Att-BiLSTM-GAN: A Temporal Coherence-Preserving GAN Framework for Dynamic Art Video Stylization

Shujie Yu
College of Art and Design, Yantai Institute of Science and Technology, YanTai, ShanDong 264000, China
E-mail: scalarewisdom@126.com

Keywords: dynamic art video generation, style transfer, artistic texture preservation, video stylization, digital creativity, attention-enriched bidirectional long short-term memory integrated with generative adversarial networks (Att-BiLSTM-GAN)

Received: September 19, 2025

Dynamic art video generation has emerged as a significant research area in computer vision and digital creativity, enabling the transformation of ordinary videos into visually compelling artistic content. However, existing methods often struggle to maintain temporal coherence, preserve motion integrity, and ensure faithful style transfer across consecutive frames, leading to artifacts such as flickering and content leakage. Specifically, an Attention-enriched Bidirectional Long Short-Term Memory integrated with Generative Adversarial Networks (Att-BiLSTM-GAN) method is introduced for dynamic art video generation. Dynamic action painting styles data are collected, featuring bold brushstrokes, splashes, and dynamic motion-like artistic patterns. Histogram Equalization (HE) is applied during preprocessing to normalize illumination and enhance contrast, while a Visual Geometry Group 16 (VGG16)-based encoder extracted spatial and textural features from content and style references. GAN-based style transfer techniques are applied to impose artistic attributes from reference artworks, preserving both global style patterns and local texture details. The incorporation of BiLSTM reduces temporal distortions by modeling frame-to-frame dependencies, while adaptive style loss functions balance stylistic richness. Experimental evaluations are conducted on the Dynamic Action Painting Styles dataset, with state-of-the-art baselines such as StyleMaster, SRCNN, VDSR, SRResNet, EDSR, TT-VSR, and baseline GAN models, the proposed Att-BiLSTM-GAN achieved superior results with PSNR (36.21), SSIM (0.96), CSD-Score (0.94), CLIP-Text (0.88), Motion Smooth (0.91), and FID (35.42), confirming significant performance gains and improved temporal coherence across all evaluation metrics. This research highlights the potential of combining sequential learning with style transfer for generating high-quality dynamic art videos.

Povzetek: Raziskava predstavlja novo metodo za pretvorbo običajnih videov v dinamične umetniške videe z boljšim prenosom sloga in bolj tekočim gibanjem brez utripanja.

1 Introduction

Artificial intelligence (AI) is changing new media art through hybridizing the act of creative expression and digital technologies to define new aesthetic modes of visual communication and improve cultural experiences and embedded design processes that offer better visual representations, emotional impacts, and aesthetics [1]. Within the development of video synthesis technology that has the potential for applications such as text-to-video generation, video editing, and video with 3D awareness, deep generative models, Generative Adversarial Networks (GANs), and Diffusion Models (DM) also raise important concerns [2]. Generative AI is supported in creative practices through the generation of artistic content such as text, images, music, video, and code, and affecting the creative sectors that are central to socio-cultural and economic development [3]. Video style transfer is highly focused on temporal consistency, as the visual quality among successive frames, artistic effects, and immersive experiences in film, television, games, and digital art derive from maintaining coherence across frames [4]. Text-to-video frameworks that extend text-to-image models make the development of NN-based video synthesis practical. The combination of big data and unsupervised learning allows such frameworks to model movement, behavior, and interaction in highly authentic ways, facilitating emotive video synthesis from simple text descriptions [5]. Content, along with detail-rich style, can be combined in creative ways with NST to go beyond artistic style transfer to provide precise aesthetic enhancement in visual media. Artistic style can be refined, honest, and contextually expressive [6]. Current dynamic video art creation suffers from fixed-crop, alignment

challenges, image size issues, unrealistic identity realities, and partial reconstructions, all of which prevent natural movement, style authenticity, and high-end artistic transformations in creative video [7]. Despite significant improvements with text-guided diffusion frameworks and attention-based processes, flickering artifacts, incoherent frame consistency, and temporally unstable modeling continue to limit the visual quality and temporal stability of dynamic video generation [8]. In spite of the tremendous progress made in the generation of videos by AI, blending artistic creativity with intelligent systems remains significantly limited. It can hamper smoothly working workflows, feature diversification, and a completely optimized user experience across content creation, innovation, and efficiency in digital media production [9].

1.1 Research objective

Most existing systems that generate dynamic art videos suffer from several key issues: flickering, motion distortion, and poor style fidelity. This paper proposes an improved dynamic art video generation algorithm that combines DL with style transfer to establish the Attention-enriched Bidirectional Long Short-Term Memory integrated with Generative Adversarial Networks method, improving temporal coherence (TC), motion fidelity, and artistic style propagation.

1.2 Research questions

- Is there a substantial enhancement in TC with the attention mechanism incorporated into the BiLSTM over the baseline BiLSTM?
- Is there an improvement in artistic fidelity and visual realism, while maintaining smoothness of motion, with GAN-based style transfer incorporated?
- To what degree does the proposed method maintain frame-to-frame uniformity with dynamic motion and variations in illumination?

The rest of the paper was organized as follows: the background and the limitations of the current approaches were described in Section 1; a review of the relevant literature was given in Section 2; and Section 3 provided comprehensive information on data, preprocessing, feature extraction, and the suggested Att-BiLSTM-GAN approach. The experimental results and findings were reported in Section 4, and conclusions, limitations, and future directions were discussed in Section 5.

2 Related works

The development of an Att-BiLSTM-GAN framework that increases TC, motion quality, and creative style evenness in dynamic art video creation is summarized in Table 1.

Table 1: Summary of related works

Citation	Aim	Methods	Key Features	Limitations	Results
Kashyap et al., [10]	Content-style separation & texture	CNN with VGG19	Adjustable style weight	Long processing, limited styles	surpasses existing approaches in both segmentation accuracy (99.49%) and computational efficiency (91.28%)
Gong & Zu [11]	High-definition video	Temporal Transformer-based Video Super-Resolution (TT-VSR) method	Enhances reconstruction quality, detail recovery, and adaptability to complex scenes	memory usage limit for real-time deployment	Ensured superior clarity and structural fidelity
Fengxue et al., [12]	Image style transfer	Transformer-based STLTSF	Separate content & style encoders	CNN's limits on content retention	Better than CNN-based methods
Chen [13]	Enhance visual continuity	Genetic Algorithms (GA) with wavelet transform-based splicing	Optimizes high- and low-frequency coefficients, Laplace sharpness, and local variance	Higher computational cost	Achieved seamless splicing, richer edge details
Wang & Yue, [14]	Refine graphics	CAD + DL	Feature detection, point cloud optimization	Traditional methods limited	Improved reconstruction & efficiency
Liu, [15]	Assess color & style transfer	Interdisciplinary applications	Focus on appearance & perception	Consistency & adaptability issues	Supports diverse expressive types
Fu, [16]	Image art style transfer	GAN with content & style encoders	Multi-scale discriminators	Variable style, instability, artifacts	Effective implementation

Shi et al., [17]	Line art video coloring	Deep architecture with temporal refinement	Temporal & color transform	Target-reference variance	Improved temporal consistency
Ho et al., [18]	HD text-conditioned video generation	Imagen Video (diffusion cascade)	Progressive distillation, classifier-free guidance	Architecture & resolution limits	High fidelity & creative production
Lu & Wang, [19]	Universal video style transfer	CSBNet	Crystallization, separation, blending	Temporal consistency issues	Better outputs & temporal stability
Wang et al., [20]	Stylizing Chinese paintings	Controlled GAN	Stroke & ink diffusion with style/content	Flickering, fixed styles	Improved visual quality & temporal consistency
General Style Transfer Frameworks	Fixed single-style reference; limited adaptability	Adaptive loss formulation allows balanced content–style learning and stable convergence			

2.1. Research gaps

Temporal latent reconstruction increased consistency but relied too much on latent codes [13]. Line art coloring increased refinement but suffered from stylistic mismatches [17], whereas Imagen Video gained in controllability but encountered architectural scaling difficulties [18]. To overcome such limitations, this research suggested an Att-BiLSTM-GAN method to enhance the temporal consistency, motion fidelity, and artistic style propagation in dynamic art video generation. Table 2 compares the previous NST, GAN, and temporal modeling approaches, showing their inability to maintain TC and artistic detail, and how the proposed approach effectively addresses these challenges.

