NADM: Noise-Aware Diffusion Model for Landscape Painting Video Generation

Ding-Ming Liu, Shao-Wei Li, Ruo-Yan Zhou, Li-Li Liang, Yong-Guan Hong, Yuan-Ze Zeng, Xiang Chang, *Member, IEEE*, Li-Jiang Li, Tian-Shuo Xu, *Member, IEEE*, Fei Chao, *Senior Member, IEEE*, Changjing Shang, and Qiang Shen

***Abstract*—Landscape painting is a gem of cultural and artistic heritage that showcases the splendor of nature through the deep observations and imaginations of its painters. Limited by traditional techniques, these artworks were confined to static imagery in ancient times, leaving the dynamism of landscapes and the subtleties of artistic sentiment to the viewer's imagination. Recently, emerging text-to-video (T2V) diffusion methods have shown significant promise in video generation, providing hope for the creation of dynamic landscape paintings. However, current T2V methods focus on generating natural videos, emphasizing the capture of details and the authenticity of physical laws. In contrast, landscape painting videos emphasize the overall dynamic aesthetic. Besides, challenges, such as the lack of specific datasets, the intricacy of artistic styles, and the creation of extensive, high-quality videos pose difficulties for these models in generating landscape painting videos. In this article, we propose landscape painting videos-high definition (LPV-HD), a novel T2V dataset for landscape painting videos, and noise-aware diffusion model (NADM), a T2V model that utilizes Stable Diffusion. Specifically, we present a motion module featuring a dual attention mechanism to capture the dynamic transformations of landscape imageries, alongside a noise adapter to leverage unsupervised contrastive learning in the latent space to ensure the overall beauty of the landscape painting video. Following the generation of keyframes, we employ optical flow for frame interpolation to enhance video smoothness. Our method not only retains the essence of the landscape painting imageries but also achieves dynamic transitions, significantly advancing the field of artistic video generation. Source code and dataset are available at https://github.com/llzlh21/NADM.**

***Index Terms*—Dynamic artistic videos, landscape painting, text-to-video diffusion.**

Received 23 December 2024; revised 30 March 2025 and 25 May 2025; accepted 28 May 2025. Date of publication 24 June 2025; date of current version 24 July 2025. This work was supported by the Key Program of the National Natural Science Foundation of China Joint Fund under Grant U23A20383. This article was recommended by Associate Editor X. Li. *(Corresponding author: Fei Chao.)*

Ding-Ming Liu, Shao-Wei Li, Li-Li Liang, Yong-Guan Hong, Yuan-Ze Zeng, and Li-Jiang Li are with the Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen 361005, China.

Ruo-Yan Zhou is with the Department of Computer Science, University of Hong Kong, Hong Kong.

Xiang Chang, Changjing Shang, and Qiang Shen are with the Institute of Mathematics, Physics, and Computer Science, Aberystwyth University, SY23 3-DB Aberystwyth, U.K.

Tian-Shuo Xu is with the AI Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511458, China.

Fei Chao is with the Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen 361005, China, and also with the Institute of Mathematics, Physics, and Computer Science, Aberystwyth University, SY23 3-DB Aberystwyth, U.K. (e-mail: fchao@xmu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2025.3576752.

Digital Object Identifier 10.1109/TCYB.2025.3576752

I. Introduction

LANDSCAPE painting is a vital component of cultural legacy. It stands out for blending philosophical depth with aesthetic appeal, employing distinctive light and shadow techniques, and mastering sophisticated ink methods [1], [2]. However, as modern technology evolves and people's aesthetic preferences become more diverse and modernized, the inheritance of this traditional art form faces challenges. Recently, advancements [3], [4], [5], [6], [7], [8], [9] in deep learning technology offer new possibilities for the modern inheritance of this ancient art form through innovative image generation of landscape paintings.

However, the beauty of landscape painting lies not only in its static form but also in the dynamic imagination and "artistic conception" it evokes. Unlike previous methods that focus on static landscape painting image generation, our approach operates at the frame level, explicitly modeling the dynamic evolution of landscape paintings to capture changes in light, shadow, and other temporal elements. By framing this as a Text-to-Video task, we aim to generate videos that showcase the continuous, time-based transitions of a landscape painting, preserving its depth and symbolic meaning in motion. However, Chinese landscape painting videos have their own uniqueness [10], [11]. They emphasize smooth transitions between frames and a consistent artistic atmosphere. Rather than focusing on the precise depiction of a specific object or action within the video, they seek to create a harmonious and cohesive visual experience.

While the Text-to-Video (T2V) approach [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23] has been widely studied across various fields, directly applying such methods to landscape painting videos still presents significant challenges. First, the absence of specific datasets for landscape painting styles hampers the accurate depiction of their unique style and artistic traits. Second, existing video generation methods struggle to capture the unique characteristics of landscape painting videos. Traditional approaches, which focus on photorealism, fail to preserve the "artistic conception" and "symbolism" of landscape paintings, leading to a loss of coherence between local details and global composition. Additionally, current methods, shown in Fig. 1, often fail to maintain fluid transitions between frames, impacting the video's overall smoothness and artistic continuity. Lastly, current techniques also struggle with representing dynamic elements like flowing rivers or drifting clouds, resulting in unnatural motion.

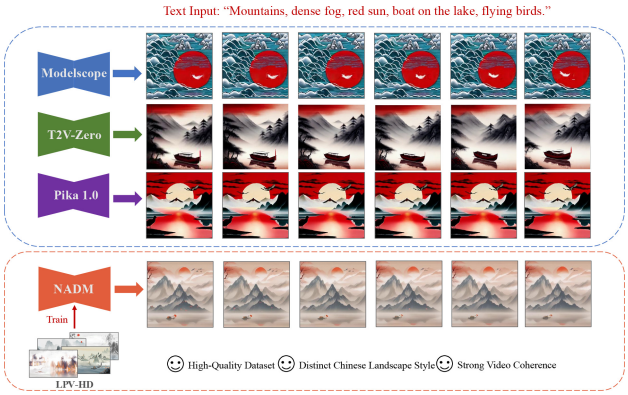

Fig. 1. Landscape painting video generation.

To tackle the specific challenges of generating landscape painting videos, notably the scarcity of datasets, we develop landscape painting videos-high definition (LPV-HD), a dedicated text-to-video dataset with around 1300 curated entries. This dataset aims to fill the technological void in emulating landscape painting's unique aesthetics, offering a vast resource to enhance deep learning models' ability to replicate this art form's distinct style. LPV-HD encompasses a range from traditional to modern styles and details the videos' dynamic elements, supporting our tech exploration and providing new tools for the art's modern evolution and innovation.

To address the challenges of smooth transitions, consistent artistic expression, and fluid motion in landscape painting videos, we propose the following solutions. Drawing inspiration from AnimateDiff [24], we integrate a bespoke motion module with frozen Stable Diffusion and combine Versatile Attention [24], [25] and Sparse-Causal Attention [26], [27], [28] to analyze temporal data. Versatile Attention captures global patterns, while Sparse-Causal Attention focuses on local causality for better event prediction. To ensure frame similarity across adjacent frames and dissimilarity across distant frames, we apply a noise contrastive learning approach using latent space, based on the predicted noise output from the de-noising U-Net. Additionally, we use optical flow-based frame interpolation for smoother transitions. Our noise-aware diffusion model (NADM) framework generates high-quality, diverse, and coherent videos that embody the style and fluidity of landscape painting.

Overall, the main contributions of this article are as follows.

1) Our work is the first to address the problem of text-to-landscape painting video generation using diffusion models, and construct a new high-quality dataset, LPV-HD, containing about 1300 landscape painting videos with detailed text descriptions.
2) We integrate contrastive learning based on video frame representations into the denoising process, significantly improving the quality and temporal coherence of generated landscape painting videos.
3) We propose a motion module that sequentially couples two types of attention in the denoising U-Net, enhancing both global-local consistency and artistic integrity.
4) Our approach only fine-tunes a small motion module on top of a frozen Stable Diffusion model, and introduces a training-free, optical flow-based frame interpolation strategy, greatly reducing computational cost.

The remainder of this article is organized as follows. Section II describes related studies of Text-to-Video Diffusion Models and landscape painting generation. Section III specified the proposed landscape painting video generation method. Section IV presents experimental results of the proposed method with current state-of-the-art methods. Finally, the conclusion and future work are presented in Section V.

II. Related Work

A. Text-to-Video Diffusion Models

Despite significant progress in Text-to-Image (T2I) generation technology, Text-to-Video (T2V) generation technology remains relatively lagging. This lag is mainly due to the lack of large-scale and high-quality text-video pairing datasets, as well as the high-dimensional modeling complexity of video data. Early T2V research [29], [30], [31], [32], [33], [34] primarily focused on generating simple video content.

Recently, several studies [22], [35], [36], [37] have explored leveraging the knowledge of pretrained T2I models to simplify the construction of T2V models, by performing the diffusion process in the latent space. In particular, these models have adopted a spatio-temporal separable architecture, inheriting the spatial operations of pretrained T2I models and reducing the complexity of constructing intricate models. Models, such as AnimateDiff [24] have enhanced the capability of generating dynamic videos by introducing innovative motion modeling modules. Champ [20] places greater emphasis on the expression of human details in generated videos. Several approaches [21], [23] have investigated T2V methods based on novel network architectures. Recent work on foundation models, such as SpectralGPT [38], offers insights for T2V models that aim to capture complex and dynamic elements. CCR-Net [39] uses a cross-channel reconstruction module for effective multimodal data fusion, offering efficient spatial-temporal feature blending for complex scene dynamics in video generation. These methods offer valuable perspectives for addressing core challenges in T2V generation, propelling further exploration and innovation in the field.

