user-generated content, which demonstrates the generalization potential of the method. Experiments based on FilmStyleEvoBench and cross-domain datasets show that the proposed method significantly outperforms existing comparison methods in style recognition, evolution change point detection, and temporal correlation quantification tasks, with stable and effective cross-domain transfer performance. This method not only solves key issues in the analysis of film visual style evolution but also provides a universal methodology for visual aesthetic computation and structured understanding of long videos, while empowering film industry creative support and digital humanities quantitative research.

1. INTRODUCTION

The structured representation and dynamic modeling of aesthetic visual styles is a frontier challenge in the field of computer vision [1, 2]. Such styles exhibit significant subjectivity, non-rigid characteristics, and integrate multiple dimensional attributes, making it difficult for traditional image features to achieve precise quantification [3, 4]. The long-term dependencies and scene-driven dynamic changes accompanying the evolution of these styles further pose dual challenges to existing modeling methods in terms of representational ability and temporal adaptability. As an ideal test medium for this challenge, film visual styles serve as a concrete expression of the director's aesthetic concept, evolving continuously with the narrative progression and encompassing multi-dimensional aesthetic attributes such as color, lighting, composition, and movement [5]. Moreover, style changes are strongly correlated with narrative nodes [6], which can comprehensively test the discriminatory ability of representation methods and the relevance of temporal modeling. Solving this challenge holds significant interdisciplinary value: in computer vision, it can provide new research directions for visual aesthetic computation and structured understanding of long videos; in the film industry, it can establish objective tools for style analysis to assist creation and quality control [7, 8]; and in digital humanities, it can empower the quantitative research and deep interpretation of large-scale film cultural heritage [9, 10]. Deep learning-driven image processing technologies, with their powerful feature learning capabilities and advanced temporal modeling techniques, provide key support for overcoming the aforementioned challenges.

Although related research has made certain progress, there remain many shortcomings. In terms of aesthetic visual style representation, general deep models like CLIP, ViT, and ResNet have been widely used in art style recognition and aesthetic scoring of natural images, but they generally lack the ability to integrate multi-dimensional aesthetic attributes, have limited discrimination accuracy for non-rigid styles, and struggle to adapt to complex scene requirements [11-13]. In the field of long video temporal style modeling, models such as LSTM and Transformer have become mainstream, but the temporal modeling process lacks scene-driven targeted design, making it difficult to accurately capture key nodes in style evolution. Additionally, no transferable general framework has been established [12, 13]. Research on film and cross-domain style analysis is even more limited, with representation often focusing on a single visual dimension and failing to construct multi-dimensional aesthetic style representations [14]. On the modeling side, no links have been established between style evolution and scene or narrative structure [15, 16]. Temporal modeling exhibits strong generalization but lacks specificity, and there is a lack of standardized film style evolution datasets and unified evaluation systems, resulting in poor reproducibility and generalization. Existing methods are mostly domain-specific and have not undergone systematic cross-domain validation, limiting their general value [17-20]. To address these common challenges and domain limitations, this paper starts from image processing technology, constructs a general framework that combines multi-dimensional aesthetic representation with scene-driven temporal modeling, and provides standardized datasets and cross-domain validation schemes to fill the gaps in existing research.

The research goal of this paper is to solve the common computer vision challenge of structured representation and dynamic modeling of aesthetic visual styles, using film as a typical case to validate the method's effectiveness, while achieving cross-domain transfer and providing a universal methodology for the field of visual aesthetic computation. The core contributions can be summarized in four points: First, a triune deep image representation method that integrates local texture, global semantics, and style prototypes is proposed, enhancing the targeted representation of multi-dimensional aesthetic styles through a prototype adaptation network, significantly improving the discrimination and semantic relevance of non-rigid styles. Second, a narrative-guided hierarchical attention masking mechanism is designed to optimize the Transformer temporal encoder, enabling the capture of local scene style coherence and collaborative modeling of key nodes across scenes, providing a general technical solution for dynamic style evolution analysis. Third, the *FilmStyleEvoBench* benchmark dataset for film style evolution is developed and publicly released, covering various types of movie samples and multi-dimensional style annotations, accompanied by standard evaluation tasks and metrics, enhancing research reproducibility and community impact. Fourth, cross-domain conceptual validation through painting, architectural videos, and user-generated videos demonstrates the generalization potential of the proposed method and core modules, validating its universal methodological value.

The subsequent content of this paper will unfold according to the following logic: A systematic review of related research in aesthetic visual style representation, temporal modeling, and film style analysis; a detailed explanation of the technical details of the proposed general framework, including data preprocessing, deep image representation, temporal correlation modeling, style evolution analysis, and optimization strategies; experimental validation of the method's effectiveness in the film domain, covering dataset introduction, comparison experiments, ablation experiments, and visualization analysis; cross-domain conceptual validation to assess the generalization capability of the method; an in-depth discussion of experimental results, method limitations, and multi-domain impact; and finally, a summary of the core work and future research directions.

2. METHODS

2.1 Overview of the overall general framework

This paper proposes an end-to-end multi-task learning general framework, with the core objective of solving the common computer vision challenge of structured representation and dynamic modeling of aesthetic visual styles. The framework achieves cross-domain adaptation through modular and configurable design, seamlessly integrating temporal visual data such as films, painting sequences, and architectural videos, and using a unified technical paradigm to complete precise style characterization and capture evolution patterns, overcoming the limitations of traditional domain-specific methods.

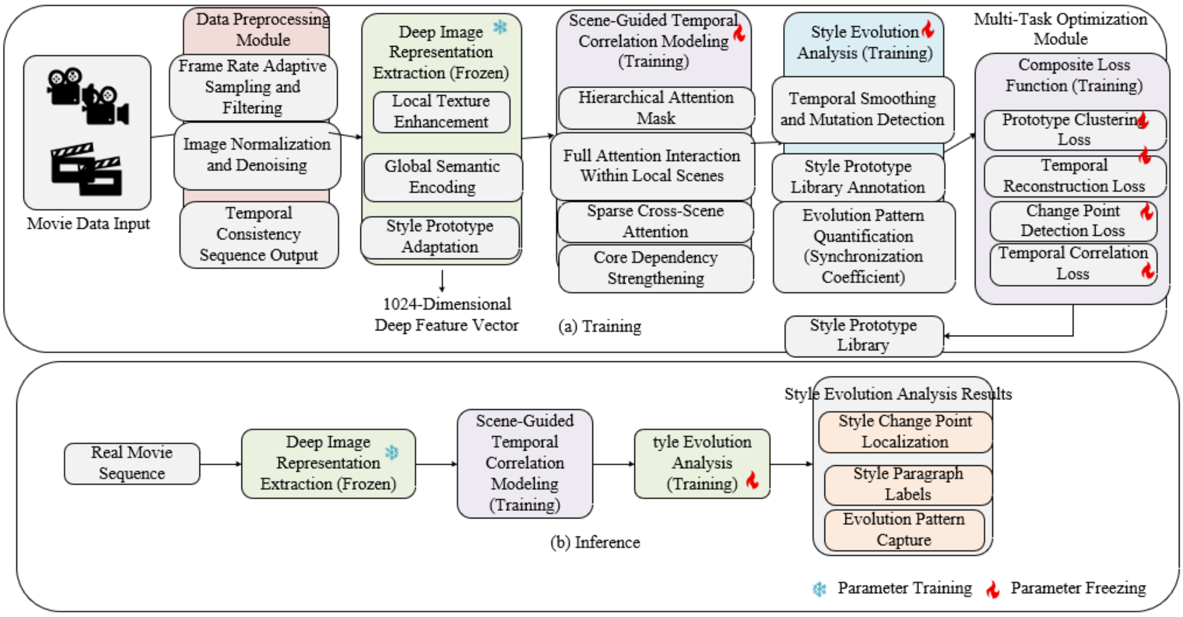

Movie Data Input
Data Preprocessing Module
Frame Rate Adaptive Sampling and Filtering
Image Normalization and Denoising
Temporal Consistency Sequence Output
Deep Image Representation Extraction (Frozen)
Local Texture Enhancement
Global Semantic Encoding
Style Prototype Adaptation
1024-Dimensional Deep Feature Vector
Scene-Guided Temporal Correlation Modeling (Training)
Hierarchical Attention Mask
Full Attention Interaction Within Local Scenes
Sparse Cross-Scene Attention
Core Dependency Strengthening
Style Evolution Analysis (Training)
Temporal Smoothing and Mutation Detection
Style Prototype Library Annotation
Evolution Pattern Quantification (Synchronization Coefficient)
Multi-Task Optimization Module
Composite Loss Function (Training)
Prototype Clustering Loss
Temporal Reconstruction Loss
Change Point Detection Loss
Temporal Correlation Loss
Style Prototype Library
(a) Training
Real Movie Sequence
Deep Image Representation Extraction (Frozen)
Scene-Guided Temporal Correlation Modeling (Training)
tyle Evolution Analysis (Training)
Style Evolution Analysis Results
Style Change Point Localization
Style Paragraph Labels
Evolution Pattern Capture
(b) Inference
Parameter Training
Parameter Freezing