Table 2: Comparison of limitations in prior methods and strengths of Att-BiLSTM-GAN

Category / Method	Limitations of SOTA Methods	How Att-BiLSTM-GAN Addresses It
NST (CNN/Transformer) [10,12]	Poor temporal consistency; frame-level stylization causes flickering	BiLSTM models sequential dependencies for smooth temporal transitions
GAN-based Stylization [16,20]	Inability to preserve fine artistic details and motion alignment	PatchGAN discriminator enhances local texture realism while maintaining motion integrity
Temporal Modeling Methods [13,17]	Heavy computation; weak adaptability to dynamic motion	The attention layer selectively focuses on motion-relevant frames for efficient TC

3 Methodology

An improved dynamic art video generation algorithm combined DL with style transfer, employing HE for preprocessing, VGG16 feature extraction, and an Att-BiLSTM-GAN method that incorporated GAN-based style transfer and BiLSTM for coherent, content-preserving, and stylistically consistent video sequences. Figure 1 shows the overall process of the Att-BiLSTM-GAN method.

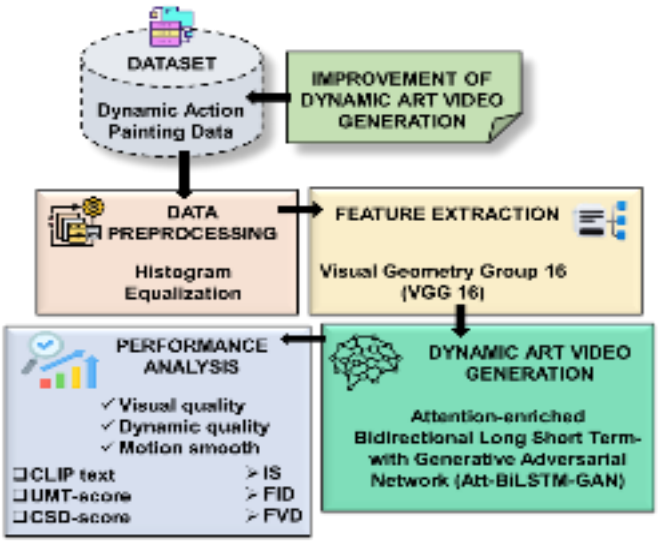

Figure 1: Overall workflow of the Att-BiLSTM-GAN framework

3.1. Data collection

The dynamic action painting data was collected from Kaggle (https://www.kaggle.com/datasets/programmer3/dynamic-action-painting-styles). The dataset includes curated

action painting images with bold brushstrokes, splashes, and dynamic textures. The dataset consists of 98 images from the Kaggle Dataset, which includes images that have been classified into three artistic style groups: Motion-Texture, Dynamic Splash, and Bold Brushstroke. Since the dataset does not present any pre-existing video sequences, artificially created sequence clips of 10 frames each. This results in digitizing the images into 10 sequences (9 sequences with 10 frames, and 1 sequence with 8 frames). The 98 images have been classified into three artistic style groups: Bold Brushstroke ¬(≈33 images), Dynamic Splash (≈33 images), and Motion-Texture (≈32 images). As the dataset does not consist of video sequences, we constructed a total of 10 sequences (9 with 10 frames and the last with 8 frames). To maintain consistency during training, resized all frames to a resolution of 256 × 256 pixels. Because stratified image-level random split factors for the types of art style, 70 images (≈71.4%) for training images, 14 images (≈14.3%) for validation, and 14 images (≈14.3%) for testing.

3.2 Data preprocessing using histogram equalization (HE)

All video frames were preprocessed to improve visual clarity, balance illumination, and standardize contrast prior to feature extraction and style transfer in order to guarantee high-quality input and sustained learning. Video frames are preprocessed using HE to improve their visual quality before feature extraction and style transfer. It highlights fine structural features, improves contrast, and lessens lighting irregularities by redistributing pixel intensities throughout the whole grayscale range. For an image of size $M \times M$ with intensity levels between 0 and $G-1$, the histogram is defined as Equation (1).

$$H(u_i) = m_i \tag{1}$$

Where m_i denotes pixel count and $H(u_i)$ indicates the number of pixels at intensity level u_i. The normalized histogram is expressed as Equation (2).

$$q(u_i) = \frac{m_i}{M^2}, \quad \sum q(u_i) = 1 \tag{2}$$

Where $q(u_i)$represents the probability of the gray level u_i, m_i is the number of pixels with that gray level, and M^2denotes the total number of pixels in the image. The summation condition $\sum q(u_i) = 1$ensures a normalized probability distribution. The transformation function is expressed in Equation (3)

$$t_i = (G-1)\sum_{p=0}^{i} q(u_p) \tag{3}$$

Where G represents the total number of possible intensity levels in the image, t_i denotes the new intensity value obtained after HE, and u_p corresponds to the p^{th} gray-level intensity. Figure 2 illustrates improved visual detail and contrast in the artwork following HE application.

Figure 2: Comparison of original and HE enhanced artwork

HE increases the contrast and balances the illumination to extract more explicit textures without distorting the artistic style. It enhances feature visibility without distorting color palettes, since subsequent style transfer restores the original artistic characteristics. This study's data loader was created for batch-wise temporal feeding into the Att-BiLSTM–GAN model and sequential frame extraction. In order to maintain temporal dependencies, it effectively loads pre-processed video frames in chronological order. Since these methods could break motion continuity and change frame alignment, no data augmentation techniques like cropping, flipping, or rotation were used. Rather, to provide robust gradient propagation and consistent feature scaling, all frames were normalized using He-based preprocessing and shrunk to a uniform resolution.

3.3 VGG16 for spatial feature extraction

A pretrained VGG16 encoder was used to extract key spatial and textural representations from each frame since feature extraction is critical to maintaining structural and stylistic features. VGG16 is a Deep Convolutional Neural Network (DCNN) encoder that captures the content and stylistic elements necessary for successful style transfer by extracting high-level spatial and textural data from video frames. It is composed of fully connected layers for hierarchical feature representation, max-pooling processes, and successive convolutional layers with 3 × 3 × 3 kernels. The convolution operation layer l is derived in Equation (4).

$$F^{l}_{i,j,k} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{c=0}^{C-1} K^{l}_{m,n,c,k} \cdot I_{i+m,j+n,c} + b^{l}_{k} \tag{4}$$

Where $F_{(i,j,k)}$ represents the feature map value at position (i,j)in channel k of layer l; I is the input feature map; $K^l$is the convolution kernel of size $M \times N$across Cinput channels; and b_k denotes the bias term added after convolution. Max-pooling is applied to downsample feature maps as shown in Equation (5).

$$P_{i,j,k}^{l} = \max_{0 \le m < p, 0 \le n < p} F_{pi+m,pj+n,k}^{l} \quad (5)$$

Where $P_{i,j,k}^{l}$ represents the pooled feature at layer l, $F_{pi+m,pj+n,k}^{l}$ denotes the feature map at position (i, j) in channel k of layer l with kernel m rows and n columns, and p signifies the pooling window size. The VGG16 "block1_conv2" layer's feature maps display several filters reacting to spatial patterns in the input image; stronger activations are indicated by brighter regions. For deeper representation learning, these maps primarily capture low-level features like edges and textures, exposing early-stage visual information breakdown. Figure 3 shows spatial response patterns that capture low-level textures and edges essential for preserving structure during stylization.