B. Landscape Painting Generation

GAN-based methods for generating images of landscape paintings have been extensively studied [4], [5], [6], [40], [41]. For example, Xue [4] uses GANs to generate landscape paintings in two stages: first creating outline sketches from random noise, then transforming these sketches into final paintings through edge-to-painting conversion. Polaca [40] is a poetry-based landscape generation model that uses GANs to transform poems into landscape images and matching calligraphy, ultimately merging them into a complete art piece. Recently, diffusion models have shown great promise in generating landscape painting images. DC-Net [42] offers a subpixel-level fusion framework that

bridges distribution gaps between data sources, providing valuable insights for integrating subtle spatial and temporal details in landscape painting video generation. HighDAN [43] suggests using domain adaptation to handle cross-scene data, offering insights into managing cross-scene consistency in landscape painting video generation. CCLAP [7] uses texts and reference images as conditions, employing latent diffusion models (LDMs) to generate landscape paintings with controllable content and style. These developments have prompted our exploration of the generation of landscape painting videos.

To the best of our knowledge, we are the first to apply diffusion models to generate dynamic landscape painting videos. Distinct from previous methods that focus on generating static images, our approach is designed to model the dynamic progression of scenes, encompassing changes in landscape, lighting, shadow, and other temporal attributes. This enables us to frame the problem as a text-to-video generation task, with the objective of synthesizing videos that reflect the continuous and temporally evolving nature of landscape paintings, rather than producing isolated static outputs.

III. Method

A. Preliminary

Stable Diffusion, based on the LDM [44], is a type of diffusion model that enhances image generation. The process starts by encoding an input image x_0 into a latent space $z_0 = E(x_0)$ using an encoder E. This latent representation z_0 is then modified through a series of steps according to

$$q(z_t|z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t I\right) \tag{1}$$

where z_t represents the latent state at step t, and $\mathcal{N}(z_t; \mu, \Sigma)$ represents a Gaussian distribution with mean μ and covariance Σ. The hyperparameter β_t controls the amount of noise added at each diffusion step t, with $\beta_t \in (0, 1)$. The process can be summarized as

$$z_t = \sqrt{\overline{\alpha_t}}z_0 + \sqrt{1-\overline{\alpha_t}}\epsilon, \epsilon \sim \mathcal{N}(0, I) \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$, with $\overline{\alpha_t} \in (0, 1)$.

Stable Diffusion adopts the vanilla training objective as proposed in DDPM [45], which is expressed as

$$\mathcal{L} = \mathbb{E}_{E(x_0),y,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|^2\right] \tag{3}$$

where y is the text description and $\tau_\theta(\cdot)$ is a text encoder based on the CLIP [46] ViT-L/14 model. The architecture features a UNet [47] with downscaling and upscaling blocks, incorporating 2-D convolutional, self-attention, and cross-attention layers.

DDIM [48] is proposed to accelerate the sampling process by optimizing the noise reduction steps. It transforms a random noise vector z_T into a coherent latent representation z_0 through a defined sequence of timesteps $t : T \to 1$, which is defined as

$$z_{t-1} = \sqrt{\alpha_{t-1}}\frac{z_t - \sqrt{1-\alpha_t}\epsilon_\theta(z_t, t, \tau_\theta(y))}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(z_t, t, \tau_\theta(y)) \tag{4}$$

where α_t is a parameter for noise scheduling, $\epsilon_\theta(z_t, t, \tau_\theta(y))$ is the predicted noise within the networks' latent space.

AnimateDiff [24] extends the Stable Diffusion model for T2V tasks by integrating a motion modeling module for video data handling. It progresses from processing four-dimensional image batches to handling five-dimensional video tensors (batch × channels × frames × height × width). The transformation adapts each 2-D convolution and attention layer from the base image model into pseudo-3-D layers that focus solely on spatial aspects. This adaptation involves reorganizing the frame dimension to merge with the batch dimension, which permits independent frame analysis. Subsequently, feature maps undergo a transformation to a (batch × height × width) × frames × channels configuration. This step sets the stage for the motion module, aimed at ensuring consistent motion and content stability across frames.

Motion modules are embedded throughout the U-shaped network, using vanilla temporal transformers with self-attention blocks operating along the temporal axis. The training objective is to minimize the following loss function

$$\mathcal{L} = \mathbb{E}_{E(x_0^{1:N}),y,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta\left(z_t^{1:N}, t, \tau_\theta(y)\right)\|_2^2\right] \tag{5}$$

where $x_0^{1:N}$ represents a sample of video data, and $z_t^{1:N}$ is the latent code obtained by adding noise to the initial latent code $z_0^{1:N}$ at time step t. During training, the pretrained weights of the base T2I model are frozen to maintain consistency in the feature space.

B. LPV-HD: Text-to-Landscape Painting Video Dataset

Diverse text-video datasets are essential for developing high-quality T2V generation models. However, there is a notable shortage of such datasets for landscape painting videos, limiting the fusion of traditional art with modern technology. The lack of comprehensive datasets with textual annotations hampers the creation of high-quality landscape painting videos.

Our work fills this notable gap by introducing LPV-HD, a groundbreaking text-to-landscape painting video dataset, now publicly available. It includes 1298 text-video pairs, sourced from open domains in high-definition. We collected 210 high-resolution clips from water ink animations and over 400 original videos from YouTube. To manage complex scene transitions, we meticulously segmented these into 1300 single-scene clips, enhancing training utility. Acknowledging the videos' intricate details, we manually annotate each text to accurately match the video content. Due to the diverse sources and editing processes, we present only the common characteristics of the dataset. Out of the 1298 videos, 1088 have a duration of 6 s; 1054 have a resolution of 1920 × 1080; and 1099 have a frame rate of 25 fps.

Our LPV-HD dataset effectively addresses the scarcity of specialized datasets for generating landscape painting videos from textual descriptions. This pioneering effort fosters the convergence of traditional art and artificial intelligence, catalyzing new avenues for scholarly inquiry and practical application.

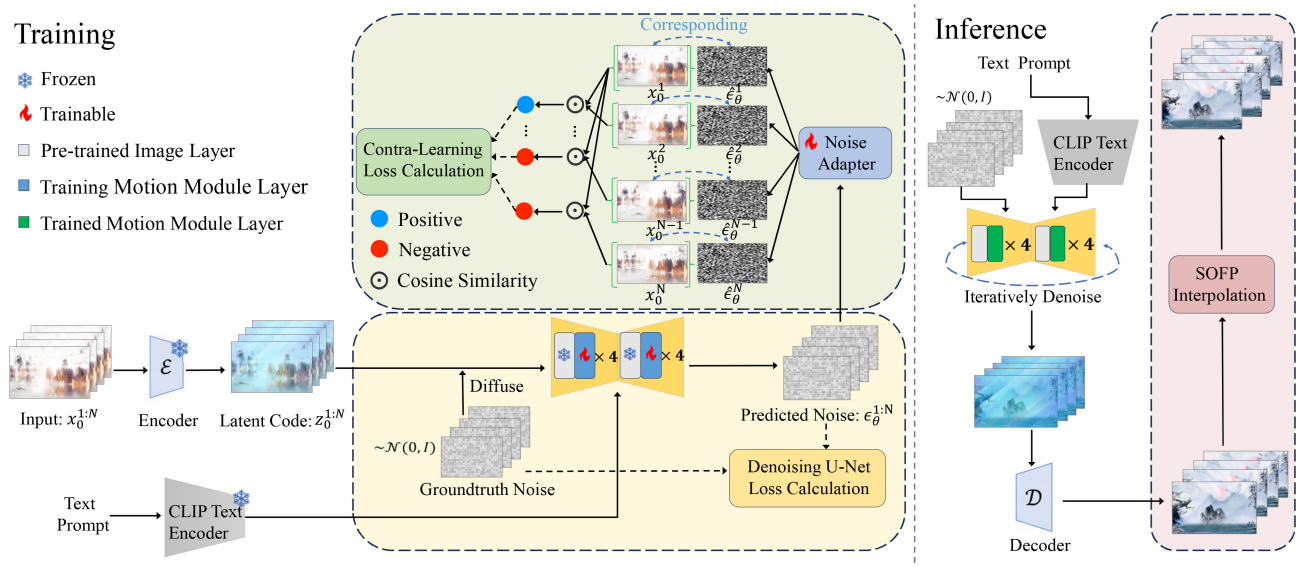

Fig. 2. Overview of NADM. Left: the architecture. NADM integrates a trainable motion module based on a frozen Stable Diffusion and introduces a noise adapter to accommodate contrast learning of noise in latent space. Right: the inference framework. The video is generated by the denoising U-Net integrated with the motion module and enhanced by sparse optical flow projection (SOFP) interpolation technology.

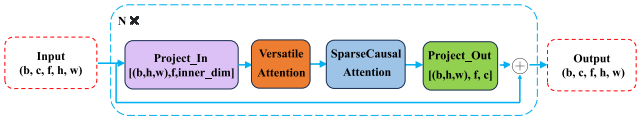

Fig. 3. Design of Motion Module. The motion module is inserted following each image layer of the pretrained SD to process video data.