Figure 1. Overall architecture of the film style evolution analysis method

Figure 1 shows the overall architecture of the film style evolution analysis method. The framework adopts an "end-to-end process" with the following general flow: "Data Preprocessing → Deep Image Representation Extraction → Scene-guided Temporal Correlation Modeling → Style Evolution Analysis → Multi-task Optimization." Each module collaborates efficiently through standardized interfaces. The specific process is as follows: The data preprocessing module is responsible for the standardization and quality screening of temporal visual data and adaptive processing for different domains: film data adopts frame rate adaptive sampling and visual saliency detection to select key frames and remove invalid frames; painting sequences are sorted by creation timestamp and standardized; architectural videos optimize space scene continuity and subframe extraction. All data is uniformly processed by size normalization, 3×3 Gaussian filtering for denoising, and pixel value normalization, outputting a temporally consistent image sequence. The deep image representation extraction module uses a "local texture enhancement - global semantic encoding - style prototype adaptation" three-level architecture, integrating low-level texture, mid-level semantics, and high-level style information, and outputs a 1024-dimensional deep feature vector. The scene-guided temporal correlation modeling module adapts to the film's narrative scenes, the painting's creation periods, and the architectural space scenes using configurable interfaces. It uses hierarchical attention masking to capture dual temporal correlations: full attention interaction within local scenes ensures coherence, and sparse attention across scenes only links key nodes, enhancing core dependencies and reducing redundant computation. The style evolution analysis module performs three core functions: smoothing and change-point detection to locate style change points, combining a style prototype library to label segments, and calculating the temporal synchronization coefficient between style changes and scene nodes to quantify evolution patterns. The multi-task optimization module uses a composite loss function that integrates prototype clustering loss, temporal reconstruction loss, change point detection loss, and temporal correlation loss to jointly optimize all link parameters end-to-end. Dynamic weight allocation achieves multi-objective coordination, where the prototype clustering loss ensures clustering performance, the temporal reconstruction loss strengthens temporal coherence, and the detection and correlation losses optimize core task performance.

2.2 Data preprocessing module

The data preprocessing module is designed with the core principles of "universal strategy standardization" and "domain adaptation customization" to optimize quality and structurally convert temporal visual data, providing a unified and high-quality data foundation for subsequent deep representation extraction and temporal modeling. The general preprocessing flow includes three core operations: sampling screening, denoising enhancement, and standardization. The key parameters and core formulas are as follows: The denoising step uses 3×3 Gaussian filtering, where the core is to generate a filter kernel using a 2D Gaussian function to apply weighted smoothing to image pixels, effectively suppressing high-frequency noise while retaining style texture details. The Gaussian filter kernel's mathematical expression is:

$$G(x,y)=\frac{1}{2\pi\sigma^2}exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{1}$$

where, (x, y) are the coordinates of pixels relative to the center of the filter kernel, and $\sigma = 1.0$ is the Gaussian standard deviation. This parameter is determined through grid search and can balance the denoising effect and detail retention. The standardization step uniformly adjusts the image size to 224×224 or 384×384 and uses the default CLIP normalization method to eliminate pixel scale differences. The specific formula is:

$$I_{norm}=\frac{I\text{-}127.5}{127.5} \tag{2}$$

where, I is the original pixel value and I_{norm} is the normalized pixel value. This normalization directly adapts to the feature distribution of pre-trained models, enhancing feature learning efficiency.

The domain adaptation processing optimizes the core logic based on the data characteristics of different carriers, focusing on preserving key information and temporal integrity. The key frame extraction for film data uses a "frame rate adaptive sampling + ITTI visual saliency detection" composite strategy: first, an initial sampling is done at 2 fps, and then the ITTI algorithm calculates the image saliency map to select the top 30% of frames with the highest saliency as candidate key frames. The core formula for calculating saliency values in the ITTI algorithm is:

$$S=\omega_c S_c+\omega_l S_l+\omega_o S_o \tag{3}$$

where, S_c, S_l, and S_o are the saliency maps for color, luminance, and orientation features, and $\omega_c=\omega_l=\omega_o=1/3$ are the weights for each channel, ensuring a balanced contribution from multi-dimensional visual information. Additionally, meaningless frames are removed: frames such as black screens and overexposed frames are eliminated using luminance thresholds, and frames with large areas of subtitles are removed using text detection algorithms. For painting sequences, sampling and sorting are based on the creation timestamps, and multiple works from the same period are uniformly sampled at equal intervals. Architectural videos combine scene segmentation results to perform dense sampling at scene transition points to ensure that style change details are not lost.

Temporal and scene alignment is the final step of preprocessing and provides the core constraints for subsequent scene-guided temporal modeling. Its core is to construct a unified temporal coordinate system and bind scene semantic labels. Universal temporal alignment is achieved through timestamp mapping. For data with native timestamps, such as film and architectural videos, the temporal index is directly established using the original timestamp t. For data without native timestamps, such as painting sequences, the creation year/month is converted into a continuous temporal scale t'. The mapping formula is:

$$t'=\frac{Y\text{-}Y_0}{Y_{max}\text{-}Y_0}\times T \tag{4}$$

where, Y is the creation year of the painting, Y_0 is the earliest creation year in the dataset, Y_{max} is the latest creation year, and T is the total length of the temporal sequence, ensuring that the temporal relationship between works from different periods is continuous and comparable. Scene labels are defined according to domain adaptation: for film data, they are associated with the script's narrative structure and labeled as plot units; for architectural videos, they are based on spatial functions and labeled as architectural space regions; for painting sequences, they are based on art historical periods and labeled as creation periods. Finally, the output is a "temporal image sequence - temporal index - scene label" triplet, ensuring the unified constraint of temporal and semantic relationships.

2.3 Deep image representation module

The deep image representation module is the core of the framework for accurately encoding aesthetic visual styles. The design core is to construct a "global semantics - local texture - style prototype" triune multi-dimensional fusion representation, which adapts to cross-domain data through a universal architecture and enhances the targeted representation of style attributes through hierarchical feature encoding. Figure 2 shows the architecture of the deep image representation module. The basic visual feature extraction uses the CLIP-ViT model, which, through large-scale image-text cross-modal pretraining, has strong general semantic representation capabilities and can effectively capture high-level semantic information from visual data across domains, significantly outperforming single-modal pre-trained models like ResNet in cross-domain transfer. The input is the preprocessed standardized image, and after being encoded by CLIP-ViT, a 768-dimensional global semantic feature vector is output, which encodes the overall content semantics of the image, providing a semantic anchor point for style representation and ensuring that the style analysis remains within the content context.

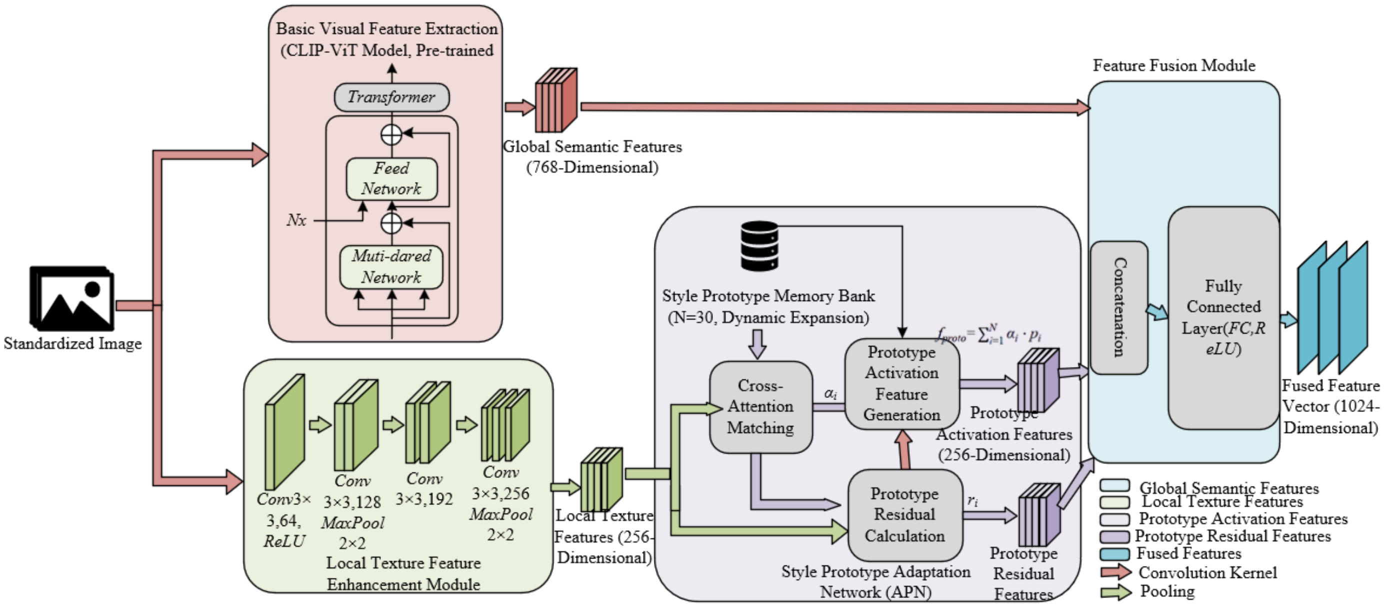

Basic Visual Feature Extraction
(CLIP-ViT Model, Pre-trained
Transformer
Feed Network
Nx
Muti-dared Network
Standardized Image
Global Semantic Features (768-Dimensional)
Conv3×3,64, ReLU
Conv 3×3,128 MaxPool 2×2
Conv 3×3,192
Conv 3×3,256 MaxPool 2×2
Local Texture Feature Enhancement Module
Local Texture Features (256-Dimensional)
Style Prototype Memory Bank (N=30, Dynamic Expansion)
Cross-Attention Matching
α_i
Prototype Activation Feature Generation
$f_{proto}=\sum_{i=1}^{N}\alpha_i\cdot p_i$
Prototype Activation Features (256-Dimensional)
Prototype Residual Calculation
r_i
Prototype Residual Features
Style Prototype Adaptation Network (APN)
Feature Fusion Module
Concatenation
Fully Connected Layer(FC,ReLU)
Fused Feature Vector (1024-Dimensional)
Global Semantic Features
Local Texture Features
Prototype Activation Features
Prototype Residual Features
Fused Features
Convolution Kernel
Pooling