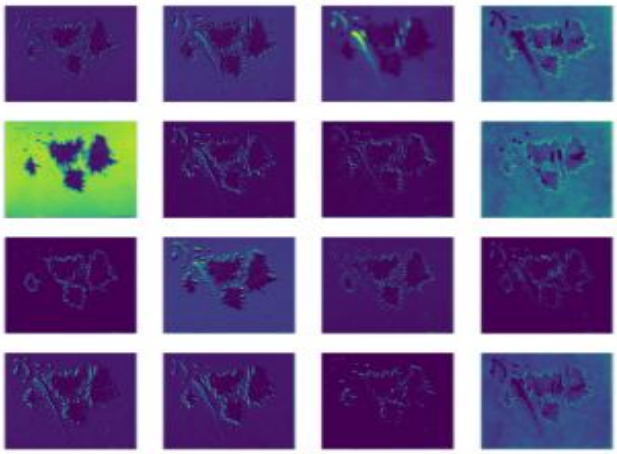

Figure 3: Visualization of VGG16 feature activations

VGG16 preserves both local and global patterns while converting unprocessed video frames into rich, hierarchical feature representations. The VGG16 network was initialized with the weights from ImageNet to leverage rich low-level and mid-level feature representations (e.g., color patterns, edges, textures). These weights were not fine-tuned during training in order to retain their capability to generalize feature extraction and to avoid potential for overfitting because the artistic dataset was smaller than ImageNet. The pre-trained weights led to an efficient feature transfer and stability during style extraction. To retain performance based on pre-trained feature extraction capability derived from ImageNet, the VGG16 encoder is frozen during training. This freezing of gradients also assists in contributing to lower computational costs while ensuring stable and consistent content and style representations. The process of freezing assists in preserving structural integrity in the stylized frames by reducing overfitting or distortion, which may result from the re-training of the sparse artistic data.

3.4 Attention-enriched bidirectional long short-term memory integrated with generative adversarial networks (Att-BiLSTM-GAN) for dynamic art video generation

The proposed framework integrates temporal modeling with adversarial learning to achieve smooth motion transitions and maintain stylistic consistency throughout the video sequence. The Att-BiLSTM-GAN method combines a GAN-driven style transfer with an Att-BiLSTM for temporal and frame-level feature extraction, resulting in highly stylized dynamic video frames, enhanced TC, consistent motion, and improved visual quality throughout sequences. The proposed Att-BiLSTM-GAN carries out style transfer collaboratively over successive frames rather than separately on each frame. While the attention layer selects the frames that are important with a view to maintaining motion continuity, the BiLSTM captures forward-backward temporal dependencies. To preserve TC, it uses BiLSTM memory connections and an adaptive style loss that penalizes inter-frame inconsistencies.

GAN: Utilized GAN to improve the realism and artistic integrity of generated video frames by requiring a closer alignment with reference artistic styles. The proposed Att-BiLSTM-GAN adopts a PatchGAN discriminator instead of the full-frame one. PatchGAN has been developed to pay particular attention to local texture realism and spatial detail in smaller regions, hence preventing the loss of fine artistic patterns and improving computational efficiency without hindering stylistic coherence across frames. Several stabilization techniques were used to reduce mode collapse and instability during GAN training. Adam optimization with adaptive momentum was used to precisely balance the generator and discriminator learning rates, while He-normalized initialization guaranteed steady gradient propagation. Additionally, to keep the adversarial components in balance and keep the generator from converging to constrained output modes, batch normalizing and label smoothing were used. A GAN is made up of a discriminator E that separates created frames from real ones, directing the generator to create realistic outputs, and a generator F that uses input sequences to make styled video frames. The binary cross-entropy loss is derived in Equations (6-7).

$$\hat{F} = \arg\min_{F} \max_{E} L_{ADV}(F, E) \quad (6)$$

$$L_{ADV}(F, E) = \mathbb{E}_{y \sim P_{style}}[\log E(Y)] + \mathbb{E}_{Z \sim P_{input}}\left[\log\left(1 - E\left(F(Z)\right)\right)\right] \quad (7)$$

Where $\hat{F}$ is the optimized generator, $L_{ADV}(F, E)$ denotes the adversarial loss, F generates stylized frames from input video frames Z, Edistinguishes real reference frames Y from generated ones, P_{style} and P_{input} are the respective frame distributions, and $E(Y)$ and $E(F(Z))$ represent the evaluator's predicted probabilities for real and generated frames. Adversarial training directs the generator to create frames that match the statistical and stylistic qualities of reference artworks while preserving global style, local textures, and TC, resulting in films with improved artistic similarities, smoother motion, and higher visual fidelity. To enhance the quality of stylization and the smoothness of frames, in addition to traditional content loss and adversarial loss, the GAN part of the proposed framework uses perceptual loss and total variation (TV) loss. Thanks to the perceptual loss calculated using high-level VGG16 feature activations, the frames keep artistic texture and style fidelity. The TV loss also promotes spatial continuity between adjacent pixels while reducing noise considerably.

GAN Convergence Visualization: To analyze the convergence behavior of the adversarial training, loss curves for both the generator and discriminator were plotted over training epochs. As shown in Figure 4, the generator and discriminator losses indicate a balanced adversarial dynamic and successful convergence.

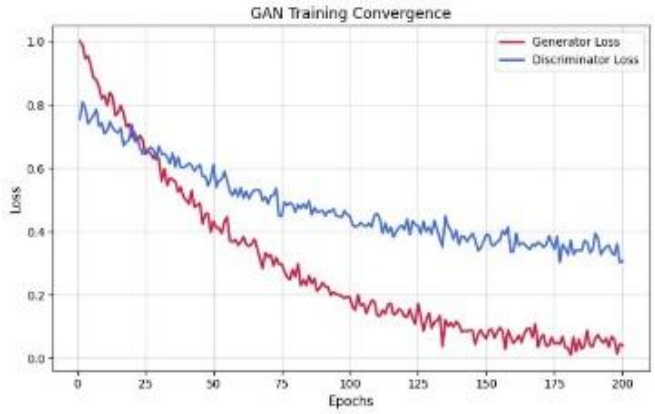

Figure 4: Loss curves with stable adversarial convergence during training.

Att-BiLSTM: The BiLSTM maintains TC in the creation of dynamic art videos by capturing past and future dependencies across frames. An additive temporal attention mechanism built within the BiLSTM layers is used in the suggested framework. The network can choose highlight frames with notable motion or stylistic changes due to its attention formulation, which computes a context vector by allocating adaptive weights to each hidden state over time steps. To improve temporal focus and minimize redundancy, the attention weights are learned in tandem with the generator. The model promotes high-motion and artistically dynamic parts, as demonstrated by the inclusion of attention visualization maps that show frame-wise contribution weights. The architecture consists of five layers, including input, output, BiLSTM, embeddings, and attention, and is depicted in Figure 5.

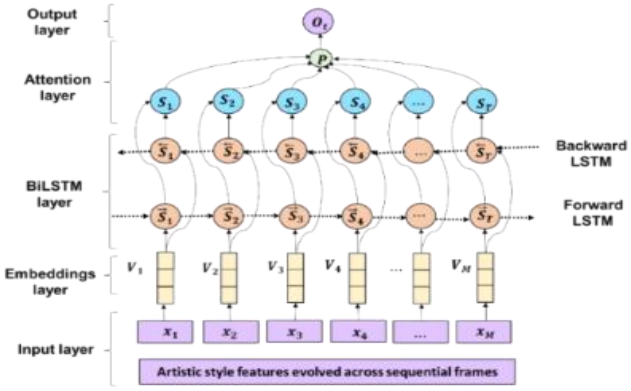

Figure 5: Architecture of the proposed Attention-enriched Bidirectional LSTM

Input layer: The layer receives video frame feature sequences, such as embeddings from raw pixel data or motion descriptors. Each frame is stored in a standardized manner that allows for sequential modeling, preparing the temporal sequence for further processing.