C. Controllable Landscape Painting Video Generation

Our model, shown in Fig. 2, integrates two key components: a denoising U-Net and a Noise-Adapter. The denoising U-Net uses pretrained Stable Diffusion weights for enhanced image layer processing. To capture motion, a motion module layer follows each image layer, extending the T2I functionality to T2V and enabling dynamic motion synthesis. Our training incorporates unsupervised contrastive learning through the Noise-Adapter, which refines the noise output from the U-Net into contrastive learning samples, enhancing noise representation in the latent space. The structure of the motion module and the contrastive learning strategy are detailed in Section III-D.

During the training process, we keep the image layers of the denoising U-Net frozen and only update the motion module layers and the Noise-Adapter. During the inference phase, the video's length and smoothness are then further refined using a projection frame interpolation technique based on sparse optical flow, as detailed in Section III-E. This procedure not only amplifies the visual appeal of the animation but also enhances the video's coherence and aesthetic value, providing a novel method for dynamically presenting landscape paintings.

D. Motion Module and Training

1) Motion Module Design: In our motion module, the core design principle is to capture dynamic changes between frames through effective information exchange. To achieve this, we integrate two key attention mechanisms: Versatile Attention [24], [25] and Sparse-Causal Attention [26], [27], [28], which are specifically optimized for time series data and analyze the temporal dimension by processing feature tensors in a particular shape configuration. The dual-attention mechanism reserves the "artistic conception" across all frames, ensuring the visual harmony and coherence of the landscape video by balancing the global and local structure of the artwork. The structure of the motion module is demonstrated in Fig. 3.

The motion module is inserted between the image layers of the pretrained Stable Diffusion. When a batch of data passes through our motion module, the video data undergoes reshaping and dimension transformation in Project_In module, converting the input tensor from [batch, channels, frames, height, width] to [(batch × height × width) × frames × inner_dim]. Through this approach, the model can meticulously analyze interactions between individual frames and capture local and global features using the designed attention mechanism. Finally, the Project_Out module reverts the processed data back to its original dimensions and adds it to the initial input as the output employing a residual architecture.

The Versatile Attention [24], [25] module extends the traditional attention framework to all frames of the video. It computes relationships across all frames, enabling the model to capture broad contextual information and understand the overall dynamics of the video. This global attention mechanism establishes connections between distant frames, providing a comprehensive understanding of the sequence's context.

Building upon this, the Sparse-Causal Attention [26], [27], [28] mechanism simulates causal relationships within the frame sequence. It focuses specifically on the current frame and its preceding frame, ensuring that temporal causality is respected by limiting the attention to previous frames

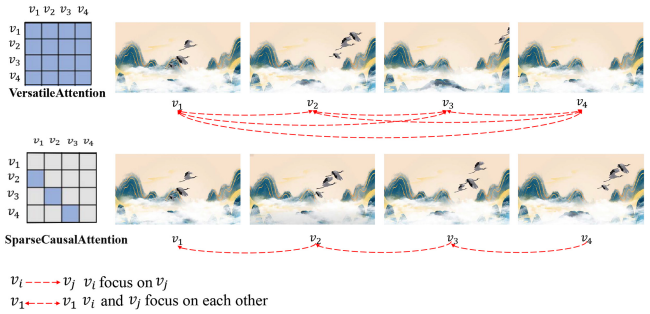

Fig. 4. Detailed explanation of the attention mechanism. The above represents Versatile Attention, where each frame is related to every other frame; the below represents Sparse-Causal Attention, where each frame only focuses on its previous frame.

only. This mechanism allows the model to effectively predict variations without introducing information from future frames.

While Versatile Attention captures the global context across all frames, Sparse-Causal Attention focuses on the immediate temporal dependencies between consecutive frames. The two modules are designed to complement each other: Versatile Attention provides a global understanding of the video, and Sparse-Causal Attention enforces temporal causality by ensuring that the model respects the flow of time. Therefore, these mechanisms work in a progressive manner, where the first establishes broad context and the second ensures smooth transitions and causal consistency within the frame sequence.

The mathematical expressions for both attention mechanisms can be unified as follows:

$$z = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \tag{6}$$

where $Q = W^Q z_{v_i}$, $K = W^K z_{v_j}$, $V = W^V z_{v_j}$, with z_{v_i} and z_{v_j} being the processed latent representations of the ith and the jth frames, respectively. W^Q, W^K, and W^V are learnable matrices that project the input into query, key, and value, and d is the dimensionality of the key and query features. For Sparse-Causal Attention, a masking mechanism is applied to include only information from the previous frame that has a causal influence on the current frame. Thus, the Versatile Attention mask is defined as

$$(M_{\text{Versatile}})_{ij} = 1. \tag{7}$$

The Sparse-Causal Attention mask is defined as

$$(M_{\text{Sparse-Causal}})_{ij} = \begin{cases} 1 & \text{if } j = i-1 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Specifically, when we implement the Versatile Attention mechanism, we derive the query feature from frame z_{v_i}, and the key and value features from the first frame z_{v_1} to the last frame z_{v_T}. When we implement the Sparse-Causal Attention mechanism, we derive the query feature from frame z_{v_i}, and the key and value features from the preceding frame $z_{v_{i-1}}$ (See Fig. 4 for a detailed visual depiction).

Through the combination of these two attention mechanisms, our motion modeling can not only capture the temporal dimensional features within video data but also ensure the consistency of content and smoothness of motion.

2) Unsupervised Contrastive Learning of Noise Corresponding to Video Frames in Latent Space: We introduce a contrastive learning strategy to improve the model. In video generation, our goal is to maintain similarity between adjacent frames while allowing for overall changes. Contrastive learning, which maximizes similarity between positive samples and contrasts them with negative ones, is well-suited for this task. Drawing inspiration from [49], which investigates how beneficial noise can simplify tasks and enhance model performance, we incorporate a noise-based strategy to guide the learning process in video frame generation. In particular, since video frames are created through successive denoising of predicted noise, we use contrastive learning to control the noise distribution, affecting the final frames. This aligns with previous work showing that noise can be an asset, rather than an impediment in model training [50]. It ensures smooth and coherent frame transitions while preserving the aesthetic and artistic flow of the landscape painting.

A common approach is to apply contrastive learning directly to the video frame z_0^i. However, we focus on the noise ε_θ^i instead, for two main reasons: First, the noise injection varies across time steps, and learning from ε_θ^i allows the model to capture the temporal dynamics crucial for denoising. Second, since denoising progresses from time step T to 1, focusing on noise aligns with this process, ensuring better learning of noise transitions and improving video coherence.

We designed a noise adapter to obtain noise samples. Adaptive methods [51], [52] play an important role in ensuring system performance. Specifically, the predicted noise is processed through a noise adapter(inspired by SimCLR [53] and SeCo [54]) to generate contrastive learning samples. The noise adapter adjusts the predicted noise instead of using it directly for contrastive learning, bridging the noise for denoising with the samples for contrastive learning. This approach is informed by [55], which highlights how learning reliable noise by contrastive learning can improve task performance, making it particularly well-suited in diffusion models.

Our noise-based contrastive learning strategy treats adjacent frames as positive pairs and frames where the frame number difference exceeds a threshold as negative pairs, using noise vectors for targeted training. Notably, we found that a simple MLP design for the noise adapter is sufficient. Specifically, for the predicted noise ϵ_θ^i of the ith frame output by the model, we assume that it passes through a noise adapter to obtain the corresponding contrastive learning sample $\hat{\epsilon}_\theta^i$. We calculate the similarity at step t using the cosine similarity

$$r_t^{(i,j)} = \frac{\hat{\epsilon_\theta}^i \cdot \hat{\epsilon_\theta}^j}{\|\hat{\epsilon_\theta}^i\| \|\hat{\epsilon_\theta}^j\|} \tag{9}$$

where $\hat{\epsilon_\theta}^i$ and $\hat{\epsilon_\theta}^j$ are two vectors output by noise adapter for the ith frame and the jth frame. Specifically, we calculate the contrastive learning loss for the noise vectors corresponding to all frames except the last one as anchors and calculate the average. For the noise vector corresponding to the ith frame,

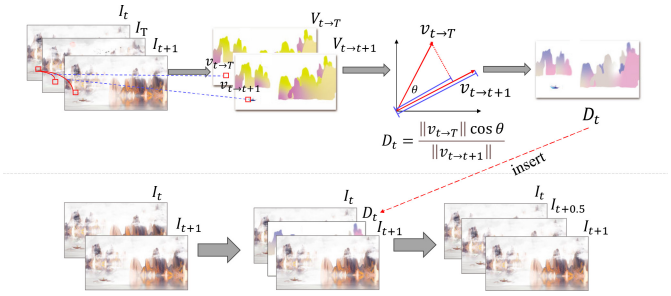

Fig. 5. Illustration of SOFP Interpolation. $V_{(t\to T)}$ is the optical flow from I_t to I_T and $V_{(t\to t+1)}$ is the optical flow from I_t to $I_{(t+1)}$. The projection result of $V_{(t\to T)}$ onto $V_{(t\to t+1)}$ is used to insert a new frame $I_{(t+0.5)}$ between two consecutive frames I_t and $I_{(t+1)}$.

we have the following definition:

$$l_t^{(i)} = -\log \frac{\exp\left(r_t^{(i,i+1)}/\tau\right)}{\exp\left(r_t^{(i,i+1)}/\tau\right) + \sum_{k=1}^{N} 1_{\{|k-i|>m\}} \exp\left(r_t^{(i,k)}/\tau\right)} \tag{10}$$

where m represents the frame sequence threshold for negative samples, τ denotes the temperature parameter for contrastive learning. The contrastive learning loss can be defined as

$$\mathcal{L}_{\text{con}} = \frac{1}{N-1} \sum_{k=1}^{N-1} l_t^{(k)} \tag{11}$$

where N is the total number of frames in the video. Through this contrastive learning of noise in the latent space, the model is encouraged to learn rich and robust temporal sequence feature representations without explicit annotations.