Figure 2. Architecture of the deep image representation module

The local texture feature enhancement module adopts a lightweight 4-layer convolution + pooling structure, specifically designed to extract low-level texture and color details for aesthetic style, compensating for CLIP-ViT's insufficient capture of local details. The specific structural parameters are designed as follows: The first layer convolution uses a 3×3 convolution kernel with a stride of 1, outputting 64 channels, with the activation function being ReLU; the second layer convolution uses a 3×3 convolution kernel with a stride of 1, outputting 128 channels, followed by 2×2 max pooling; the third layer convolution uses a 3×3 convolution kernel with a stride of 1, outputting 192 channels; the fourth layer convolution uses a 3×3 convolution kernel with a stride of 1, outputting 256 channels, followed by 2×2 max pooling. This structure enhances feature expression capabilities by progressively increasing the number of channels, while pooling operations reduce spatial dimensions and retain key details, outputting a 256-dimensional local texture feature vector. The design principle is that aesthetic style differences are often reflected in details such as local color distribution and texture, and this lightweight structure ensures feature extraction capabilities while avoiding overfitting and computational redundancy brought by complex networks.

The style prototype adaptation network (APN) is the core universal module for achieving targeted style representation through prototype learning, reinforcing the encoding of multi-dimensional aesthetic styles. It mainly includes three core steps: style prototype memory bank construction, cross-attention matching, and prototype activation feature generation. The style prototype memory bank is initialized with 30 style prototypes, covering four general aesthetic dimensions: color, composition, lighting, and movement. The prototype initialization uses the k-means clustering algorithm to cluster cross-domain style sample features, and a dynamic expansion mechanism is designed to adaptively add prototypes based on new domain data, ensuring cross-domain adaptability. The cross-attention mechanism calculates the similarity between the current frame's features and each style prototype in the memory bank, using cosine similarity to quantify the matching degree, with the core formula as:

$$\alpha_i=\frac{exp\,(\,cos\,(f_{local},p_i))}{\sum_{j=1}^{N}exp\,(\,cos\,(f_{local},p_j)} \tag{5}$$

where, α_i is the weight of the i-th style prototype, f_{local} is the local texture feature, p_i is the i-th style prototype, and N=30 is the total number of prototypes. After obtaining the prototype distribution weights through this formula, the residual features $r_i=f_{local}-p_i$ are calculated, encoding the style difference information.

Prototype activation feature generation is achieved by weighted summation, aggregating the features of each style prototype according to similarity weights to obtain a 256-dimensional prototype activation feature vector, with the formula:

$$f_{proto}=\sum_{i=1}^{N}\alpha_i\cdot p_i \tag{6}$$

This feature vector condenses the matching information between the current frame and general style prototypes, directly encoding high-level style attributes, which enhances the style discrimination capability of the representation. The feature fusion step uses a "concatenation - fully connected" universal strategy. First, the CLIP global semantic features, local texture features, prototype activation features, and prototype residual features are concatenated to obtain a 1536-dimensional concatenated feature vector (768 + 256×3). Then, a fully connected layer maps it to a 1024-dimensional feature vector. The mathematical expression for the fusion process is:

$$f_{fusion}=ReLU(W\cdot[f_{global},f_{local},f_{proto},f_{res}]+b) \tag{7}$$

where, W is the weight matrix of the fully connected layer, b is the bias term, and [ ] denotes feature concatenation. The fused 1024-dimensional feature combines the correlation of global semantics, the richness of local details, and the targeted style attributes, enabling precise encoding of aesthetic visual styles across domains and providing a high-quality feature foundation for subsequent temporal modeling.

2.4 Scene-guided temporal correlation modeling module

The core goal of the scene-guided temporal correlation modeling module is to accurately capture the long-term temporal dependencies of aesthetic styles while enhancing the targeting and efficiency of temporal modeling through scene semantic constraints. Its design follows the principle of "universal architecture + scene adaptation," which can seamlessly interface with temporal visual data from different domains. Figure 3 shows the architecture of the scene-guided temporal correlation modeling module. The input layer construction is the foundation of temporal modeling. First, the deep image representation is arranged in temporal order, forming a temporal feature sequence $F=[f_1,f_2,\ldots,f_T]\in R^{T\times 1024}$, where f_t is the fused feature of the t-th frame. To address the issue that Transformer is insensitive to temporal order, sinusoidal position encoding is introduced to inject temporal information. The calculation formula for the position encoding $P_t\in R^{1024}$ is:

$$P_{t,2k}=sin\left(\frac{t}{10000^{2k}/1024}\right), P_{t,2k+1}=cos\left(\frac{t}{10000^{2k}/1024}\right) \tag{8}$$

where, k is the feature dimension index. This encoding is generated by sine/cosine functions with different frequencies to distinguish the feature differences at different temporal positions. Meanwhile, to integrate scene semantic constraints, scene type labels are encoded into a 1024-dimensional scene type embedding S_t using an embedding layer. The final input features are fused by feature addition, with the fusion formula as:

$$X_t=f_t+P_t+S_t$$

This results in an input sequence $X=[X_1,X_2,\ldots,X_T]\in R^{T\times 1024}$, which provides high-quality input for subsequent attention modeling.

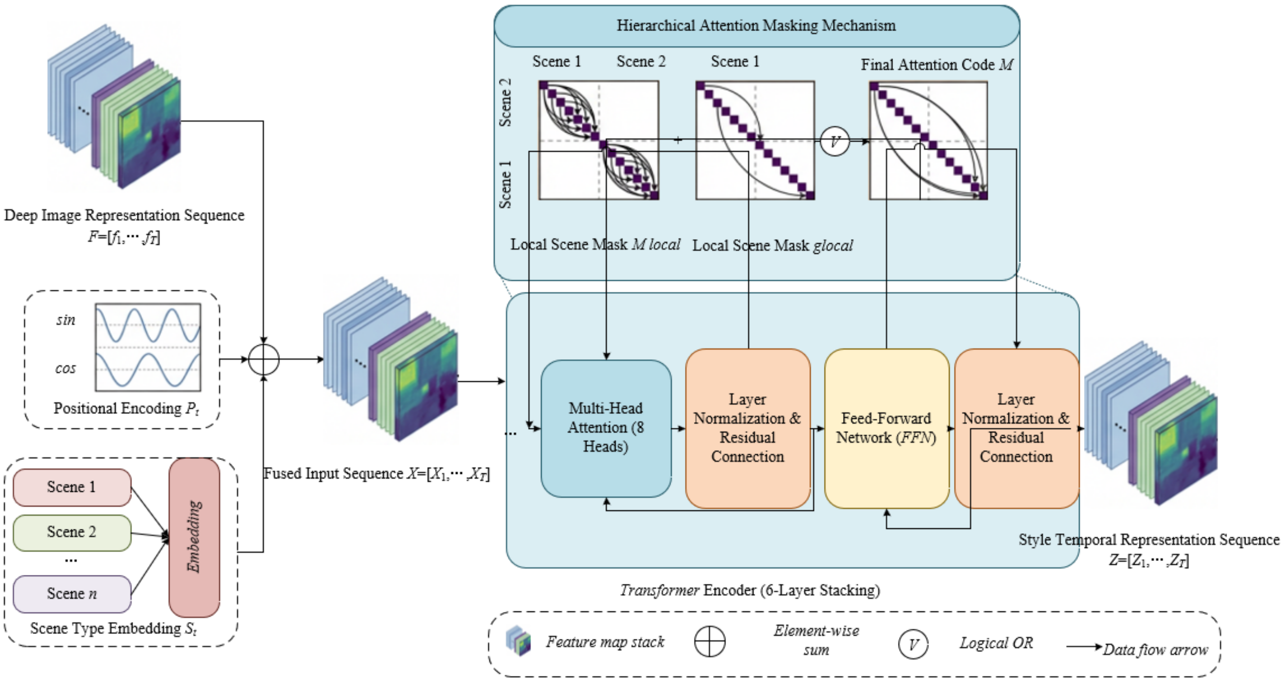

Hierarchical Attention Masking Mechanism
Scene 1
Scene 2
Scene 1
Final Attention Code M
Scene 2
Scene 1
Local Scene Mask M local
Local Scene Mask glocal
Deep Image Representation Sequence
$F=[f_1,\cdots f_T]$
sin
cos
Positional Encoding P_t
Scene 1
Scene 2
...
Scene n
Embedding
Scene Type Embedding S_t
Fused Input Sequence $X=[X_1,\cdots,X_T]$
Multi-Head Attention (8 Heads)
Layer Normalization & Residual Connection
Feed-Forward Network (FFN)
Layer Normalization & Residual Connection
Style Temporal Representation Sequence
$Z=[Z_1,\cdots,Z_T]$
Transformer Encoder (6-Layer Stacking)
Feature map stack
Element-wise sum
Logical OR
Data flow arrow