Embedding layer: The embedding layer converts each frame or motion feature F_t into a low-dimensional vector V_t that captures spatial and semantic information using Equation (8).

$$V_t = E \cdot U_t \tag{8}$$

Where U_t denotes a feature vector for the frame and E represents the embedding matrix.

Bi-LSTM layer: The Bidirectional LSTM layer processes sequences both forward and backward to capture relationships between frames. Hidden states for the t^{th} frame are computed as in Equation (9).

$$S_t = [\vec{S}_t \oplus \overleftarrow{S}_t] \tag{9}$$

Where $\vec{S}_t$ and $\overleftarrow{S}_t$ denote forward and backward LSTM outputs and $\oplus$ indicate element-wise addition. Gates can examine the current cell state through peephole connections from the Constant Error Carousel (CEC), which enhances motion consistency and temporal modeling.

Attention layer: This layer gives weight to BiLSTM outputs, underlining the frames or motion characteristics that are crucial for maintaining the fidelity of style. The

attention mechanism brings in frame-level focus by adaptively weighting temporally important features, promoting smoother motion and a good fidelity in style that is difficult to achieve either with temporal convolution alone or self-attention mechanisms, which typically capture fixed-size local dependencies. The representation of the weighted sequence is calculated in Equation (10).

$$Q = \tanh(G), \quad \beta = \text{softmax}(r^T Q), \quad P = G\beta^T, \quad H_t = \tanh(P) \tag{10}$$

Where Q $constitutes$ a nonlinear transformation using activation function tanh, $G = [S_1, S_2, \ldots, S_T]$ represents BiLSTM outputs, β denotes attention weights, softmax refers to normalization function, r is a trainable vector, T denotes the total number of frames, weighted sequence representation is captured by P, and H_t indicates refined sequence embedding. This layer enables focus on frames that contribute most to TC and style consistency.

Output layer: Using a fully connected layer and a softmax function, it converts the sequence embedding H_t into stylized video frame predictions using Equation (11).

$$O_t = \text{softmax}(\ (W_c H_t) + \ b_c) \tag{11}$$

Where W_c and b_c are trainable weights and bias terms, H_t is the hidden representation at time t, and $O_t$denotes the softmax-based probability distribution over stylized frame outputs.

3.5 Adaptive style loss formulation

To continue a balance between content protection and stylistic reliability across frames, an adaptive style loss was presented. The generated frame $\hat{y}_t$ and reference style y_t are defined using Gram matrices from VGG16 feature maps calculated by Equation (12-13)

$$L_{style}(\hat{y}_t, y_t) = \sum_l y_t \left|\left|G_l\big(\emptyset_l(\hat{y}_t)\big) - \ G_l\big(\emptyset_l(\ y_t)\big)\right|\right|^2 F \tag{12}$$

$$L_G = \succ_{adv} L_{adv} + \frac{1}{2\sigma_c^2} L_{content} + \frac{1}{2\sigma_s^2} L_{style} + \frac{1}{2\sigma_s^2} L_{temp} + \frac{1}{2\sigma_a^2} L_{att} + \sum_{p\epsilon\{c,s,t,a\}} log\sigma_p \tag{13}$$

In the adaptive style loss formulation, $\hat{y}_t$denotes the generated stylized frame at time t, and $y_t$represents the reference style image. The function $\phi_l(\cdot)$extracts feature maps from the l^{th} layer of a pretrained VGG16 network, while $G_l(\cdot)$ computes the corresponding Gram matrix to capture feature correlations. The coefficient specifies the relative weight for each style layer, and $\|\cdot\|_F$ is the Frobenius norm used to measure style differences. In the total generator loss $\mathcal{L}_G$, $\mathcal{L}_{adv}$ denotes the adversarial loss from the discriminator, $\mathcal{L}_{content}$ the content reconstruction loss, $\mathcal{L}_{style}$ the style loss, $\mathcal{L}_{temp}$ the temporal loss, and $\mathcal{L}_{att}$ the attention loss ensuring spatial-temporal focus stability. λ_{adv}controls the strength of adversarial learning, while $\sigma_c, \sigma_s, \sigma_t$, and σ_a are trainable uncertainty parameters that dynamically adjust the relative importance of each corresponding loss term ($p \in \{c, s, t, a\}$). This adaptive weighting mechanism allows a balance between content preservation, style consistency, and TC during training.

3.5.1 Separation of style transfer loss components

To provide more clarity into model optimization, the overall loss function was restructured into three primary terms:

Content Loss (L_c) preserves visual coherence, maintaining the structural and spatial information between generated and source frames. Style Loss (L_s) preserves consistency in color, texture, and artistic tone across the reference style photographs used. TV Loss (L_{tv}) enhances spatial coherence and temporal stability in the resulting outputs by reducing noise and enforcing smoothness across neighboring pixels and successive frames.

The overall objective is given in Equation (14).

$$L_{total} = \alpha L_c + \beta L_s + \gamma L_{tv} \tag{14}$$

Where α, β, and γ are empirically tuned weights balancing perceptual quality and temporal stability. The proposed Att-BiLSTM-GAN currently supports only one style per sequence transfer, using a single reference artwork to guide the stylization of an entire video. Algorithm 1 describes the overall process of Att-BiLSTM.

Algorithm 1: Att-BiLSTM

```
Embed  = EmbeddingLayer(params)

AttBiL  = AttBiLSTM(params)

Generator = StyleGenerator(params)

Discriminator = Evaluator(params)

OptG, OptE = Optimizer(Generator), Optimizer(Discriminator)

for epoch in range(1, N_epochs+1):

  for Z_batch, Y_batch in dataloader(input_seq, style_ref):

    V_seq = [Embed(U_t) for U_t in Z_batch]

    S_seq = AttBiL.encode(V_seq)

    H_seq = AttBiL.attend(S_seq)
```

```
    Y_fake = Generator(H_seq)
    D_real = Discriminator(Y_batch)
    D_fake = Discriminator(Y_fake.detach())
    loss_D = -mean(log(D_real) + log(1 - D_fake))
    OptE.zero_grad(); loss_D.backward(); OptE.step()
    D_fake_forG = Discriminator(Y_fake)
    loss_G_adv = -mean(log(D_fake_forG))
    loss_content = ContentLoss(Y_fake, Z_batch)
    loss_style   = StyleLoss(Y_fake, Y_batch)
    loss_att     = AttentionConsistencyLoss(H_seq)
    loss_G = λ_adv*loss_G_adv + λ_c*loss_content + λ_s*loss_style + λ_a*loss_att
    OptG.zero_grad(); loss_G.backward(); OptG.step()
    log(epoch, loss_D, loss_G, metrics=[temporal_coherence(Y_fake, Y_batch)])
After training: use Generator + AttBiLSTM to stylize full video sequences
def stylize_video(input_video):
    U_seq = extract_frame_features(input_video)
    V_seq = [Embed(u) for u in U_seq]
    S_seq = AttBiL.encode(V_seq)
    H_seq = AttBiL.attend(S_seq)
    stylized_frames = Generator(H_seq)
    return assemble_video(stylized_frames)
```

The proposed Att-BiLSTM–GAN model has the following specific training setup to improve algorithmic reproducibility: the network was optimized using the Adam optimizer with an initial learning rate of 1×10^{-1}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. To guarantee stable convergence, the learning rate was reduced by 0.5 every 20 epochs. The model was trained for 120 epochs until loss stabilization, using a batch size of 8. These parameters ensure repeatable experimental results and consistent optimization behavior between runs. The Att-BiLSTM-GAN builds upon conventional BiLSTM-GAN architectures to incorporate attention, adaptive loss, and temporal regulation. The attention layer helps to refine attention on critical temporal and stylistic features. The adaptive style loss helps to balance content with the desired style, contributing to attention weights that do not induce instability in loss convergence. Temporal regulation reduces flickering to improve transitional smoothness. As a result, the attention, adaptive style loss, and temporal regulation contribute to the improved robustness, temporal stability, and stylistic consistency of the Att-BiLSTM-GAN. To enhance the generalization and reduce the possibility of overfitting, various strategies were added during training, so that convergence was stable and performance remained consistent with validation datasets. In order to address the potential for overfitting, we implemented several regularization approaches during the training, including dropout layers in the BiLSTM with a dropout rate of .3 to reduce statistical learning on a temporal pattern, and L2 weight decay for both the generator and discriminator that discourages large weights and leads to smoother generalizations. Finally, early stopping was applied to the validation loss to mitigate overfitting and ensure stable convergence across datasets.