3) Training Objectives: In this section, we detail the training objectives of NADM. We sample videos to obtain a sequence of frames $x_0^{1:N}$, and encode each frame into the latent space $z_0^{1:N}$ using a pretrained Variational Autoencoder initially. Next, the latent codes are noised using the defined forward diffusion schedule: $z_t^{1:N} = \sqrt{\bar{\alpha}_t} z_0^{1:N} + \sqrt{1-\bar{\alpha}_t}\epsilon$. During the training process, our network receives noisy latent codes and corresponding textual prompts as input, predicting the noise intensity added to the latent codes. This process uses the L2 loss for calculation, and the loss is defined as

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbb{E}(x_0^{1:N}),y,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta\left(z_t^{1:N}, t, \tau_\theta(y)\right)\|_2^2\right]. \tag{12}$$

Note that during optimization, the pretrained weights of the base Stable Diffusion are frozen. Only the motion module and noise adapter are trainable. Combining the diffusion model with the contrastive learning strategy, the overall objective function of the NADM can be formulated as follows:

$$\mathcal{L} = \lambda_{\text{diff}} \cdot \mathcal{L}_{\text{diff}} + \lambda_{\text{con}} \cdot \mathcal{L}_{\text{con}}. \tag{13}$$

The hyperparameters λ_{diff} and λ_{con} are used to balance different losses and achieve improved performance.

E. Frame Interpolation Based on Sparse Optical Flow Projection

In the field of video processing, frame interpolation based on optical flow [56], [57], [58] is commonly used. We propose a frame interpolation strategy which is demonstrated in Fig. 5, to enhances the fluidity of dynamic elements in the landscape, ensuring smooth motion and visual continuity. This approach uses information from two adjacent frames along with the last frame of the video to interpolate additional frames between the two adjacent ones.

Optical flow is defined as the motion vector of pixel points between two frames. For two images, the optical flow vector indicates the new position of each pixel in the first image within the second frame. Given a pair of consecutive frames I_t and I_{t+1}, we calculate the optical flow $V_{t\to t+1}$ between the two frames. Similarly, we calculate the optical flow $V_{t\to T}$ between frame I_t and frame I_T.

In our research, we employ the advanced optical flow estimation algorithm RAFT [59] to precisely calculate the motion distance between pixels. Given a set of image triplets $\{I_t, I_T, I_{t+1}\}$, our primary task is to estimate two sets of optical flows: one from I_t to I_T, denoted as $V_{t\to T}$, and the other from I_t to I_{t+1}, denoted as $V_{t\to t+1}$. For each pixel point (x, y) in the image, we project the motion vector $V_{t\to T}(x, y)$ from I_t to I_T onto the motion vector $V_{t\to t+1}(x, y)$ from I_t to I_{t+1}. Thus, we define the distance mapping $D_t(x, y)$ as the ratio between these two vectors

$$D_t(x, y) = \frac{\|V_{t\to T}(x, y)\|\cos\theta}{\|V_{t\to t+1}(x, y)\|} \tag{14}$$

where θ is the angle between these two vectors. The projection results are used for interpolation between frame I_t and frame I_{t+1}.

The method of NADM is summarized in Algorithm 1.

IV. Experimentation

A. Experimental Setup

All our experiments are conducted on a single NVIDIA GeForce RTX 3090 with 24GB of VRAM. We perform all computations on a Linux-based system with Python 3.10.13, GCC 11.2.0, and NVIDIA CUDA 12.0 for parallel processing.

1) Training Settings: We directly use the trained Stable Diffusion v1.5 as the base model. We train the motion module and noise adapter using the proposed LPV-HD dataset. The video clips in the dataset are first sampled at a stride of 4, then resized and center-cropped to a resolution of 256 × 256. The frame length is set to 8, the batch size to 1, the learning rate to 1×10^{-5}, and the total number of training steps to 40k. In the contrastive learning of noise, we set the threshold for negative samples to 4, and the temperature parameter τ to 0.07. We use a linear beta schedule as in AnimateDiff [24], where $\beta_{\text{start}} = 0.00085$ and $\beta_{\text{end}} = 0.012$. For loss function, we set λ_{diff} as 1 and λ_{con} as 0.07.

2) Inference Settings: At inference, we use DDIM [48] sampler. We only use our frame interpolation module during inference. Our frame interpolation strategy expands an N-frame video to 2N-1 frames by inserting new frames between the original adjacent frames. Both the number of sampling steps and the scale for text guidance are selectable

Algorithm 1: NADM's Main Learning Algorithm

Input:
– $z_0^{1:N}$: Latent code from source video
– $\tau_\theta(y)$: CLIP-encoded text prompt for the input video
Hyperparameters:
– τ: Temperature parameter used in contrastive learning
– m: Threshold for negative sample
– λ_{diff}, λ_{con}: Balance parameters for the loss function
– T: Total number of sampling times
– α_t: Parameters controlling the noise added

leftmargin=*, label=▷
- Training:

for $t \sim Uniform(\{1, \ldots, T\})$ **do**
  $z_t^{1:N} = \text{forward}(z_0^{1:N}, \epsilon^{1:N}, t)$ ▶ *Eq.* (2)
  $\epsilon_{\theta'*}^{1:N} \leftarrow d_{frozen}(z_t^{1:N}, t, \tau_\theta(y))$
  Δcalculate $M_{\text{Versatile}}$ ▶ *Eq.* (7)
  Δcalculate $M_{\text{Sparse-Causal}}$ ▶ *Eq.* (8)
  $\epsilon_\theta^{1:N} \leftarrow f(\epsilon_{\theta'}^{1:N}, M_{\text{Versatile}}, M_{\text{Sparse-Causal}})$
  $\mathcal{L}_{diff} = \text{Diffusion\_loss}(\epsilon^{1:N}, \epsilon_\theta^{1:N})$ ▶ *Eq.* (12)
  for *all* $i \in \{1, \ldots, N\}$ **do**
    $\hat{\epsilon}_\theta^i = g(\epsilon_\theta^i)$
  end
  for *all* $i \in \{1, \ldots, N-1\}$ **do**
    Calculate $l_t^{(i)}$ ▶ *Eq.* (10)
  end
  $\mathcal{L}_{\text{con}} = \text{Con\_loss}(l_t^{(1)}, l_t^{(2)}, \ldots, l_t^{(N-1)})$ ▶ *Eq.* (11)
  $\mathcal{L} = \lambda_{\text{diff}} \cdot \mathcal{L}_{\text{diff}} + \lambda_{\text{con}} \cdot \mathcal{L}_{\text{con}}$
  update network f and noise-adapter g to minimize $\mathcal{L}$
end

leftmargin=*, label=▷
- Inference:

for $t = T$ **to** 1 **do**
  $\epsilon_{\theta'*}^{1:N} \leftarrow d_{frozen}(z_t^{1:N}, t, \tau_\theta(y))$
  Δcalculate $M_{\text{Versatile}}$ ▶ *Eq.* (7)
  Δcalculate $M_{\text{Sparse-Causal}}$ ▶ *Eq.* (8)
  $\epsilon_\theta^{1:N} \leftarrow f(\epsilon_{\theta'}^{1:N}, M_{\text{Versatile}}, M_{\text{Sparse-Causal}})$
  $z_{t-1}^{1:N} = \text{DDIM\_sample}(z_t^{1:N}, \epsilon_\theta^{1:N})$ ▶ *Eq.* (4)
end
$z_0^{1:2N-1} = \text{SOFP\_Interpolation}(z_0^{1:N})$

for users. When conducting qualitative and quantitative comparisons, we use DDIM with 25 sampling steps, and the scale for text guidance is 8.

B. Results

In this article, we present several qualitative results from our NADM model in Fig. 6, displaying only four frames per animation due to space limitations. We encourage readers to visit our website for higher-quality visuals. Our method skillfully combines brushwork beauty with the dynamic play of light and shadow typical of landscape painting into the T2V model. It uses the "five shades of ink"[1] technique, evident in the second row's right-side images, where varying ink depths create nuanced color changes, beautifully transitioning across elements like mountains and water. Additionally, our model employs the "combined colored ink"[2] technique to blend multiple colors harmoniously, as seen in the third row, offering a synchronized and aesthetically cohesive landscape portrayal. In addition, we find our method effectively differentiates main subjects from their surroundings. For example, the animation in the first row's right-side showcases raindrops and a house moving at distinct speeds and blurs, creating vivid scenes. Figs. 7–9 display more video clips. Results affirm that NADM adeptly transfers landscape painting's brushwork, techniques, and style to digital media, enriching modern digital content with traditional artistic values.

C. Baseline Comparisons

We evaluate our method against seven prominent baselines: AnimateLCM [60], which accelerates diffusion models via consistency decoupling; Gen-2 [61], a multimodal AI system for novel video creation from text; Text2Video-Zero [62], enhancing video consistency with frame-inter attention for zero-shot text-based video generation; Pika 1.0 [63], specializing in Text-to-3-D effects with parallel sampling; AnimateDiff-lightning [64], focusing on rapid video production through progressive adversarial diffusion distillation; ModelScope [65], which leverages a 3-D-Unet architecture to iteratively denoise Gaussian noise into video content and Lavie [66], which develops cascaded video LDMs.