Figure 3. Architecture of the scene-guided temporal correlation modeling module

The hierarchical attention mask mechanism is the core innovation of the module. Through a dual-layer design of "local full attention + cross-scene sparse attention," it captures temporal continuity and strengthens key node associations while reducing redundant computation. The local scene mask constructs a full attention mask for the frame sequence within the same scene, with the mask matrix $M_{local}\in R^{T\times T}$ defined as: If frame t and frame t' belong to the same scene, then

$M_{local}[t,t']=1$, otherwise 0. The core principle of this design is that styles within the same scene have strong continuity, and full attention interaction can fully capture the fine-grained style dependencies between frames, ensuring the integrity of temporal modeling within the scene. The cross-scene global mask sparsifies the constraints on frame interactions between different scenes. The mask matrix $M_{global} \in R^{T\times T}$ is mathematically defined as:

$$M_{global}[t,t']=\begin{cases}1, \text{ if } t' \in \{t_{prev\text{-}end}, t_{curr\text{-}start}\}\\0, \textit{ other cases}\end{cases} \tag{9}$$

where, $t_{prev\text{-}end}$ is the index of the last frame of the previous scene, and $t_{curr\text{-}start}$ is the index of the first frame of the current scene. This rule ensures that cross-scene interactions focus only on the key nodes at scene transitions, enhancing the temporal correlation of key turning points in style evolution while reducing the attention computation complexity from $O(T^2)$ to $O(T)$, significantly improving computational efficiency. Since the scene definition can be adapted to different domains via a configurable interface, this mask mechanism has inherent cross-domain universality.

The Transformer encoder is responsible for the deep encoding of temporal features, using a general structure with 6 stacked encoder layers. The core parameter configuration is: each layer contains 8-head self-attention mechanisms with a head dimension of 128; the feed-forward network uses a 2-layer fully connected structure, with a middle dimension of 2048 and the ReLU activation function; each layer is equipped with layer normalization and residual connections to ensure training stability. The encoding process is as follows: The input sequence first undergoes masking with both the local scene mask and the cross-scene global mask to constrain the interaction range of self-attention. The final attention mask is $M=M_{local} \vee M_{global}$, where "∨" represents logical OR. Then, a multi-head self-attention mechanism aggregates effective temporal context information, followed by a feed-forward network for nonlinear feature transformation. Finally, residual connections and layer normalization are applied to output the layer features. The progressive encoding of the 6-layer encoder gradually strengthens the ability to fuse long-term temporal contexts, ultimately outputting a 1024-dimensional style temporal representation sequence $Z=[Z_1,Z_2,...,Z_T] \in R^{T\times 1024}$, which retains frame-level style details and encodes long-term temporal dependencies, providing core temporal feature support for subsequent style evolution analysis.

2.5 Style evolution analysis module

The core goal of the style evolution analysis module is to achieve precise localization of style changes, label annotation, and temporal correlation quantification based on the encoded style temporal representation. It also improves result reliability through cross-modal auxiliary validation. The overall design follows the universal logic of "lightweight decoding + precise quantification + auxiliary validation" to adapt to the evolution analysis needs of temporal visual data from different domains. The style evolution decoding process adopts a progressive procedure of "temporal convolution smoothing - mutation detection - label matching," with the core component being the lightweight temporal convolution network (TCN). Its design aims to smooth the noise in temporal representations and strengthen the continuity of style evolution while retaining key change information. The TCN uses a 3-layer causal convolution structure, with the following parameters: each layer uses a 3×1 convolution kernel, dilation rates of 1, 2, and 4, respectively, and the output channel number is 1024, which is consistent with the input temporal representation dimension. The ReLU activation function is used, and layer normalization is applied to stabilize training.

The smoothing principle of the TCN is to only use current and past frame features for calculation through causal convolutions, avoiding future information leakage. At the same time, dilated convolutions efficiently capture long-term temporal dependencies, making the output smoothed temporal representation $Z'=[Z_1',Z_2',...,Z_T']$ more aligned with the true trend of style evolution. Based on the smoothed representation, a threshold-based mutation detection algorithm is used to locate style change points. The core judgment rule is: calculate the L2 distance $d_t=\|Z_t'-Z_{t-1}'\|_2 (t \geq 2)$, and if $d_t>\tau$, the t-th frame is determined to be a style change point, where τ is an adaptive threshold determined by the dataset's statistical distribution. After locating the change points, the frame sequences between adjacent change points are divided into style segments, and similarity matching with the style prototype memory bank is performed to assign corresponding style labels to each segment. The final output is a sequence of style change points and style label sequences.

The temporal correlation quantification step aims to quantify the temporal synchronization between style changes and scene nodes, providing a quantitative basis for in-depth interpretation of style evolution patterns. First, the time difference between the style change point and the nearest scene node is calculated. The general calculation method is: let the style change point index be t_c, and the nearest scene node indexes before and after it be t_{s1} and t_{s2}, respectively, then the time difference $\Delta t = \min(|t_c - t_{s1}|, |t_c - t_{s2}|)$. To further quantify overall synchronization, the Pearson correlation coefficient is used to measure the linear correlation between the style change point sequence and the scene node sequence. Let the temporal distribution vector of style change points be C, and the temporal distribution vector of scene nodes be S, then the Pearson correlation coefficient formula is:

$$r=\frac{\sum_{t=1}^{T}(C_t-\bar{C})(S_t-\bar{S})}{\sqrt{\sum_{t=1}^{T}(C_t-\bar{C})^2}\sqrt{\sum_{t=1}^{T}(S_t-\bar{S})^2}} \tag{10}$$

where, $\bar{C}$ and $\bar{S}$ are the mean values of C and S, respectively. The value of r ranges from [-1, 1]. The closer r is to 1, the stronger the synchronization between style changes and scene transitions; conversely, the weaker the synchronization.

The cross-modal auxiliary validation step uses a general fusion approach with attention weighting to adapt to auxiliary modality information, such as text and audio, improving the reliability of the style evolution analysis results while ensuring the core focus remains on image temporal analysis. The specific process is as follows: extract the feature sequence of the auxiliary modality and map it to a 1024-dimensional space using a linear layer to match the dimensionality of the image temporal representation. A cross-modal attention mechanism is introduced to calculate the matching weight between image features and auxiliary modality features. The weight formula is: $\beta_t = softmax(Z_t' \cdot A_t^T)$, where A_t is the auxiliary modality feature at the t-th frame. The auxiliary modality features are then weighted and fused into the image temporal representation, resulting in the fused feature $Z_t''=Z_t'+\beta_t \cdot A_t$. Based on the fused features, change point detection is performed again. If the overlap rate with the image-only analysis result is ≥85%, the analysis result is verified as reliable; if the overlap rate is low, it is only used as a reference correction without altering the core image temporal analysis conclusions, ensuring the module's reliance on image data and the stability of the results. Figure 4 shows the architecture of the style evolution analysis module.