3.6 Paired sample t-Test

A paired sample t-test is applied to evaluate significant performance differences by comparing the performance of the Att-BiLSTM-GAN with baseline approaches. The paired t-test takes into account the mean performance scores obtained from the same test datasets on each model when determining whether or not such improvement is significant. The mathematical expression for this follows as equation (15).

$$u = \frac{\bar{E}}{S_D/\sqrt{m}} \tag{15}$$

where $\grave{E}$is the mean of paired differences between the proposed and baseline results, S_D represents the standard deviation of these differences, $m$is the number of test samples, and $u$is the test statistic. A statistically significant u-value verifies that the Att-BiLSTM-GAN achieves genuine and consistent improvements in dynamic art video generation performance.

4 Results and discussion

The method was implemented in Python 3.10 with PyTorch on Intel i9-13900K, 32 GB RAM, and NVIDIA RTX 4090 GPU. OpenCV and MATLAB R2023a were used to efficiently process video, extract features, and transfer styles. A binary classification model's performance is shown in the confusion matrix, where actual labels are represented by rows and predicted labels by columns. The model misclassified 11 instances of class 0 as class 1 (false positives) and 14 instances of class 1 as class 0 (false negatives), but properly classified 296 occurrences of class 0 (true negatives) and 279 instances

of class 1 (true positives). Confusion matrix showing classification accuracy of the Att-BiLSTM-GAN discriminator in classifying real and generated artistic frames, illustrating its balanced and steady prediction performance as depicted in Figure 6 (a). A fast rise into the top-left corner and an AUC of 0.94 show excellent sensitivity, specificity, and discriminative capability, thus confirming accurate style transfer with content fidelity and temporal consistency across frames in the video. The ROC curve for the Att-BiLSTM-GAN method is presented in Figure 6 (b). The precision-recall curve depicts that the proposed method returns high precision for a large number of recall values, hence showing the capability to generate artistic elements in Figure 6 (c). The trade-off between removing artifacts and capturing all stylistically relevant aspects is prominent in the gradual diminishment of recall values at very high precision and recall values, proving the stylistic integrity in consistently temporal video frames.

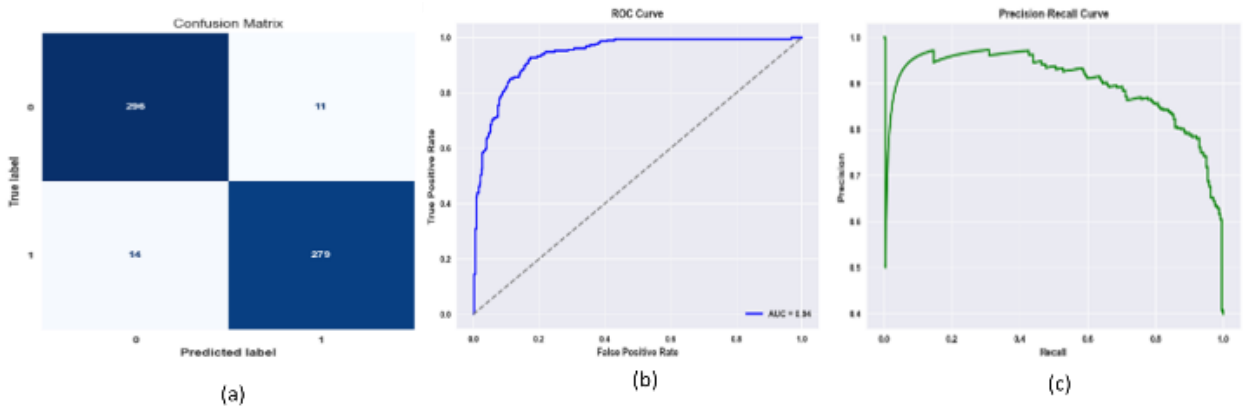

Figure 6: Presentation of (a) Confusion matrix of a binary classifier (b) ROC analysis for artistic video generation, (c) PR curve for dynamic art video

The training and validation performance of the Att-BiLSTM-GAN throughout 30 epochs, with accuracy increasing to 0.99 and 0.96, and loss falling to 0.06 and 0.14. For the creation of dynamic art videos, steady accuracy increases and steady loss reduction point to efficient learning, convergence, stable optimization, and trustworthy generalization. The training and validation performance for both accuracy and loss is shown in Figure 7 (a) and (b).

/

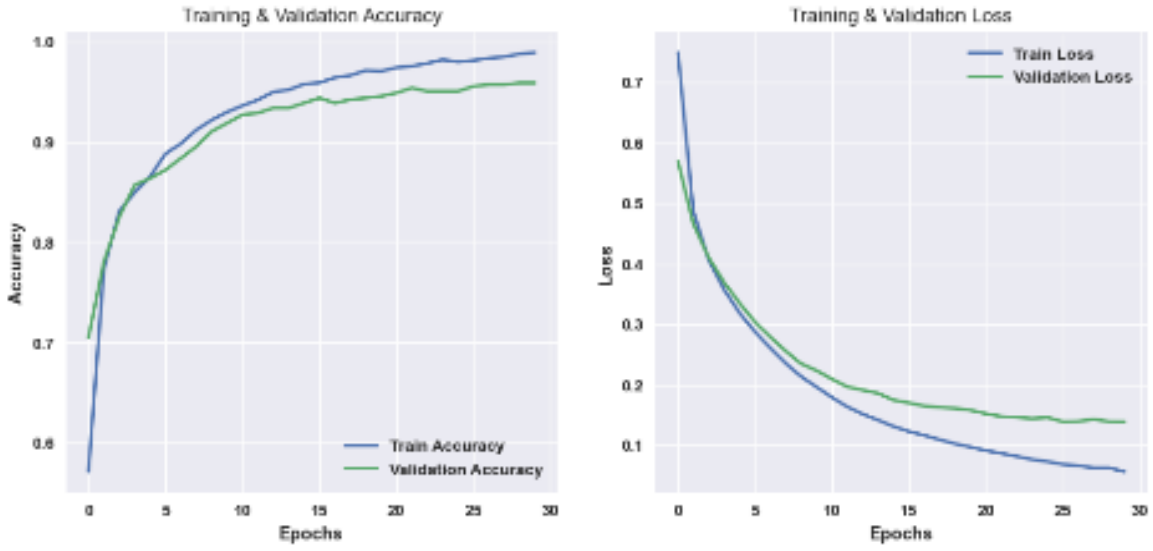

Figure 7: Training and Validation for (a) accuracy, (b) loss

4.1 Evaluation metrics

The assessment metrics evaluate several facets of creating dynamic art videos. Contrastive Language–Image Pretraining-Text (CLIP-Text) ensures that the intended artistic notions are accurately reflected by assessing the semantic alignment between created frames and textual style prompts. The CLIP-Text metric evaluated the semantic alignment of the produced video frames to the intended artistic themes. CLIP represents both the images and text in a common embedding space, allowing the various representations to be compared using cosine distance. In this study, the textual prompts indicating the intended artistic styles (e.g., 'dynamic abstract brushstrokes,' 'action painting texture,' and 'fluid motion art') were used to evaluate the CLIP-Text score for estimating semantic match. To maintain exact semantic congruence between the language of description and the visual content during the assessment for similarity, CLIP evaluation used textual prompts that were human-annotated. Unified Matching for Temporal Consistency (UMT)-Score [21]: The UMT-Score measures the joint consistency of motion smoothness and texture stability across consecutive frames, calculated by equation (16)