1) Qualitative Results: Fig. 10 presents our qualitative results, comparing our NADM model with other generation models. The style of Gen-2 [61] deviates from landscape painting, evident in unrealistic details like the fisherman's attire. Similarly, AnimateLCM [60] and AnimateDifflightning [64] introduce modern elements and an overly detailed depiction, misaligning with the traditional painting style. Text2Video-Zero [62] aligns closer to the landscape style but suffers from instability and inconsistency in temporal context, like fluctuating pine tree shapes. Pika 1.0 [63] achieves consistency but lacks dynamic variation, and ModelScope [65] fails in vivid and dynamic depiction, rendering water waves as mere lines. Lavie [66] lacks the concept of mountains, and there is almost no dynamic variation between frames. In contrast, our NADM excels by incorporating raindrops, mist, and boats moving at varied speeds, capturing the dynamic essence of ink wash painting and harmoniously blending movement with stillness, accurately reflecting landscape painting aesthetics.

2) Quantitative Results: Our NADM is assessed against baselines using both automatic metrics and user evaluations. For quantitative indicators, we leverage CLIP-temp [67] for temporal consistency by computing cosine similarities between consecutive frames and CLIP-text [36] for text alignment by averaging CLIP scores between video frames and prompts. In addition, we report the commonly used FVD [68] and KVD [68] for video generation. Specifically, we calculate FVD

*θ' indicates that the results are obtained from a model with frozen parameters.

[1]The "five shades of ink" technique is a traditional painting method that uses varying depths of ink to create a spectrum of shades.

[2]The "combined colored ink" technique refers to a method in painting where different colored inks are blended together to achieve harmonious and unified effects.

Fig. 6. Main results. Our NADM creates high-quality videos in the style of landscape painting.

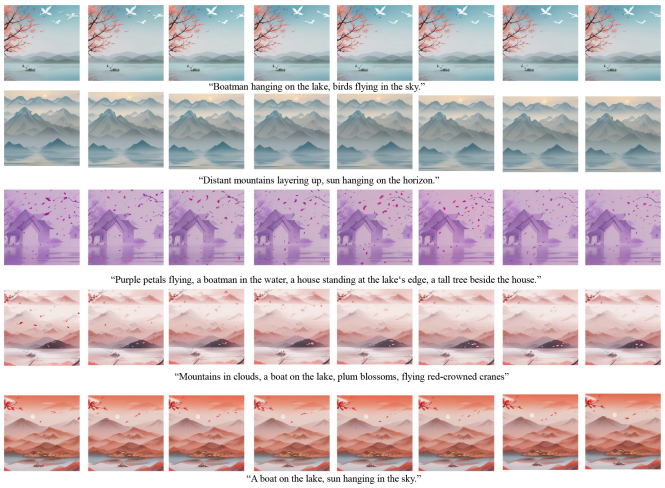

Fig. 7. Additional sample results of NADM (1/3).

and KVD between real and generated videos with 16 frames, which we refer to as FVD_{16} and KVD_{16}, for video evaluation. Besides, we report inception score (IS) [69], which reflects high quality of samples by calculating conditional entropy. We further evaluate the training computational requirements of NADM with baselines.

Table I presents a detailed comparison of NADM with baselines. Quantitative metric analysis shows that NADM performs better in all metrics compared to baselines. Besides, our strategy enables efficient training on an RTX 3090, balancing technical performance with training practicability.

D. Result Show

1) Cost-Effective With High Efficacy: The "GPUs for Training" column in Table I compares the computational requirements of NADM and baselines. Pika 1.0 and Gen-2, nonopen-source commercial models, are presumed to have much higher GPU demands. According to Tim Dettmers [70], ModelScope's parameter count (1.7 billion) requires at least 16GB GPU memory, such as an RTX 3090. Text2Video-Zero needs no training but lacks smooth interframe transitions. AnimateLCM uses eight A800 GPUs, Lavie uses eight A100

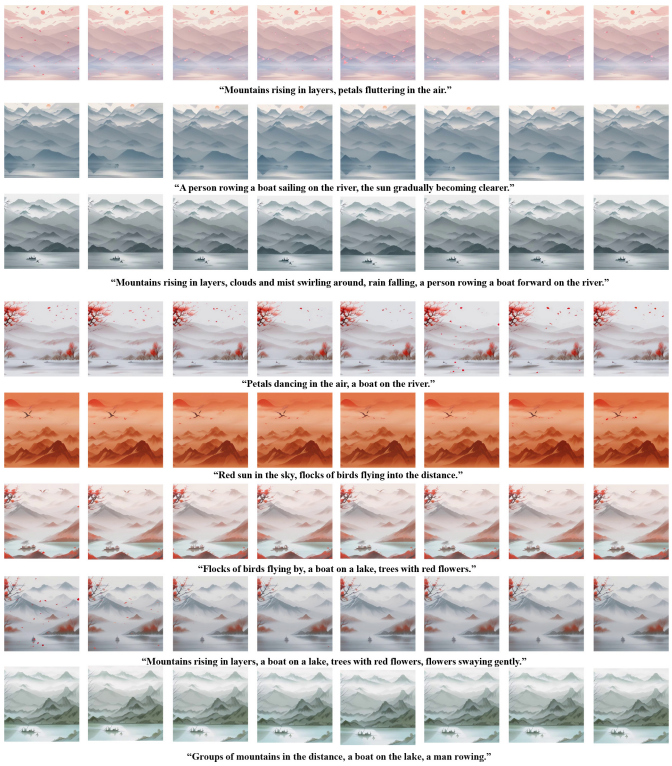

Fig. 8. Additional sample results of NADM (2/3).

GPUs, and AnimateDiff-lightning requires 64 A100 GPUs for training, while NADM operates on a single RTX 3090.

This efficiency is achieved by incorporating small, trainable modules into the frozen Stable Diffusion model, significantly reducing the number of parameters to train. This design minimizes GPU requirements while maintaining high-quality landscape painting video generation. Compared to ModelScope, which has similar hardware needs, NADM produces videos with superior fidelity and artistic quality, offering a cost-effective and accessible solution.

2) User Study: 500 participants, including creators, enthusiasts, and aesthetes of landscape painting, were surveyed to evaluate our generated videos for Text Alignment, Temporal

TABLE I
QUANTITATIVE RESULTS WITH BASELINES. ADL STANDS FOR ANIMATEDIFF-LIGHTNING. * REPRESENTS INFERRED RESULT. ↑ AND ↓ IMPLY HIGHER AND LOWER VALUES ARE BETTER, RESPECTIVELY

Model	FVD_{16}↓	KVD_{16}↓	Clip-Text↑	Clip-Temp↑	IS↑	GPUs for Training
Pika 1.0 [63]	$137.8_{\pm 5.4}$	$35.0_{\pm 2.4}$	24.564	0.972	$21.39_{\pm 0.25}$	Details not disclosed
AnimateLCM [60]	$133.3_{\pm 5.3}$	$25.2_{\pm 1.7}$	29.987	0.970	$17.51_{\pm 0.13}$	A800 × 8
Gen-2 [61]	$79.4_{\pm 1.9}$	$23.2_{\pm 2.3}$	30.309	0.978	$26.43_{\pm 0.23}$	Details not disclosed
Lavie [66]	$84.3_{\pm 3.2}$	$21.2_{\pm 3.7}$	29.143	0.928	$20.76_{\pm 0.23}$	A100 × 8
T2V-Zero [62]	$90.3_{\pm 2.7}$	$39.2_{\pm 2.9}$	27.633	0.980	$18.26_{\pm 0.18}$	No training required
ADL [64]	$94.5_{\pm 4.4}$	$28.7_{\pm 2.2}$	30.421	0.962	$17.83_{\pm 0.14}$	A100 × 64
ModelScope [65]	$104.8_{\pm 5.5}$	$32.2_{\pm 1.9}$	29.562	0.961	$13.81_{\pm 0.10}$	Single RTX 3090*
NADM(ours)	$\mathbf{53.7}_{\pm 0.8}$	$\mathbf{10.5}_{\pm 0.5}$	**30.526**	**0.981**	$\mathbf{21.97}_{\pm 0.28}$	Single RTX 3090

Fig. 9. Additional sample results of NADM (3/3).

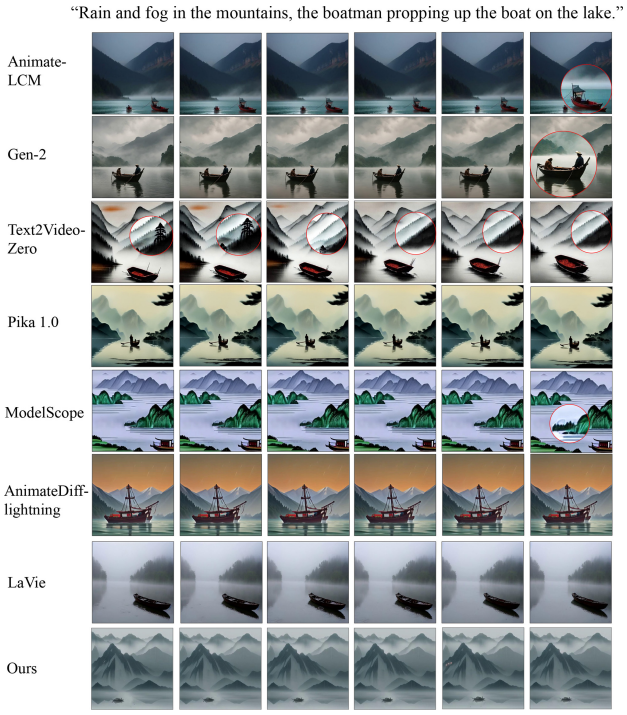

Fig. 10. Qualitative comparisons with baseline methods. Best viewed with zoomed-in.