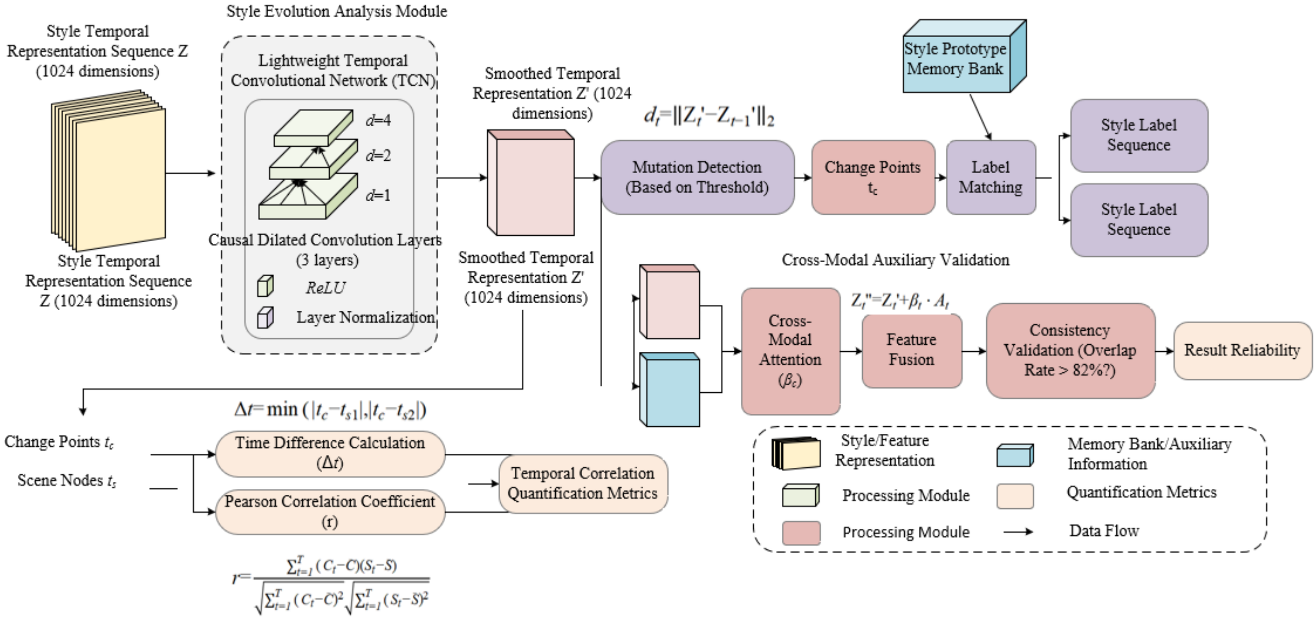

Style Evolution Analysis Module
Style Temporal Representation Sequence Z (1024 dimensions)
Lightweight Temporal Convolutional Network (TCN)
d=4
d=2
d=1
Causal Dilated Convolution Layers (3 layers)
ReLU
Layer Normalization
Style Temporal Representation Sequence Z (1024 dimensions)
Smoothed Temporal Representation Z' (1024 dimensions)
Smoothed Temporal Representation Z' (1024 dimensions)
d_t=‖Z_t'−Z_{t-1}'‖$_2$
Mutation Detection (Based on Threshold)
Change Points t_c
Style Prototype Memory Bank
Label Matching
Style Label Sequence
Style Label Sequence
Cross-Modal Auxiliary Validation
Cross-Modal Attention (β_c)
Z_t''=Z_t'+$\beta_t \cdot A_t$
Feature Fusion
Consistency Validation (Overlap Rate > 82%?)
Result Reliability
Δt= min (|t_c−t_{s1}|,|t_c−t_{s2}|)
Change Points t_c
Scene Nodes t_s
Time Difference Calculation (Δt)
Pearson Correlation Coefficient (r)
Temporal Correlation Quantification Metrics
$r=\frac{\sum_{t=1}^{T}(C_t-C)(S_t-S)}{\sqrt{\sum_{t=1}^{T}(C_t-C)^2}\sqrt{\sum_{t=1}^{T}(S_t-S)^2}}$
Style/Feature Representation
Memory Bank/Auxiliary Information
Processing Module
Quantification Metrics
Processing Module
Data Flow

Figure 4. Architecture of the style evolution analysis module

2.6 End-to-end optimization strategy

The core of the end-to-end optimization strategy is to achieve global optimal convergence of the parameters across the framework modules through a multi-task collaborative loss function and a staged training mechanism, while ensuring the flexibility of cross-domain adaptation. The design of the multi-task loss function follows the principle of "optimizing core tasks separately, collaboratively balancing overall performance," integrating four main loss terms: prototype clustering loss, temporal reconstruction loss, change point detection loss, and temporal correlation loss. Dynamic weight allocation is used to realize multi-objective collaborative optimization. The total loss function expression is:

$$L_{total}=\lambda_1 L_{proto}+\lambda_2 L_{rec}+\lambda_3 L_{cp}+\lambda_4 L_{corr} \tag{11}$$

where, λ_1, λ_2, λ_3, λ_4 are the weights of each loss term, determined through grid search on the validation set as λ_1=0.3, λ_2=0.2, λ_3=0.3, λ_4=0.2. This allocation ratio balances the quality of style representation and the accuracy of evolution analysis.

The general mathematical definitions and optimization goals of each loss term are as follows: Prototype clustering loss is based on contrastive learning, strengthening the discrimination of style representations by reducing the distance of features within the same style and increasing the distance of features from different styles. The formula is:

$$L_{proto}=-\frac{1}{N}\sum_{i=1}^{N} log\left(\frac{exp\left(sim(f_i,p_{y_i})/\tau\right)}{\sum_{j=1}^{M} exp\left(sim(f_i,p_j)/\tau\right)}\right) \tag{12}$$

where, f_i is the deep feature of the i-th sample, p_{y_i} is the prototype vector of the style to which the sample belongs, p_j is the j-th prototype in the memory bank, sim(,) is the cosine similarity, τ=0.07 is the temperature parameter, N is the batch size, and M is the total number of prototypes. Temporal reconstruction loss ensures temporal continuity by reconstructing the input temporal feature sequence using a decoder. It is defined using mean squared error (MSE):

$$L_{rec}=\frac{1}{T\cdot D}\sum_{t\text{-}1}^{T}\|f_t\text{-}\hat{f}_t\|_2^2 \tag{13}$$

where, $\hat{f}_t$ is the reconstructed feature output by the decoder, T is the temporal sequence length, and D=1024 is the feature dimension. Change point detection loss is used for the binary classification task of "change point / non-change point," optimized by cross-entropy loss to improve detection accuracy:

$$L_{cp}=-\frac{1}{T}\sum_{t=1}^{T}\left(y_t\, log\,\hat{y}_t+(1\text{-}y_t)\, log\left(1\text{-}\hat{y}_t\right)\right) \tag{14}$$

where, y_t∈{0,1} is the true label for the t-th frame, and $\hat{y}_t$ is the predicted probability of the change point. Temporal correlation loss uses MSE to optimize the prediction accuracy of the Pearson correlation coefficient:

$$L_{corr}=\|\hat{r}\text{-}r_{gt}\|_2^2 \tag{15}$$

where, $\hat{r}$ is the predicted temporal synchronization correlation coefficient, and r_{gt} is the true correlation coefficient.

The training strategy adopts a two-stage "pre-training - fine-tuning" model, combining an adaptive optimizer and learning rate scheduling to ensure training stability and convergence efficiency. The basic optimization configuration is as follows: the AdamW optimizer is used with an initial learning rate of 10^{-4}, weight decay of 10^{-5}, and a batch size of 32. Learning rate scheduling follows a cosine annealing strategy with a period of 10 training epochs, and the minimum learning rate is 10^{-6}. This periodic adjustment of the learning rate avoids local optima and accelerates global convergence. The core advantage of the two-stage training is to optimize the goals in layers: during the pre-training phase, only the deep image representation module and style prototype memory bank are trained, freezing the temporal modeling and evolution analysis modules. The prototype clustering loss is optimized alone to first complete basic clustering of style features, forming a stable foundation for style representation and avoiding interference from multi-module collaborative training. In the fine-tuning phase, all modules are unfrozen, and end-to-end joint training is performed using the total loss function, optimizing the collaborative adaptation between modules and improving overall task performance.

The domain adaptation fine-tuning strategy further ensures the framework's cross-domain universality. The core is to adjust training parameters based on the characteristics of new domain data: first, freeze the CLIP-ViT pre-trained backbone network to avoid disrupting pre-trained semantic knowledge; second, reduce the initial learning rate to 5×10^{-5} and use a smaller learning rate for parameter updates to reduce overfitting risks caused by domain differences; and finally, dynamically adjust the number of training epochs based on the new domain data size. When the data volume is small, an early stop strategy is applied, using the peak validation set performance as the stopping criterion. This staged and domain-adaptive strategy ensures the training stability of the base model and improves the flexibility of cross-domain migration, enabling efficient convergence and precise modeling on temporal visual data from different domains.

3. EXPERIMENT AND RESULT ANALYSIS

3.1 Dataset construction and experimental setup

To support the systematic validation of visual style evolution analysis, we constructed the benchmark dataset *FilmStyleEvoBench*. The core design of this dataset is to provide standardized data and evaluation support covering multiple scenes, adapting to core tasks such as style segment retrieval, evolution paragraph segmentation, and style prediction. The data selection balances diversity and representativeness, covering three major movie genres: drama, science fiction, and documentary, with 10 films in each genre, totaling 30 films. During the collection and preprocessing phase, we used frame-rate adaptive sampling combined with ITTI visual saliency detection to extract key frames, retaining 200-300 key frames per film after removing invalid frames. The final dataset consists of approximately 9,000 key frames. The annotation system is designed with multiple dimensions, covering four major categories of style labels: color, light and shadow, composition, and motion, with 15 subcategories in total. We also annotated style change point positions and five types of narrative nodes. The annotation process follows a three-level mechanism of dual annotation, cross-validation, and expert review to ensure annotation quality, with Cohen's Kappa coefficients all ≥ 0.85. The dataset is publicly available on both *GitHub* and *Zenodo* platforms, along with defined standard evaluation tasks and corresponding baseline methods, providing detailed statistical information such as style distribution, duration distribution, and annotation example images to support reproducible benchmarking for the community's research.