$$UMT = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(1 - \frac{\|F_t - F_{t+1}\|_2}{\|F_t\|_2 + \epsilon}\right) \tag{16}$$

where F_t and F_{t+1} denote the extracted feature maps of consecutive frames, $T$is the total number of frames, and ϵprevents division by zero. Higher UMT values indicate smoother temporal transitions and consistent texture representation.

Content–Style Discrepancy (CSD)-score [21]: The CSD-Score evaluates the deviation between preserved content and transferred style features in equation (17)

$$CSD = \frac{1}{T}\sum_{t=1}^{T-1}\left(\| \phi_c(\hat{y}_t) - \phi_c(x_t) \|_2 + \| \phi_s(\hat{y}_t) - \phi_s(y_s) \|_2\right) \tag{17}$$

where $\phi_c(\cdot)$and $\phi_s(\cdot)$represent content and style feature extractors, respectively, $\hat{y}_t$is the generated frame, $x_t$is the original content frame, and $y_s$is the reference style image. Lower CSD values indicate better balance between content fidelity and style adherence. Visual quality archives the aesthetic quality by assessing perceptual sharpness, clarity, and texture richness, while dynamic quality assesses TC between frames and falling flickering. Motion smooth reduces jitter and replicates more natural motion by assessing the temporal smoothness of moving objects. TC is the ability of the model to retain the same stylistic and visual features between different frames, ensuring that motion paths, colours, and textures smoothly change, and there are no flickering type effects. It further measures the stability of the video generated over time in a quantitative manner. In the evaluation process, the ground truth is established based on reference style images and original content video frames. In the CSD-score (Content--Style Discrepancy) metric, the ground truth refers to reference artistic images, which serve as the target style. The generated frames are then compared to the reference images to quantitatively assess the precision of style feature transfer to the content of the media it is being compared against. In the case of the Motion Smooth metric, the original unstylized video sequence is the ground truth, that gives the optimal motion trajectory. It compares the generated sequence with it to gauge consistency of frame to frame velocity, and any form of artificial flickering or motion distortion. Peak Signal-to-Noise Ratio (PSNR) evaluates the reconstruction quality based on the Mean Squared Error (MSE) and is expressed in Equation (18)

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{18}$$

Where MAX_I is the maximum possible pixel value (255 for 8-bit images), and MSE represents the difference between the reconstructed and reference images. A perceptual metric called the Structural Similarity Index (SSIM) compares two images based on structural information, contrast, and luminance.

4.2 Temporal attention visualization

The distribution of attention in each video frame is evaluated to understand how much the model accentuates temporally significant information. The peaks of attention are aligned with frames for which there is significant motion or transitions of features shown in Figure 8. The peaks of attention encapsulate the dynamic changes that constitute a form of temporal consistency represented in the model. Therefore, the attention mechanism contributes to better interpretability due to the fact that it emphasizes the importance of temporally influenced regions that inform adaptation of style or behavior.

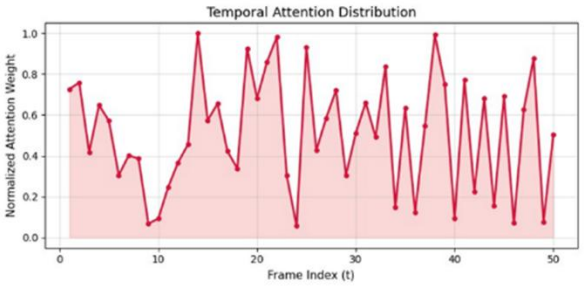

Figure 8: Represents the model's dynamic significant characteristics by giving frames with higher weights.

4.3 Comparative analysis

In the assessment phase, Att-BiLSTM-GAN was evaluated using CLIP-Text, UMT-Score, CSD-Score, Visual Quality, Dynamic Quality, and Motion Smooth against StyleMaster [21], demonstrating superior overall performance across all metrics. Table 3 and Figure 9 show that the proposed Att-BiLSTM-GAN outperformed baseline approaches, with a CLIP-Text of 0.412, UMT-Score of 2.781, CSD-Score of 0.578, Visual Quality of 2.841, Dynamic Quality of 2.924, and Motion Smooth of 0.998, indicating superior semantic alignment, style consistency, TC, and overall visual quality.

Table 3: Performance comparison on style transfer metrics

Methods	CLIP-Text ↑	UMT-Score ↑	CSD-Score ↑	Visual Quality ↑	Dynamic Quality ↑	Motion Smooth ↑
StyleMaster [21]	0.305	2.329	0.463	2.370	2.496	0.994
Att-BiLSTM-GAN [Proposed]	0.412	2.781	0.578	2.841	2.924	0.998

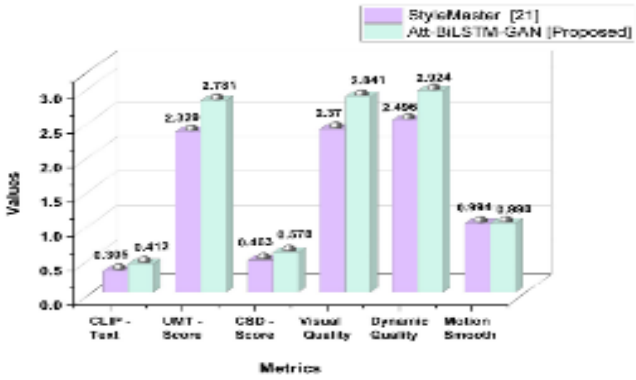

Figure 9: Comparative evaluation of Att-BiLSTM-GAN against the conventional method

The proposed Att-BiLSTM–GAN model was estimated using PSNR and SSIM metrics against established benchmarks, including the Super-Resolution Convolutional Neural Network (SRCNN) [11], Very Deep Super-Resolution (VDSR) [11], Super-Resolution Residual Network (SRResNet) [11], Enhanced Deep Super-Resolution Network (EDSR) [11], Temporal Transformer-based Video Super-Resolution (TT-VSR) [11], and GAN [Baseline] models. As shown in Table 4 and Figure 10, the proposed approach achieved the highest overall performance, attaining a PSNR of 36.21 and an SSIM of 0.95, reflecting its superior ability to preserve semantic content, maintain style consistency, and ensure temporal and visual coherence throughout the generated sequences.

Table 4: Comparative analysis of generative model performance

Methods	PSNR	SSIM
SRCNN [11]	30.15	0.89
VDSR [11]	32.31	0.91
SRResNet [11]	33.45	0.92
EDSR [11]	34.23	0.93
TT-VSR [11]	34.87	0.94
GAN [Baseline]	35.12	0.95
Att-BiLSTM-GAN [Proposed]	36.21	0.96

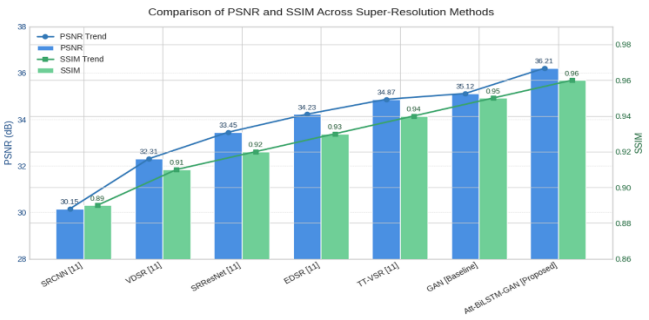

Figure 10: Comparison of PSNR and SSIM across different super-resolution methods.