Consistency, adherence to landscape painting style, and overall quality. Due to the lack of uniform quantitative indicators for style fidelity, these participants' insights are crucial for our research. Fig. 11 shows our NADM model significantly outshining all baseline methods in the user study. The result highlights how subjective style preferences can profoundly affect perceptions, indicating that a distinctive style may have a more lasting impact.

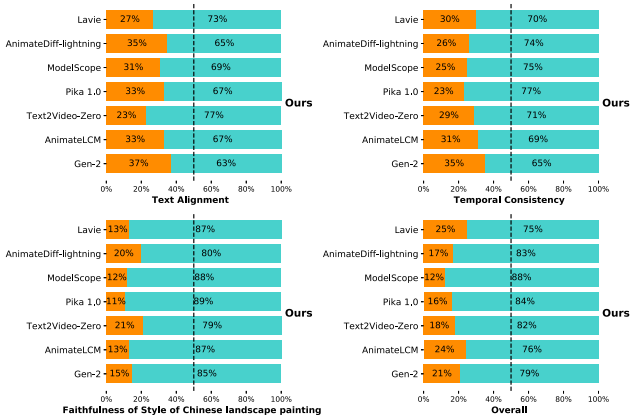

Fig. 11. User preference studies. NADM outperforms all baselines from overall aspects.

E. Ablation Studies

We primarily conducted ablation studies on the proposed two types of attention mechanisms(SC-Attn refers to Sparse-Casual Attention and Ver-Attn refers to Versatile Attention) within the motion module's residual architecture, and the contrastive learning of noise for video frames (CLofNoise), individually removing each design to assess its impact. The findings, illustrated in Fig. 12, reveal that omitting SC-Attn leads to discontinuous frame transitions while maintaining content consistency. In contrast, omitting Ver-Attn leads to leads to incoherence in the overall video. Conversely, excluding CLofNoise maintains content uniformity and smooth transitions but lacks in capturing dynamic changes, such as petal movement in the video.

TABLE II
QUANTITATIVE RESULTS OF ABLATION STUDY

Model	Clip-Text↑	Clip-Temp↑	FVD_{16}↓	KVD_{16}↓	IS↑
without SC-Attn	28.862	0.973	$92.8_{\pm 2.8}$	$28.6_{\pm 1.9}$	$15.14_{\pm 0.27}$
without Ver-Attn	28.028	0.953	$97.1_{\pm 3.2}$	$17.1_{\pm 1.6}$	$14.57_{\pm 0.13}$
without CLofNoise	29.864	0.964	$103.9_{\pm 3.6}$	$16.9_{\pm 1.2}$	$17.07_{\pm 0.11}$
NADM	**30.526**	**0.981**	$\mathbf{53.7}_{\pm 0.8}$	$\mathbf{10.5}_{\pm 0.5}$	$\mathbf{21.97}_{\pm 0.28}$

TABLE III
ABLATION STUDIES ON NEGATIVE SAMPLE THRESHOLD

Threshold	Clip-Text↑	Clip-Temp↑	FVD_{16}↓	KVD_{16}↓	IS↑
2	30.462	0.975	$57.2_{\pm 1.1}$	$18.6_{\pm 2.9}$	$18.13_{\pm 0.17}$
3	30.028	0.980	$55.3_{\pm 0.7}$	$15.3_{\pm 2.6}$	$20.54_{\pm 0.16}$
5	**30.575**	0.974	$58.4_{\pm 1.2}$	$13.3_{\pm 1.5}$	$20.07_{\pm 0.11}$
6	29.942	0.964	$60.9_{\pm 2.1}$	$16.6_{\pm 2.2}$	$18.56_{\pm 0.27}$
4	30.526	**0.981**	$\mathbf{53.7}_{\pm 0.8}$	$\mathbf{10.5}_{\pm 0.5}$	$\mathbf{21.97}_{\pm 0.28}$

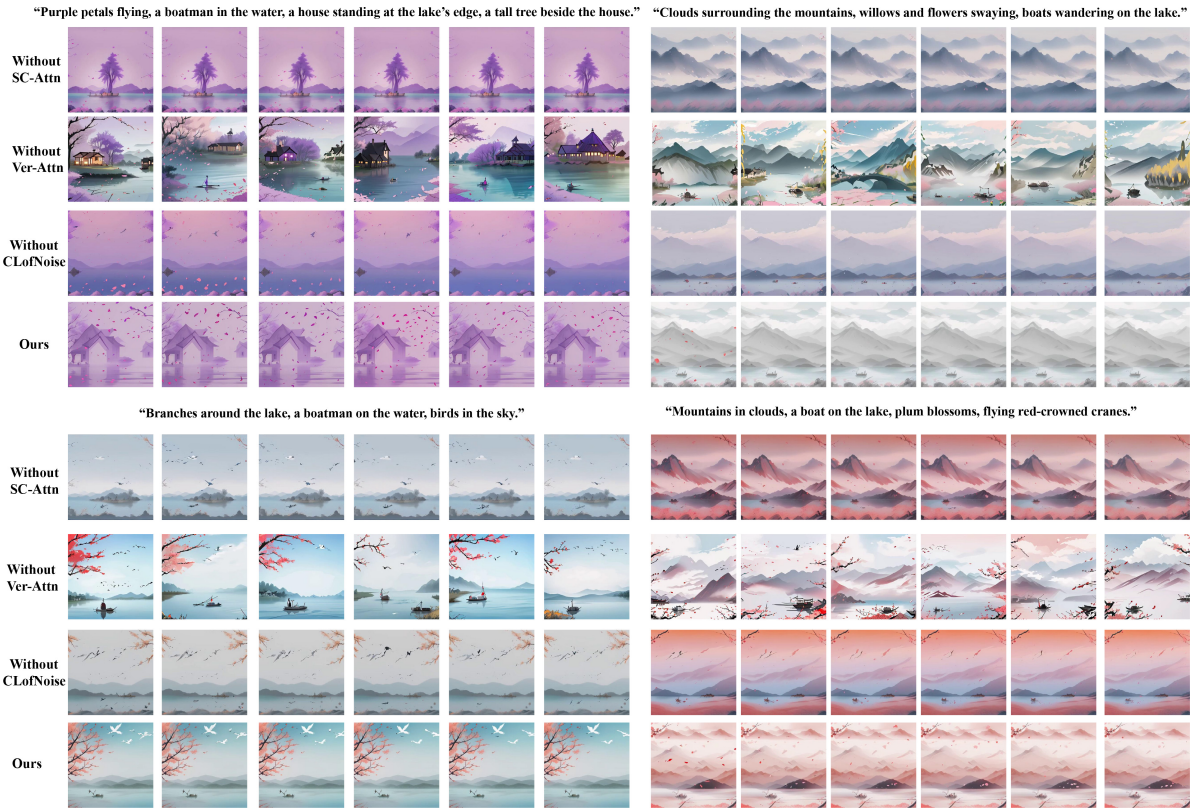

Fig. 13. Additional ablation study (1/2).

TABLE IV
ABLATION STUDIES ON POSITIVE SAMPLES

Threshold	Clip-Text↑	Clip-Temp↑	FVD_{16}↓	KVD_{16}↓	IS↑
All frames within the threshold	30.522	**0.985**	$56.2_{\pm 2.2}$	$16.3_{\pm 2.1}$	$17.26_{\pm 0.31}$
The next adjacent frame	**30.526**	0.981	$\mathbf{53.7}_{\pm 0.8}$	$\mathbf{10.5}_{\pm 0.5}$	$\mathbf{21.97}_{\pm 0.28}$

Figs. 13 and 14 present the results of our ablation study comparison. The landscape painting videos generated by our complete model are superior to those produced by the three ablated models, showing improvements in both style and detail. The quantitative results, detailed in Table II, highlight the significance of the SC-Attn, Ver-Attn and CLofNoise in enhancing textual accuracy and temporal coherence. These qualitative and quantitative assessments confirm the vital roles of our model's components in achieving high-quality outputs.