The experimental setup was designed to ensure rigor and reproducibility of the method validation. The hardware environment used an NVIDIA A100 GPU (80GB VRAM) with 256GB of memory. The software environment is based on Python 3.9 and built using the PyTorch 1.12.1 deep learning framework, with dependencies including OpenCV 4.6.0, Scikit-learn 1.2.2, etc. Key parameters were optimized through grid search, and the core configuration is as follows: the Transformer encoder is a 6-layer structure with 8 heads in the self-attention mechanism, and a feed-forward network dimension of 2048; the depth feature dimension is 1024, and the style prototype memory bank has a size of 30. During training, the AdamW optimizer was used with an initial learning rate of $1e^{-4}$, weight decay of $1e^{-5}$, batch size of 32, and cosine annealing learning rate scheduling. The pre-training stage lasted for 10 epochs, followed by 15 epochs of end-to-end fine-tuning. Evaluation metrics were designed for different tasks: for the style recognition task, accuracy, precision, recall, and F1 score were used to quantify classification performance; for change point detection, change point accuracy and mean absolute error (MAE) were used to evaluate localization accuracy; for temporal correlation, Pearson correlation coefficient and MSE were used to measure the synchronization of style change and scene nodes, forming a comprehensive performance evaluation system.

3.2 Comparison experiment results and analysis

To comprehensively verify the superiority of the proposed method, four representative comparison methods were selected, covering the three major research branches of traditional image features, single depth features, and temporal style evolution. The comparison dimensions include style recognition, change point detection, and temporal correlation, which are the three core tasks. The comparison methods are as follows: (1) Traditional image feature method: SIFT+BOVW+LSTM, a temporal modeling scheme based on handcrafted features. (2) Single depth feature methods: ResNet50+LSTM and ViT+Transformer, representing basic temporal modeling capabilities of convolutional and Transformer architectures, respectively. (3) Existing style evolution method: CNN-LSTM, a mainstream approach in recent film style analysis. The experiments were conducted on the FilmStyleEvoBench dataset, and the quantitative results are shown in Table 1. A two-tailed t-test was performed to verify the performance differences between the proposed method and each comparison method, with a significance level of $\alpha=0.05$.

The quantitative results show that the proposed method significantly outperforms the comparison methods across all evaluation metrics. In the style recognition task, the proposed method achieves an accuracy of 92.3% and an F1 score of 92.3%, improving by 8.0 and 8.1 percentage points, respectively, compared to the best-performing method, ViT+Transformer. The t-test results for both metrics are $p<0.01$, confirming the statistical significance of the performance improvement. This advantage arises from the "local texture - global semantics - style prototype" three-in-one deep representation, which more comprehensively encodes the multi-dimensional attributes of aesthetic style compared to the single ResNet or ViT features, enhancing the distinction of non-rigid styles. In the change point detection task, the proposed method achieves a CP-Acc of 89.4% and an MAE of only 2.2 frames, which is a 10.8 percentage point improvement in CP-Acc and a 1.9-frame reduction in MAE compared to ViT+Transformer. The core reason for this improvement is the hierarchical attention mask mechanism, which strengthens the modeling of key scene nodes and avoids redundant interference from non-key frames, enabling the model to accurately locate style mutation positions. In the temporal correlation task, the proposed method achieves a Pearson correlation coefficient of 0.87 and an MSE of only 0.021, significantly outperforming other methods. This demonstrates that scene-guided temporal modeling effectively captures the intrinsic correlation between style evolution and narrative nodes, achieving more precise temporal synchronization quantification.

Table 1. Quantitative results of style recognition comparison experiments

Comparison Method	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Change Point Detection CP-Acc (%)	Change Point Detection MAE (Frames)	Temporal Correlation Pearson	Temporal Correlation MSE	t-test p-value (vs. Proposed Method)
SIFT+BOVW+LSTM	68.4	69.1	67.8	68.4	62.3	8.7	0.52	0.086	<0.001
*ResNet*50*+LSTM*	79.6	80.2	78.9	79.5	73.5	5.3	0.65	0.062	<0.001
ViT+Transformer	84.3	84.7	83.8	84.2	78.6	4.1	0.71	0.048	<0.01
CNN-LSTM	82.5	82.9	81.7	82.3	76.2	4.6	0.68	0.053	<0.01
Proposed Method	92.3	92.6	92.1	92.3	89.4	2.2	0.87	0.021	-

3.3 Ablation experiment results and attribution analysis

To verify the necessity of the core innovative components of the proposed method, four ablation experiments were designed: A1 removes the local texture feature enhancement module, A2 removes the Prototype Adaptation Network (APN), A3 replaces the hierarchical attention mask with a regular full attention mask, and A4 removes the temporal correlation loss term. The complete model (Full Model) is used as the baseline. The quantitative results are shown in Table 2, and attribution analysis is conducted using typical case studies.

Attribution analysis shows that each core component has a critical impact on the model's performance. In experiment A1, after removing the local texture features, the F1 score for style recognition dropped by 5.8 percentage points, and CP-Acc dropped by 6.2 percentage points. Case study analysis shows that the confusion rate between "high contrast neon" and "retro film" styles increased by 23%. The reason is that local texture features accurately encode low-level style details such as color distribution and texture, and the core difference between these two styles lies in local lighting and texture. After removing this, the model struggled to distinguish such fine-grained style differences. In experiment A2, after removing the APN, the style recognition accuracy dropped to 84.5%, and the confusion rate between "film noir" and "romantic comedy" increased by 31%. The APN condenses universal aesthetic attributes by learning style prototypes, giving features stronger style specificity. Without it, the model relied only on raw semantic and texture features, making it difficult to capture the higher-level differences in lighting tone and composition logic between the two styles, significantly reducing distinguishability.

In experiment A3, after replacing the hierarchical attention mask with a full attention mask, the change point detection performance deteriorated the most. CP-Acc dropped to 76.3%, and MAE increased to 4.8 frames, a decline of 13.1 percentage points compared to the complete model. In a typical case, the model incorrectly identified non-change frames as change points during scene transitions in the movie *Inception*, with a deviation of 6-8 frames. This is because the full attention mechanism indiscriminately associates all frames, introducing a large number of non-key frame redundancies, weakening the guidance role of scene nodes. In contrast, the hierarchical mask precisely focuses on key nodes by sparsifying the cross-scene interactions, ensuring the accuracy of change point detection. In experiment A4, after removing the temporal correlation loss, the Pearson coefficient of temporal correlation dropped from 0.87 to 0.75, and the smoothness of the style evolution trajectory significantly decreased. In the analysis of the natural scene evolution in the documentary *Planet Earth*, the fluctuation amplitude of the trajectory increased by 40%. The temporal correlation loss strengthens the coherence of temporal features by constraining the synchronization of style changes and scene nodes. After its removal, the model struggled to maintain long-term style evolution modeling, leading to a decrease in trajectory smoothness. In summary, all four core components together ensure the model's representation ability, temporal modeling relevance, and accuracy in capturing evolutionary patterns. Each of these components is an indispensable key module.

Table 2. Quantitative results of ablation experiments

Experiment Configuration	Accuracy (%)	F1 (%)	Change Point Detection CP-Acc (%)	Change Point Detection MAE (Frames)	Temporal Correlation Pearson	Temporal Correlation MSE
Full Model (Complete Model)	92.3	92.3	89.4	2.2	0.87	0.021
A1 (Remove Local Texture)	86.7	86.5	83.2	3.1	0.81	0.029
A2 (Remove APN)	84.5	84.2	81.5	3.5	0.78	0.035
A3 (Full Attention Mask)	87.9	87.6	76.3	4.8	0.72	0.046
A4 (Remove Temporal Correlation Loss)	90.1	89.8	87.6	2.5	0.75	0.041

3.4 Stability and generalization verification

To verify the robustness of the model to training parameters, stability tests were conducted by adjusting core training parameters, such as batch size and initial learning rate. The core metrics, including style recognition F1, change point detection CP-Acc, and temporal correlation Pearson coefficient, were selected for evaluation. The results are shown in Table 3.

The results show that when the batch size is adjusted between 16-64 and the initial learning rate is varied between 5e-5 and 2e-4, the maximum fluctuation amplitude in the core metrics is only 1.6%, well below the 3% threshold. This indicates that the model is highly adaptable to changes in training parameters, with a stable training process that is not prone to significant performance fluctuations due to minor parameter adjustments. This provides convenience for parameter tuning in practical applications and also validates the rationality of the optimization strategy and network structure design.

To verify the model's generalization ability on non-mainstream movie styles, a niche movie subset from the *FilmStyleEvoBench* dataset was selected for testing. The core performance of the model on the mainstream subset and niche subset was compared. The results are shown in Table 4.