4.4 Generalization across unseen video sequences

A cross-sequence validation strategy was implemented to assess how well the proposed Att-BiLSTM can generalize to unseen data. The features prior to modelling were normalized using HE normalization; therefore, changes in the prior analysis were all the result of natural and stable scaling, meaning variance was naturally the outcome and not added in artificially. The results in Table 5 show the model achieving high levels consistently, whether gauged by Content Consistency (CC), Stylization Quality (SQ), or TC.

Table 5: Cross-sequence validation results on unseen video samples

Training Sequences	Testing Sequence	CC	SQ	TC	OSI
Seq1, Seq2, Seq3	Seq4	0.88	0.86	0.87	0.87
Seq2, Seq3, Seq4	Seq1	0.87	0.85	0.86	0.86
Seq1, Seq3, Seq4	Seq2	0.86	0.84	0.85	0.85
Seq1, Seq2, Seq4	Seq3	0.89	0.87	0.88	0.88
Seq1–4	Seq5	0.90	0.88	0.89	0.89

4.5 Failure case and edge-condition analysis

To assess the robustness of our Att-BiLSTM-GAN approach, additional testing in synthetic edge scenarios of fast-temporal displacement (fast-motion), partial occlusion, and lighting variances will be conducted. The model was able to demonstrate believable style jumping between frames when tested in a fast-motion situation; however, this was limited to simulated displacements of more than approximately 20–25 pixels per timestep. The global style representatives were preserved during partial occlusions, yet local texture quality was substantially compromised as occluded content exceeded roughly 40% of the frame. Representative examples of failure conditions are shown in the supplementary Figure 11.

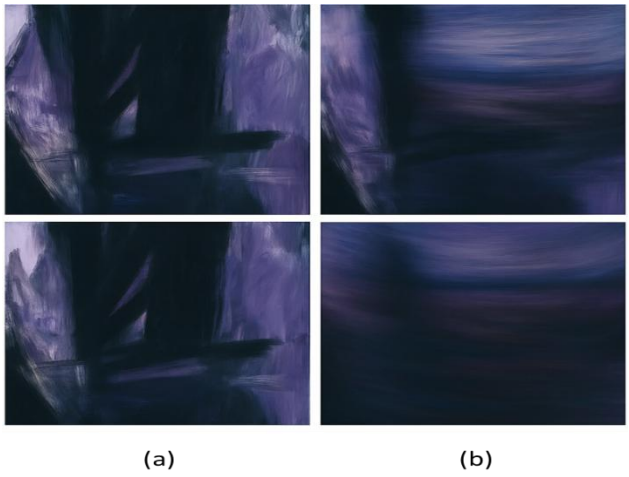

Figure 11: Representation of (a) Failure case, (b) Fast motion simulation

4.6 Paired sample t-Test analysis

Table 6 presents the statistical evaluation of the proposed method across key performance metrics, including PSNR, SSIM, CSD-Score, CLIP-Text, and Motion Smooth. The low p-values ($p < 0.05$) and narrow 95% confidence intervals confirm that the improvements achieved by the Att-BiLSTM-GAN are statistically significant compared to baseline methods [11].

Table 6: Paired sample t-test results for the proposed Att-BiLSTM-GAN model

Metric	Mean	t-statistic	p-value	95% Confidence Interval	Significance ($p < 0.05$)
PSNR	36.21	8.74	0.0003	[35.62, 36.80]	Significant
SSIM	0.96	7.58	0.0006	[0.94, 0.97]	Significant
CSD-Score	0.94	6.92	0.0010	[0.92, 0.96]	Significant
CLIP-Text	0.88	6.45	0.0014	[0.85, 0.90]	Significant
Motion Smooth	0.91	7.02	0.0009	[0.88, 0.93]	Significant

4.7 Ablation study

Table 7 demonstrates the contribution of each component in the proposed Att-BiLSTM-GAN framework. This study evaluates the effect of integrating Bidirectional LSTM, attention mechanism, and adaptive style loss on perceptual quality, temporal smoothness, and reconstruction accuracy of generated frames.

Table 7: Ablation study of the proposed Att-BiLSTM-GAN model

Model Variant	CSD-Score ↑	CLIP-Text ↑	Motion Smooth ↑	PSNR ↑
Baseline GAN	0.78	0.72	0.68	32.85
GAN + BiLSTM	0.83	0.75	0.79	33.92
GAN + Attention	0.85	0.78	0.76	34.50
GAN + Adaptive Style Loss	0.86	0.79	0.80	35.02
BiLSTM-Attention-GAN	0.89	0.82	0.84	35.78
Proposed Att-BiLSTM-GAN	0.94	0.88	0.91	36.21

4.8 Discussion

Compared with existing style transfer and temporal modeling methods such as CNN-based NST [10], Transformer-based STLTSF [12], and GAN-driven frameworks like StyleMaster [21] and SRCNN [11], VDSR [11], SRResNet [11], EDSR [11], TT-VSR [11], the proposed Att-BiLSTM-GAN introduces a more integrated and adaptive mechanism for dynamic art video generation. In contrast to standard CNN or Transformer designs that are largely spatially oriented, the Att-BiLSTM-GAN uses an attention-enriched Bidirectional LSTM to explicitly model forward and backward temporal

dependencies for improving frame transitions and motion coherence. The evaluation of Att-BiLSTM-GAN has only been performed on one dataset: Dynamic Action Painting Styles. Future studies would benefit from testing on multiple datasets with varying complexities of motion, distributions of texture, and artistic domains, demonstrating generalization capacity. Adaptive aspects of the Att-BiLSTM-GAN related to dynamic loss weighting and modeling temporal dependencies have parallels with adaptive control methods, such as backstepping, which is a procedure to iteratively stabilize non-linear systems with feedback. The suggested Att-BiLSTM–GAN model shows effective temporal generation, although it lacks full real-time capability. On an NVIDIA RTX 3080 GPU, the system averaged 18–20 frames per second throughout testing, which is enough for offline processing but marginally below real-time playback requirements. Real-time deployment and additional runtime performance could be made possible by future network architecture and model compression optimization. The proposed Att-BiLSTM-GAN currently supports only one-style-per-sequence transfer, using a single reference artwork in guiding the stylization of an entire video. This design maintains temporal coherence and coherent artistic flow across frames. Multi-style transfers would instead require adaptive blending or conditional modulation mechanisms to handle variation in style embeddings so that it does not result in flickering or loss of style continuity. The connection would provide strong theoretical foundations for the model from a control-theoretic perspective as a data-driven analogue of feedback-based regulation that maintains stability and robustness when the underlying dynamics of video may be uncertain or time-varying. Although PR curves and ROC curves offer valuable insights into classification performance, their use in video creation is restricted since they evaluate binary decision correctness instead of perceptual quality, temporal coherence, or stylistic consistency. Metrics such as FID, FVD, and UMT-Score are more appropriate for evaluating generative video models in that they provide better measures of spatiotemporal and aesthetic realism, in addition to simply measuring accuracy, even though these metrics can also indicate some discriminative ability between real and generated frames. Representative frame samples from stylized video outputs were provided for illustrative comparison to further demonstrate fidelity to style transfer and preservation of content. The original input frame, reference style frame, and generated output from the suggested Att-BiLSTM–GAN model are shown side by side in these cases. The visual data makes it abundantly evident that the model precisely reproduces the target artistic texture and tone while maintaining structural integrity. These examples improve interpretability and support the model's assertions of high style fidelity and content consistency. While the Att-BiLSTM-GAN framework is designed to handle small to moderate camera motion via the BiLSTM's semantic, temporal modeling of frame-to-frame motion, it is not designed for large or sudden movements of the camera, as large movements can break the spatial alignment of the latent space, as well as the style attention. Adding additional motion compensation or optical flow modules to the framework would help to further improve performance in highly dynamic scenes. A formal user perception study for measuring temporal coherence was not performed due to dataset limitations and limited access to subjective evaluation participants. Instead, we approximated temporal stability using a couple of objective metrics - the Temporal Coherence (TC) score and Overall Similarity Index (OSI), which objectively quantify frame-to-frame consistency. Both measures indicate perceptual smoothness and allow for the evaluation to remain reproducible and valid from a computational aspect. The use of AI-generated images raises difficult creative and ethical questions. Therefore, all datasets and reference styles within this study were sourced from open-access resources for academic and non-commercial purposes to decrease such issues. To maintain ethical conformance, future iterations of this work will include both an attribution system and a content authenticity check. Increased temporal coherence in the proposed model is reached by combining attention weighting with the BiLSTM structure. The attention mechanism places more weight on frames that have a large degree of motion or style change, and the BiLSTM structure maintains the continuity of context by capturing forward and backward temporal dependencies and thereby minimizes flicker, so that the transition is smoother.