We conducted ablation experiments on different schemes for selecting positive and negative samples in contrastive learning, with results shown in Tables III and IV. We explored negative sample thresholds of 2, 3, 4, 5, and 6, and considered positive samples as either the adjacent next frame or all frames within the threshold. A threshold that is too high may result in minimal content variation, while one that is too low can cause excessive variation and a lack of coherence. Using all frames within the threshold as positive samples may lead to excessive

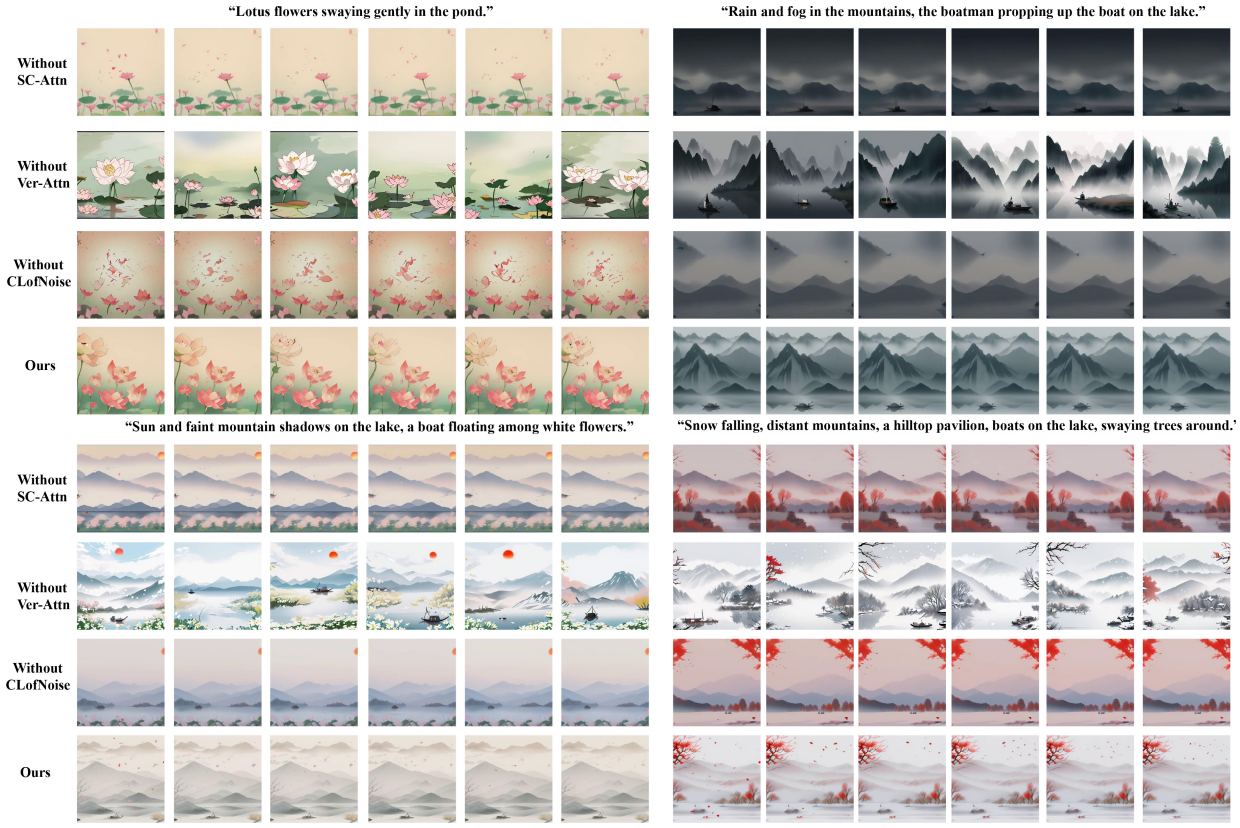

Fig. 14. Additional ablation study (2/2).

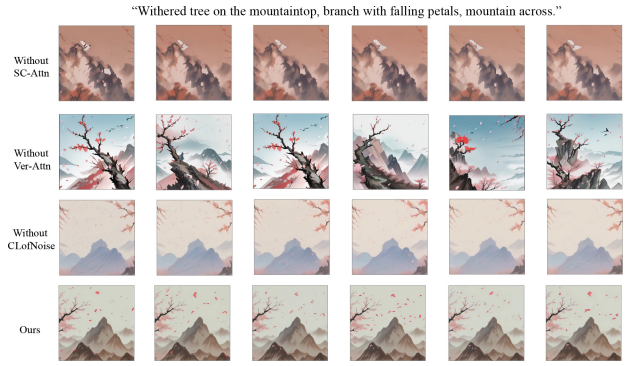

Fig. 12. Ablation study on sparse-casual attention, versatile attention and contrastive learning of noise.

similarity, resulting in a loss of differences and transitions. Our experiments confirm the effectiveness of our chosen approach.

V. Conclusion

In this article, we proposed a new text-to-video conversion network, NADM, designed to generate videos in the style of landscape paintings. Our approach introduces key innovations, including the LPV-HD dataset, a specialized motion module, contrastive learning of noise in latent space, and a frame interpolation strategy based on optical flow projection. These advancements enable our framework to effectively capture the dynamic beauty and poetic essence of landscape painting imagery. Extensive experiments demonstrate significant improvements in visual quality and temporal coherence. We hope our framework contributes to the preservation and promotion of landscape painting, while inspiring more studies that bridge traditional art with modern technology.

However, there are still some limitations. First, freezing the parameters of our underlying model during training preserves generation capabilities but limits flexibility in depicting fine details. Second, while NADM excels in video content and style diversity, it still needs improvement in controlling complex landscape lines. Third, our frame interpolation relies on optical flow projections, and due to the limitations of current optical flow estimators, this affects the smoothness and coherence of the video output. We leave these challenges for future work.

Appendix
User Study Details

We conducted a user study on generated videos to compare our method against six baselines. In this study, we invited 500 raters to assess video pairs, each generated by two different methods, including our method and one of the baseline models. After viewing each video pair, raters answered the following four questions.

1) Which video clip matches the text better? Please select the one that better represents the given text description.

2) Which video clip has higher consistency? Please select the one that looks more smooth as a video clip.
3) Which video clip exhibits the most distinct landscape painting style? Please select the best one.
4) Which video gives you the best overall impression? Please consider all aspects.

ACKNOWLEDGMENT

The authors sincerely appreciate the constructive comments from the anonymous reviewers, which significantly contributed to the revision of this article.

REFERENCES

[1] J. Zhang, Y. Han, J. Tang, Q. Hu, and J. Jiang, "Semi-supervised image-to-video adaptation for video action recognition," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 960–973, Apr. 2017.
[2] L. Zhang, P. Jing, Y. Su, C. Zhang, and L. Shaoz, "SnapVideo: Personalized video generation for a sightseeing trip," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3866–3878, Nov. 2017.
[3] X. Chang, F. Chao, C. Shang, and Q. Shen, "Sundial-GAN: A cascade generative adversarial networks framework for deciphering oracle bone inscriptions," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1195–1203.
[4] A. Xue, "End-to-end Chinese landscape painting creation using generative adversarial networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 3863–3871.
[5] D. Lin, Y. Wang, G. Xu, J. Li, and K. Fu, "Transform a simple sketch to a Chinese painting by a multiscale deep neural network," *Algorithms*, vol. 11, no. 1, p. 4, 2018.
[6] Y. Wang, W. Zhang, and P. Chen, "ChinaStyle: A mask-aware generative adversarial network for Chinese traditional image translation," in *Proc. SIGGRAPH Asia Tech. Briefs*, 2019, pp. 5–8.
[7] Z. L. Wang, J. Zhang, Z. Ji, J. Bai, and S. Shan, "CCLAP: Controllable Chinese landscape painting generation via latent diffusion model," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2023, pp. 2117–2122.
[8] L. Li et al., "AutoDiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 7082–7091.
[9] P. Jiang, M. Lin, and F. Chao, "Move and act: Enhanced object manipulation and background integrity for image editing," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2025, pp. 4039–4047. [Online]. Available: https://arxiv.org/abs/2407.17847
[10] M. Turner, "Classical Chinese landscape painting and the aesthetic appreciation of nature," *J. Aesthet. Educ.*, vol. 43, no. 1, pp. 106–121, 2009.
[11] S. E. Lee, "Chinese landscape painting," *Bull. Clevel. Museum Art*, vol. 41, no. 9, pp. 199–201, 1954.
[12] J. An et al., "Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation," 2023, *arXiv:2304.08477*.
[13] U. Singer et al., "Make-a-video: Text-to-video generation without text-video data," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–13.
[14] W. Wang et al., "VideoFactory: Swap attention in spatiotemporal diffusions for text-to-video generation," in *Proc. ICLR*, 2023, pp. 1–30.
[15] J. Xing et al., "Make-your-video: Customized video generation using textual and structural guidance," *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 2, pp. 1526–1541, Feb. 2025.
[16] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng, "MagicVideo: Efficient video generation with latent diffusion models," 2023, *arXiv:2211.11018*.
[17] Z. Deng, X. He, and Y. Peng, "Efficiency-optimized video diffusion models," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 7295–7303.
[18] Z. Deng, X. He, Y. Peng, X. Zhu, and L. Cheng, "MV-diffusion: Motion-aware video diffusion model," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 7255–7263.
[19] D. Jin, Q. Yu, L. Yu, and M. Qi, "SAW-GAN: Multi-granularity text fusion generative adversarial networks for text-to-image generation," *Knowl.-Based Syst.*, vol. 294, Jun. 2024, Art. no. 111795.
[20] S. Zhu et al., "Champ: Controllable and consistent human image animation with 3D parametric guidance," in *Proc. 18th Eur. Conf. Comput. Vis.*, 2024, pp. 145–162.
[21] Y. Jin et al., "Pyramidal flow matching for efficient video generative modeling," 2025, *arXiv:2410.05954*.
[22] S. Yuan et al., "MagicTime: Time-lapse video generation models as metamorphic simulators," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 8, 2025, doi: 10.1109/TPAMI.2025.3558507.
[23] S. Chen et al., "GenTron: Diffusion transformers for image and video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 6441–6451.
[24] Y. Guo et al., "AnimateDiff: Animate your Personalized text-to-image diffusion models without specific tuning," 2024, *arXiv:2307.04725*.
[25] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[26] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," 2019, *arXiv:1904.10509*.
[27] A. Hertz, R. Mokady, J. M. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022, *arXiv:2208.01626*.
[28] Y. Ma et al., "Follow your pose: Pose-guided text-to-video generation using pose-free videos," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 4117–4125.
[29] G. Mittal, T. Marwah, and V. N. Balasubramanian, "Sync-DRAW: Automatic video generation using deep recurrent attentive architectures," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1096–1104.
[30] Y. Pan, Z. Qiu, T. Yao, H. Li, and T. Mei, "To create what you tell: Generating videos from captions," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1789–1798.
[31] T. Marwah, G. Mittal, and V. N. Balasubramanian, "Attentive semantic video generation using captions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1426–1434.
[32] T. Gupta, D. Schwenk, A. Farhadi, D. Hoiem, and A. Kembhavi, "Imagine this! scripts to compositions to videos," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, pp. 610–626.
[33] Y. Liu, X. Wang, Y. Yuan, and W. Zhu, "Cross-modal dual learning for sentence-to-video generation," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1239–1247.
[34] Z. Li, H. Liang, H. Wang, X. Zheng, J. Wang, and P. Zhou, "A multi-modal vehicle trajectory prediction framework via conditional diffusion model: A coarse-to-fine approach," *Knowl.-Based Syst.*, vol. 280, Nov. 2023, Art. no. 110990.
[35] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, "TokenFlow: Consistent diffusion features for consistent video editing," 2023, *arXiv:2307.10373*.
[36] C. Qi et al., "FateZero: Fusing attentions for zero-shot text-based video editing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 15932–15942.
[37] R. Wu, L. Chen, T. Yang, C. Guo, C. Li, and X. Zhang, "LAMP: Learn a motion pattern for few-shot-based video generation," 2023, *arXiv:2310.10769*.
[38] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.
[39] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517010.
[40] S. Yuan et al., "Learning to generate poetic Chinese landscape painting with calligraphy," in *Proc. 31st Int. Joint Conf. Artif. Intell. (IJCAI)*, 2022, pp. 1–4.
[41] L. Zhou, Q.-F. Wang, K. Huang, and C.-H. Lo, "An interactive and generative approach for Chinese Shanshui painting document," in *Proc. Int. Conf. Doc. Anal. Recognit. (ICDAR)*, 2019, pp. 819–824.
[42] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5527812.
[43] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," 2023, *arXiv:2309.16499*.
[44] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10674–10685.
[45] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," 2021, *arXiv:2105.05233*.