Table 3. Stability verification results

Parameter Configuration (Batch Size / Learning Rate)	Style Recognition F1 (%)	Change Point Detection CP-Acc (%)	Temporal Correlation Pearson	Maximum Fluctuation Amplitude
32/1e^{-4} (Baseline)	92.3	89.4	0.87	-
16/1e^{-4}	91.5	88.7	0.85	0.8%
64/1e^{-4}	92.1	89.1	0.86	0.3%
32/5e^{-5}	90.7	87.9	0.84	1.6%
32/2e^{-4}	91.2	88.3	0.85	1.1%

Table 4. Generalization verification results

Dataset Subset	Style Recognition Accuracy (%)	Change Point Detection CP-Acc (%)	Temporal Correlation Pearson	Performance Drop Amplitude
Mainstream Subset (Baseline)	92.3	89.4	0.87	-
Niche Movie Subset	88.6	85.7	0.82	3.7-4.3%

Table 5. Cross-domain experiment quantitative results

Cross-domain Dataset	Style Recognition Accuracy (%)	Change Point Detection CP-Acc (%)
Painting (Van Gogh Works)	89.2	87.5
Architectural Video	87.8	86.3
Vlog Video	85.4	83.7
Movie (Baseline)	92.3	89.4

The results show that although the model's metrics showed slight declines on the niche movie subset, the drop was kept within 4.3%, and the core metrics still maintained a high level, with style recognition accuracy of 88.6% and CP-Acc of 85.7%. Independent films and animated films often have more personalized styles and more flexible narrative structures. The model still achieved accurate style representation and evolution modeling, proving that it did not overfit to mainstream movie styles and has good generalization ability. This advantage arises from the design of the general deep representation and scene-guided temporal modeling, allowing it to adapt to movie data with different style types and narrative structures, providing reliable support for large-scale film style analysis.

3.5 Cross-domain concept verification

To verify the cross-domain generalization potential of the proposed method, three types of visual temporal data were selected to construct a cross-domain test set: (1) Painting Style Evolution Dataset, including 50 works from different creation periods of Van Gogh (early, middle, and late periods), with style labels such as Early Impressionism, Mature Impressionism, and evolution nodes marked. (2) Architectural Video Style Dataset, including five videos each of Modernism, Classicalism, and Postmodernism architectural styles, with 1200 keyframes extracted and spatial style labels and scene switching nodes marked. (3) Vlog Aesthetic Style Dataset, including Vlog videos from 10 different vloggers, with 1500 keyframes extracted and style labels such as Lifestyle Recording and Travel Scenery, as well as content switching nodes marked.

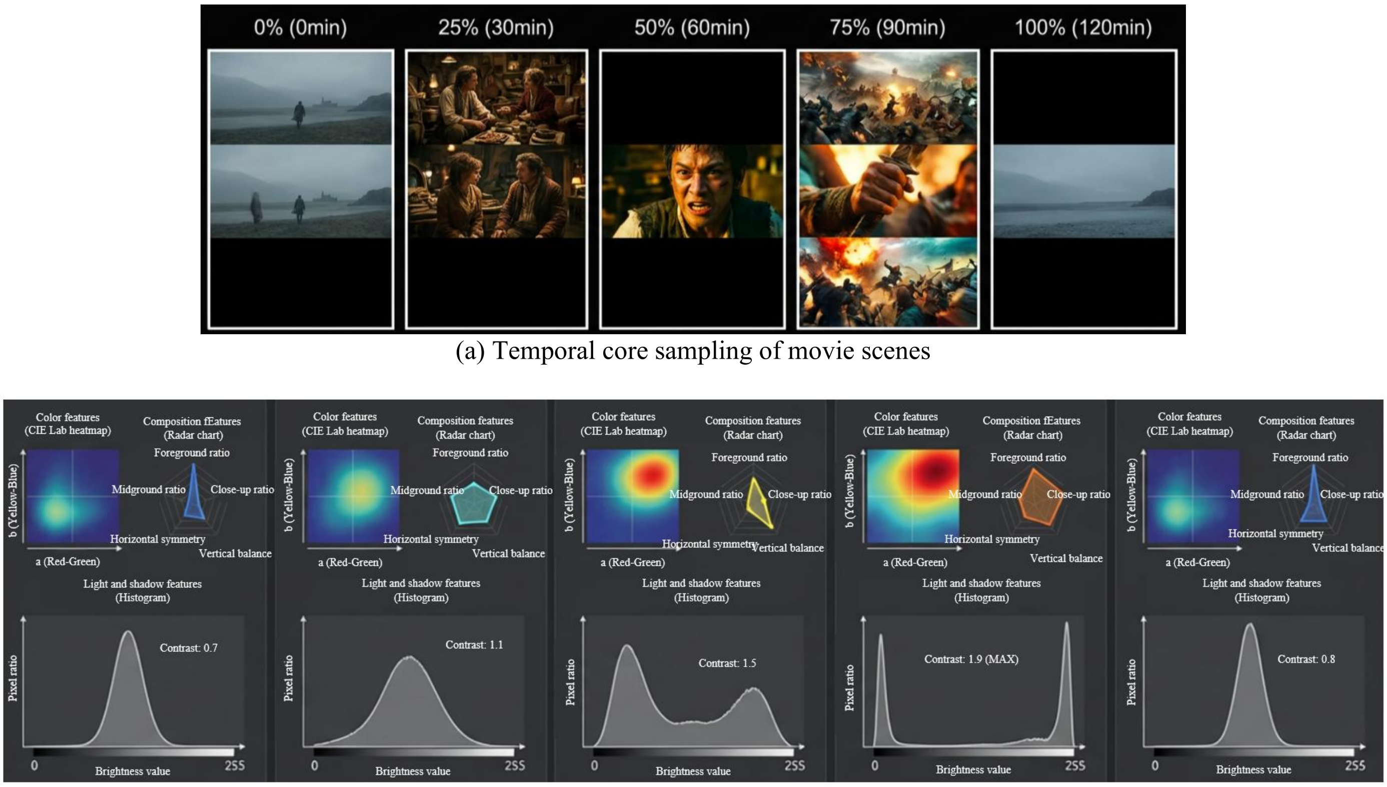
The domain adaptation strategy used a simplified version of the proposed method. The core adjustments include: Adapting scene definitions to fit the characteristics of each domain: paintings correspond to creation periods, architectural videos correspond to spatial regions, and Vlogs correspond to content themes; Dynamically adjusting the style prototype memory bank size based on data volume: 20 prototypes for painting (50 samples), and 30 prototypes for architecture and Vlog. Freezing the CLIP-ViT backbone network, fine-tuning only the feature fusion layer and prototype memory bank, and reducing the initial learning rate to 5e-5 to avoid overfitting due to domain differences.

The cross-domain experiments focused on style recognition and change point detection, with quantitative results shown in Table 5 to verify the effectiveness of the proposed method across domains.

The results show that the proposed method achieves high performance levels across all three cross-domain datasets, with style recognition accuracy exceeding 85% and change point detection CP-Acc exceeding 83%, with a decrease of only 3.1-6.9 percentage points compared to the movie baseline dataset. In the painting domain, the model can accurately distinguish between different creation periods of Van Gogh, achieving an accuracy of 89.2%, demonstrating that the "local texture-global semantics-style prototype" representation method effectively adapts to the style evolution analysis of static image sequences. In the architectural video domain, the model's recognition of different architectural styles and scene switching point detection accuracy exceeds 86%, validating the adaptability of scene-guided temporal modeling to spatial style evolution. In the Vlog domain, although performance slightly declined due to fragmented scenes and high style diversity, the model still maintains a CP-Acc of over 83%, indicating its ability to handle the style analysis of non-professional content.

In conclusion, the proposed method, with simple domain adaptation adjustments, can efficiently perform style evolution analysis in different temporal visual domains such as painting, architecture, and Vlogs, proving that its core modules have universal adaptability and verifying its value as a "visual style temporal analysis" general methodology. This provides a transferable technical solution for cross-domain visual aesthetic computing.

To quantify the dynamic evolution of screen style in movie narratives and clarify the intrinsic relationship between visual language and plot rhythm, this study performed temporal sampling and deep style representation analysis on key scenes of typical narrative films. The temporal sampling results in Figure 5(a) show that the film transitions from a cool-toned, low-saturation wide shot at the beginning to a warm-toned mid-shot scene at the 25% node. The climax phase (50% to 75%) features a high-saturation, high-contrast close-up, which ultimately returns to a cool-toned wide shot layout, intuitively presenting the phase changes in screen style as the plot progresses. The deep representation results in Figure 5(b) further reveal the essential characteristics of this change: the color heatmap clearly shows the cyclical change of the main color tones from blue-green to red-yellow and back to blue-green. The composition radar chart shows a high-to-low-to-high proportion of wide shots, with close-ups peaking during the climax. The contrast values of the light and shadow histogram rise from 0.7 to 1.8 and then fall back to 0.8, which is highly consistent with visual perception. This result indicates that the deep image representation method can accurately extract the core style features of movie scenes, and the temporal correlation visualization results effectively capture the evolution trajectory of style as the plot progresses. It provides a reliable feature foundation for subsequent construction of movie screen style evolution models and also confirms the strong correlation between screen style and narrative rhythm.