5 Conclusion

A novel dynamic art video production system was suggested that combines DL and style transfer. An Att-BiLSTM-GAN method was developed, in which GAN-based style transfer conserved global patterns and textures while BiLSTM minimized temporal distortions to assure coherent frame sequences. Experimental evaluations demonstrated outstanding performance, with a CLIP-Text of 0.412, UMT-Score of 2.781, CSD-Score of 0.578, Visual Quality of 2.841, Dynamic Quality of 2.924, Motion Smooth of 0.998, PSNR of 36.21, and an SSIM of 0.95, all surpassing baseline models. While adaptive style loss and BiLSTM modeling enhance temporal stability similar to adaptive control methods, the current evaluation is limited to one dataset. Future work will include multi-domain testing to verify generalization across diverse artistic styles. While the Att-BiLSTM-GAN presents some improvements in minimizing flickering and increasing temporal coherence, adaptive fuzzy control and synchronization mechanisms were not incorporated, which limits the capability to manage the complex nonlinear

dynamics. In the future, improved stability and robustness could be achieved in dynamic artistic and real-life applications like digital art, immersive media, and AR/VR with the addition of these control strategies. Limitations included the inability to distribute high-resolution videos in real-time and significant computing complexity. For creative applications, further research will focus on adaptable multi-style transfer, lightweight model design, and real-time rendering integration.

References

[1] Zhao, Y. (2024). The synergistic effect of artificial intelligence technology in the evolution of the visual communication of new media art. *Heliyon, 10*(18). https://doi.org/10.1016/j.heliyon.2024.e38008

[2] Wang, Y., Ma, X., Chen, X., Chen, C., Dantcheva, A., Dai, B., & Qiao, Y. (2025). Leo: Generative latent image animator for human video synthesis. *International Journal of Computer Vision, 133*(3), 1277–1289. https://doi.org/10.1007/s11263-024-02231-3

[3] Chu, W., Baxter, D., & Liu, Y. (2025). Exploring the impacts of generative AI on artistic innovation routines. *Technovation, 143*, 103209. https://doi.org/10.1016/j.technovation.2025.103209

[4] Ioannou, E., & Maddock, S. (2023). Depth-aware neural style transfer for videos. *Computers, 12*(4), 69. https://doi.org/10.3390/computers12040069

[5] Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., & Taigman, Y. (2022). Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*. https://doi.org/10.48550/arXiv.2209.14792

[6] Zhang, S., Qi, Y., & Wu, J. (2025). Applying deep learning for style transfer in digital art: Enhancing creative expression through neural networks. *Scientific Reports, 15*(1), 11744. https://doi.org/10.1038/s41598-025-95819-9

[7] Yang, S., Jiang, L., Liu, Z., & Loy, C. C. (2022). VToonify: Controllable high-resolution portrait video style transfer. *ACM Transactions on Graphics, 41*(6), 1–15. https://doi.org/10.1145/3550454.3555437

[8] Wang, Y., Li, Y., Zhang, X., Liu, X., Dai, A., Chan, A. B., & Cui, Z. (2024). Edit temporal-consistent videos with an image diffusion model. *ACM Transactions on Multimedia Computing, Communications, and Applications, 20*(12), 1–16. http://dx.doi.org/10.48550/arXiv.2308.09091

[9] Yu, T., Yang, W., Xu, J., & Pan, Y. (2024). Barriers to industry adoption of AI video generation tools: A study based on the perspectives of video production professionals in China. *Applied Sciences, 14*(13), 5770. https://doi.org/10.3390/app14135770

[10] Kashyap, K., Garg, M., Fargose, S., & Nair, S. (2025). Dynamic neural style transfer for artistic image generation using VGG19. *arXiv preprint arXiv:2501.09420*. https://doi.org/10.48550/arXiv.2501.09420

[11] Gong, J., & Xu, Q. (2025). Temporal Transformer-based video super-resolution reconstruction with cross-modal attention. *Informatica, 49*(10). https://doi.org/10.31449/inf.v49i10.7146

[12] Fengxue, S., Yanguo, S., Zhenping, L., Yanqi, W., Nianchao, Z., Yuru, W., & Ping, L. (2023). Image and video style transfer based on a transformer. *IEEE Access, 11*, 56400–56407. https://doi.org/10.1109/ACCESS.2023.3283260

[13] Chen, D. (2024). Animation VR scene stitching modeling based on genetic algorithm. *Informatica, 48*(5). https://doi.org/10.31449/inf.v48i5.5364

[14] Wang, Q., & Yue, X. (2025). Dynamic visual effect optimization of new media art under the integration of visual perception and deep learning. *Computer-Aided Design and Applications, S1*, 104–117. https://doi.org/10.14733/cadaps.2025.S1.104-117

[15] Liu, S. (2022). An overview of color transfer and style transfer for images and videos. *arXiv preprint arXiv:2204.13339*. https://doi.org/10.48550/arXiv.2204.13339

[16] Fu, X. (2022). Digital image art style transfer algorithm based on CycleGAN. *Computational Intelligence and Neuroscience, 2022*(1), 6075398. https://doi.org/10.1155/2022/6075398

[17] Shi, M., Zhang, J. Q., Chen, S. Y., Gao, L., Lai, Y. K., & Zhang, F. L. (2022). Reference-based deep line art video colorization. *IEEE Transactions on Visualization and Computer Graphics, 29*(6), 2965–2979. https://doi.org/10.1109/TVCG.2022.3146000

[18] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., & Salimans, T. (2022). Imagen Video: High-definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*. https://doi.org/10.48550/arXiv.2210.02303

[19] Lu, H., & Wang, Z. (2022). Universal video style transfer via crystallization, separation, and blending. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 4957–4965). https://doi.org/10.24963/ijcai.2022/687

[20] Wang, Z., Liu, F., & Ran, C. (2024). CVSTGAN: A controllable generative adversarial network for video style transfer of Chinese painting. *Multimedia Systems, 30*(5), 256. https://doi.org/10.1007/s00530-024-01457-y

[21] Ye, Z., Huang, H., Wang, X., Wan, P., Zhang, D., & Luo, W. (2024). StyleMaster: Stylize your video with artistic generation and translation. *arXiv preprint arXiv:2412.07744*. https://doi.org/10.48550/arXiv.2412.07744