[46] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2015, pp. 234–241.

[48] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–22.

[49] H. Zhang, S.-Y. Huang, and X. Li, "Variational positive-incentive noise: How noise benefits models," 2025, *arXiv:2306.07651*.

[50] X. Li, "Positive-incentive noise," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 8708–8714, Jun. 2024.

[51] H. Jiao, Q. Shen, Y. Shi, and P. Shi, "Adaptive tracking control for uncertain cancer-Tumor-immune systems," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 18, no. 6, pp. 2753–2758, Nov./Dec. 2021.

[52] J. Yu, P. Shi, W. Dong, and H. Yu, "Observer and command-filter-based adaptive fuzzy output feedback control of uncertain nonlinear systems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 9, pp. 5962–5970, Sep. 2015.

[53] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.

[54] T. Yao, Y. Zhang, Z. Qiu, Y. Pan, and T. Mei, "SeCo: Exploring sequence supervision for unsupervised representation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10656–10664.

[55] H. Zhang, Y. Xu, S. Huang, and X. Li, "Data augmentation of contrastive learning is estimating positive-incentive noise," 2024, *arXiv:2408.09929*.

[56] J. Park, C. Lee, and C.-S. Kim, "Asymmetric bilateral motion estimation for video frame interpolation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14539–14548.

[57] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 624–642.

[58] Z. Zhong, G. Krishnan, X. Sun, Y. Qiao, S. Ma, and J. Wang, "Clearer frames, anytime: Resolving velocity ambiguity in video frame interpolation," in *Proc. 18th Eur. Conf. Comput. Vis.*, 2024, pp. 346–363.

[59] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 402–419.

[60] F.-Y. Wang et al., "AnimateLCM: Accelerating the animation of Personalized diffusion models and adapters with decoupled consistency learning," 2024, *arxiv:2402.00769*.

[61] (Runway, Manhattan, NY, USA). *Gen-2 by Runway—Research*. Accessed: Mar. 25, 2024. [Online]. Available: http://runwayml.com

[62] L. Khachatryan et al., "Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 1–11.

[63] "Pika—Pika.art." Pika Labs. Accessed: Mar. 25, 2024. [Online]. Available: https://pika.art

[64] S. Lin and X. Yang, "AnimateDiff-lightning: Cross-model diffusion distillation," 2024, *arXiv:2403.12706*.

[65] 2024, "Text-to-video-synthesis model in open domain," Dataset. [Online]. Available: http://modelscope.cn

[66] Y. Wang et al., "LaVie: High-quality video generation with cascaded latent diffusion models," *Int. J. Comput. Vis.*, vol. 133, pp. 3059–3078, May 2025.

[67] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," 2023, *arXiv:2302.03011*.

[68] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," 2022, *arXiv:1812.01717*.

[69] S. Yu et al., "Generating videos with dynamics-aware implicit generative adversarial networks," 2022, *arXiv:2202.10571*.

[70] T. Dettmers. "The best GPUs for deep learning in 2023—An in-depth analysis." 2023. [Online]. Available: https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/

Ding-Ming Liu received the B.E. degree from the Department of Artificial Intelligence, Xiamen University, Xiamen, China, in 2025.

He is currently working as a Research Assistant with Xiamen University. His research interests include computer vision and generative models.

Shao-Wei Li received the B.E. degree from the Department of Artificial Intelligence, Xiamen University, Xiamen, China, in 2025, where he is currently pursuing the M.Sc. degree.

His research interests include large-language models and model compression.

Ruo-Yan Zhou received the B.E. degree in computer science and technology from Xiamen University, Xiamen, China. She is currently pursuing the master's degree with the Department of Computer Science, University of Hong Kong, Hong Kong, in 2025.

Her research interests focus on computer vision and generative models, where she strives to explore foundational methods and their practical applications.

Li-Li Liang is currently pursuing the B.E. degree with the School of Information, Xiamen University, Xiamen, China.

Her research interests include diffusion models and machine learning.

Yong-Guan Hong is currently pursuing the B.E. degree with the School of Information, Xiamen University, Xiamen, China.

His research interests include computer vision and generative models.

Yuan-Ze Zeng is currently pursuing the B.E. degree with the School of Information, Xiamen University, Xiamen, China.

Her research interests include diffusion models and 3-D Gaussian splatting.

Xiang Chang (Member, IEEE) received the B.Eng. degree in cognitive science from the Department of Artificial Intelligence, Xiamen University, Xiamen, China, in 2019, and the Ph.D. degree in robotics from the Department of Computer Science, Aberystwyth University, Aberystwyth, U.K., in 2024.

He is currently a Lecturer with the Faculty of Business and Physical Sciences, Aberystwyth University. His research interests include deep-reinforcement learning and imitation learning-based robotic motion planning.

Li-Jiang Li received the B.S. and M.Sc. degrees in intelligence science and technology from the School of Informatics, Xiamen University, Xiamen, Fujian, China, in 2022, where she is currently pursuing the Ph.D. degree.

Her academic interests span various applications of artificial intelligence in visual perception and predictive modeling. Her research focuses on computer vision and machine learning.

Tian-Shuo Xu (Member, IEEE) received the B.Sc. and M.S. degrees in intelligence science and technology from the School of Informatics, Xiamen University, Xiamen, China, in 2020 and 2023, respectively. He is currently pursuing the Ph.D. degree in artificial intelligence with The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China.

His research interests include computer vision and machine learning.

Fei Chao (Senior Member, IEEE) received the B.Sc. degree in mechanical engineering from Fuzhou University, Fuzhou, China, in 2004, the M.Sc. degree (Hons.) in computer science from the University of Wales, Aberystwyth, U.K., in 2005, and the Ph.D. degree in robotics from Aberystwyth University, Aberystwyth, in 2009.

He is currently an Associate Professor with the School of Informatics, Xiamen University, Xiamen, China, and an Adjunct Research Fellow with the Department of Computer Science, Aberystwyth University. He has published more than 100 peer-reviewed journals and conference papers. His research interests include machine-learning algorithms, large-language model compression, and network architecture search.

Changjing Shang received the Ph.D. degree in computing and electrical engineering from Heriot-Watt University, Edinburgh, U.K., in 1995.

She is a University Senior Research Fellow with the Department of Computer Science, Aberystwyth University, Aberystwyth, U.K., and a Fellow of the Learned Society of Wales (Welsh National Academy). She has published over 220 peer-reviewed papers and supervised more than 25 Ph.Ds/PDRAs, in the areas of pattern recognition, data mining and analysis, and systems control with uncertainty.

Qiang Shen received the Ph.D. degree in computing and electrical engineering from Heriot-Watt University, Edinburgh, U.K., in 1990, and the D.Sc. degree in computational intelligence from Aberystwyth University, Aberystwyth, U.K., in 2013.

He holds the Established Chair of Computer Science and is a Pro Vice-Chancellor with Aberystwyth University. His research interests include computational intelligence and its applications.

Dr. Shen is a recipient of the IEEE Fuzzy Systems Pioneer Award and a Fellow of the Royal Academy of Engineering.