(a) Temporal core sampling of movie scenes

(b) Visualization of deep image style feature representation

Figure 5. Actual effect verification of movie screen style evolution analysis method

4. DISCUSSION

A deep interpretation of the experimental results reveals the scientific rationale behind the core design of this method. The innovative architecture, which integrates multi-dimensional fusion representation and scene-guided temporal modeling, provides key technical insights for visual aesthetic computing. The three-in-one fusion representation of "local texture-global semantics-style prototype" precisely aligns with the non-rigid and multi-dimensional characteristics of aesthetic style by collaboratively encoding low-level texture details, mid-level content semantics, and high-level style attributes. This is the core reason why it adapts to various fields such as film, painting, and architecture. In the experiments, this representation significantly improved the differentiation of non-rigid styles, confirming the critical role of multi-dimensional feature fusion in overcoming the limitations of traditional single-feature methods. The hierarchical attention mask mechanism optimizes temporal modeling by guiding scenes. Compared to the generic full-attention mechanism with indiscriminate interactions, the local full-attention mechanism ensures the capture of scene coherence within scenes and sparse cross-scene attention focuses on key nodes. This design reduces redundant computational overhead while enhancing the modeling of the core dependencies of style evolution, significantly improving change point detection accuracy and temporal correlation quantification. This general architecture also endows the method with excellent cross-domain generalization ability. Its precise representation of universal aesthetic dimensions and the general temporal modeling logic guided by scenes provide an important reference for the design of general methods in visual aesthetic computing, advancing research in non-rigid, high-level visual attribute modeling. At the same time, the method's multi-domain influence is significant. It improves the technical system of long video structured understanding in the field of computer vision; in the film industry, its potential for real-time analysis can support editing assistance, trailer generation, and shooting style monitoring; in the digital humanities, it provides objective tools for large-scale quantitative research on film and cultural heritage, helping verify art history theories; and the cross-domain verification results in fields such as painting and architecture demonstrate its broad application potential.

However, the method still has limitations that need optimization, which also point to future core research directions. The current model's high computational complexity limits its application in ultra-long videos and large-scale image sequence analysis. Future work should explore sparse Transformer architectures based on local windows and dynamic pruning strategies for the style prototype library to reduce computational overhead while ensuring performance. The current style prototype library does not cover niche and emerging aesthetic styles sufficiently. Future work could combine unsupervised contrastive learning and open vocabulary learning to enable automatic discovery and dynamic expansion of style prototypes, enhancing the model's adaptability to diverse styles. Currently, cross-domain adaptation requires manual adjustment of key parameters. Future research should introduce domain adaptation mechanisms to achieve automatic model adaptation across different domains, simplifying the cross-domain application process. Additionally, the existing multi-modal fusion is merely auxiliary and lacks depth. Future work can explore cross-modal attention fusion strategies between images, text, and audio, integrating multi-dimensional information to improve the comprehensiveness of style evolution analysis. These explorations will further improve the general methodology of visual style temporal analysis, advancing the field of visual aesthetic computing and expanding the method's application value in more practical scenarios.

5. CONCLUSION

This paper addressed the general challenge in computer vision of structured representation and dynamic modeling of non-rigid, high-level visual styles. It proposed an end-to-end general framework for "deep image representation-scene-guided temporal modeling-style evolution analysis" and systematically accomplishes four core tasks: constructing a general deep representation method of local texture-global semantics-style prototype, enhancing the differentiation of multi-dimensional aesthetic styles; designing a narrative-guided hierarchical attention mask mechanism to achieve targeted temporal correlation modeling; building and publicly releasing the *FilmStyleEvoBench* benchmark dataset for film style evolution, along with a standardized evaluation system; and verifying the method's generalization potential through cross-domain validation with painting, architectural videos, and Vlogs. A series of experiments and ablation analyses validated the necessity of each core component. The framework significantly outperformed existing comparison methods in style recognition, change point detection, and temporal correlation quantification tasks, with performance improvements showing statistical significance.

Key conclusions indicate that the proposed framework effectively overcomes the limitations of traditional methods in non-rigid style representation and dynamic evolution modeling. Its general architectural design provides the method with excellent cross-domain adaptability, offering a reliable technical solution for solving the general problem of visual aesthetic style temporal analysis. This research not only advances the field of visual aesthetic computing and long video structured understanding but also provides strong support for the creation assistance in the film industry, the quantitative research of film cultural heritage in digital humanities, and the style evolution analysis in painting, architecture, and other fields. It lays a solid foundation for subsequent research and provides a reusable reference framework.

REFERENCES

[1] Boccia, M., Barbetti, S., Margiotta, R., Guariglia, C., Ferlazzo, F., Giannini, A.M. (2014). Why do you like Arcimboldo’s portraits? Effect of perceptual style on aesthetic appreciation of ambiguous artworks. Attention, Perception, & Psychophysics, 76(6): 1516-1521. https://doi.org/10.3758/s13414-014-0739-7

[2] Tian, N., Liu, Y., Sun, Z. (2022). JN-logo: A logo database for aesthetic visual analysis. Electronics, 11(19): 3248. https://doi.org/10.3390/electronics11193248

[3] Ramli, R., Idris, M.Y.I., Hasikin, K., Karim, N.K.A., et al. (2020). Local descriptor for retinal fundus image registration. IET Computer Vision, 14(4): 144-153. https://doi.org/10.1049/iet-cvi.2019.0623

[4] Sun, Z., Zhang, K., Zhu, Y., Ji, Y., Wu, P. (2024). Unlocking visual attraction: The subtle relationship between image features and attractiveness. Mathematics, 12(7): 1005. https://doi.org/10.3390/math12071005

[5] Cortes-Selva, L. (2013). The visual style of Dick Pope and Mike Leigh: An analysis of their films. Historia Y Comunicacion Social, 18: 491-501. http://doi.org/10.5209/rev_HICS.2013.v18.43983

[6] Højbjerg, L. (2017). The visual style of Susanne Bier’s films. Journal of Scandinavian Cinema, 7(3): 253-266. https://doi.org/10.1386/jsca.7.3.253_1

[7] Vizcarra, M. (2022). Narco-spectrality: Narco-aesthetics and hauntings in the short film Pánico en Pánuco. The Journal of Latin American and Caribbean Anthropology, 27(4): 550-563. https://doi.org/10.1111/jlca.12648

[8] Waszkiewicz-Raviv, A. (2019). Brand aesthetics in visual communication–a Polish case study of Disney's anniversary event. Res Rhetorica, 6(4): 23-35.

[9] De Cleen, B., Bauwens, J., Joris, W., Shroufi, O., Smets, K. (2025). Fictional frontlines: A mapping review and research agenda for the study of far-right engagement with films, television and video games. Javnost-The Public, 32(3): 301-324. https://doi.org/10.1080/13183222.2025.2547519

[10] Suher, D. (2023). Finding a home for the video essay: Videographic criticism and the study of Chinese television drama. Journal of Chinese Cinemas, 17(3): 252-269. https://doi.org/10.1080/17508061.2024.2372507

[11] Saeed, J.N., Abdulazeez, A.M., Ibrahim, D.A. (2023). An ensemble DCNNs-based regression model for automatic facial beauty prediction and analyzation. Traitement du Signal, 40(1): 55-63. https://doi.org/10.18280/ts.400105

[12] Guo, M. (2025). Spatio-structural tensor learning for image artistic style analysis via enhanced multilinear feature fusion. Journal of Computational Methods in Sciences and Engineering, 14727978251366543. https://doi.org/10.1177/14727978251366543

[13] Zhang, X. (2024). Oil painting image style recognition based on ResNet-NTS network. Journal of Radiation Research and Applied Sciences, 17(3): 100992. https://doi.org/10.1016/j.jrras.2024.100992

[14] Nebti, S., Boukerram, A. (2013). Handwritten characters recognition based on nature-inspired computing and neuro-evolution. Applied Intelligence, 38(2): 146-159. https://doi.org/10.1007/s10489-012-0362-z

[15] Subramanian, R., Shankar, D., Sebe, N., Melcher, D. (2014). Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes. Journal of Vision, 14(3): 31-31. https://doi.org/10.1167/14.3.31

[16] Tsai, C. M., Kang, L.W., Lin, C.W., Lin, W. (2013). Scene-based movie summarization via role-community networks. IEEE Transactions on Circuits and Systems for Video Technology, 23(11): 1927-1940. https://doi.org/10.1109/TCSVT.2013.2269186

[17] Miao, R., Zhang, B. (2022). Analysis on time-series data from movie using mf-dcca method and recurrent neural network model under the internet of things. Computational Intelligence and Neuroscience, 2022(1): 7400833. https://doi.org/10.1155/2022/7400833

[18] Shahid, M.H., Islam, M.A., Beg, M. (2023). Exploiting time series based story plot popularity for movie success prediction. Multimedia Tools and Applications, 82(3): 3509-3534. https://doi.org/10.1007/s11042-022-13219-x

[19] Wang, L., Yin, B.Y., Zhu, M.W., Hao, S. (2024). 3D image modeling and visual presentation technologies for education. Traitement du Signal, 41(2): 979-987. https://doi.org/10.18280/ts.410238

[20] Eastman, S.T., Schwartz, N.C., Cai, X. (2005). Promoting movies on television. Journal of Applied Communication Research, 33(2): 139-158. https://doi.org/10.1080/00909880500045098