Animation of Historical Images: A Video Diffusion Generation Method Guided by Depth Estimation

Yingqiang Chen[1], Hongliang Du[1,*]
[1]School of Media, Weifang University, Weifang 261021, Shandong, China
Corresponding author: Hongliang Du (e-mail: 18265696603@163.com).

ABSTRACT The use of artificial intelligence and smart algorithms for the restoration, enhancement and creation of videos from historical images continues to grow. To tackle the problems of spatial structure drift, inter-frame flickering, and motion discontinuity in the animation of historical images, we propose a video diffusion generation model based on depth estimation. Based on the video diffusion model, the model includes the historical image preprocessing module, the single-frame depth estimation module, the spatial structure encoding module, and the temporal motion compensation module, using the depth feature as a conditional constraint in the reverse denoising generation of the model. Experimental results show that the proposed method achieves a PSNR of 28.76, an improvement of 1.28 over conventional video diffusion models; SSIM increases to 0.883, LPIPS decreases to 0.108, and FVD decreases to 219.54, with both generation quality and temporal stability outperforming the comparison methods.

INDEX TERMS Historical footage; Animation conversion; Depth estimation; Video diffusion generation; Temporal consistency

I. INTRODUCTION

An important challenge is the use of artificial intelligence and intelligent algorithms for image restoration, video generation and preservation of digital cultural heritage, which is still being deepened. It has become an important avenue to introduce visual computing and historical storytelling: to bring historical footage into dynamic form. Traditional historical videos generally have problems like low resolution, high noise, scratches, obstructions, gray deficiencies and space hierarchy; directly created videos have problems like character drifting, background jitter, boundary distortion, and interframe flickering. Ma et al. performed a systematic survey of the evolution of video diffusion generation, the methods of video generation control, and the remaining open issues, highlighting that temporal consistency and controllable generation are still challenging issues in the field of video generation [1]. To tackle the challenging issue of controllable video generation, Cai et al. presented a four-dimensional guidance mechanism to the two-dimensional diffusion generation process as a new way of imposing structural constraints [2]. Rayo and Tous presented a temporal consistent video cartoonization approach to animated scene generation where the frame-to-frame stability is a key in the quality of animated scenes [3]. Xu et al. and Niu et al. enhanced the control of image-to-video generation from the character animation and motion field control aspects respectively [4], [5]. Gupta et al. investigated the effectiveness of diffusion models for photorealistic video generation, and Gu et al. extended the scope of video diffusion control to include 3D perception to capture spatial structures [6], [7].

To solve the problems of instability of the spatial structure of the historical image and lack of temporal continuity in historical image animation, based on the above research, we design a video diffusion generation method based on depth estimation. In this study, image preprocessing, single frame depth estimation, spatial structure encoding, diffusion generation and temporal compensation are included. Through the application of depth constraints, we hope to improve the foreground, the background, and the depth of scene in the animated historical images, which will make the images have more structure, and make the viewing continuity more credible.

II. MODELING THE HISTORICAL IMAGE ANIMATION CONVERSION TASK

The animation conversion of historical images can be defined as a conditional generation process from a single-frame historical image to a continuous video sequence. Let the input historical image be I_0. After denoising, grayscale equalization, scratch repair, and sharpness enhancement, an enhanced image I_e is obtained. A spatial depth map D_0 is then extracted via a depth estimation network. The image texture, spatial hierarchy, and motion conditions are jointly

fed into a video diffusion model to generate an animated video sequence $V = \{F_1, F_2, \ldots, F_T\}$ of length T [8]. The task mapping can be represented as:

$$V = G_\theta(I_e, E_d(D_0), M) \tag{1}$$

where G_θ denotes the video diffusion generative model with parameters θ; $E_d(\cdot)$ is the spatial depth encoder; M represents the motion control conditions; and F_t is the generated result for frame t [9]. This modeling process is illustrated in Figure 1, which shows the complete transformation path between historical image input, visual enhancement, depth estimation, diffusion generation, and video output.

To ensure the structural stability of the generated video, a comprehensive optimization objective is formulated:

$$L = \lambda_r L_r + \lambda_d L_d + \lambda_t L_t \tag{2}$$

where L_r is the image reconstruction loss, which constrains the consistency between the generated frame and the original historical image in terms of subject texture; L_d is the depth consistency loss, used to preserve the relationships between the foreground, background, and spatial hierarchy; and L_t is the temporal consistency loss, used to reduce inter-frame flickering and subject drift [10]. The coefficients λ_r, λ_d, and λ_t represent the weights of the different loss terms, respectively. In the experiments, $\lambda_r = 1.0$, $\lambda_d \in [0.1, 0.5]$, and $\lambda_t \in [0.2, 0.6]$ were set, and a grid search on the validation set determined the optimal values to be $\lambda_d = 0.3$ and $\lambda_t = 0.4$. At these values, the generated results demonstrate a balanced performance in terms of structural preservation and inter-frame stability.

Depth consistency is further expressed as:

$$L_d = \frac{1}{T} \sum_{t=1}^{T} \|E_d(D_0) - E_d(D_t)\|_1 \tag{3}$$

where D_t is the estimated depth map corresponding to frame t, and $\|\cdot\|_1$ denotes the 1-norm distance [11]. This formula is used to verify whether the depth guidance maintains a stable spatial structure across consecutive frames, providing a constraint foundation for subsequent diffusion generation and motion compensation.

III. CONSTRUCTION OF A DEPTH-ESTIMATION-GUIDED VIDEO DIFFUSION GENERATION MODEL

A. OVERALL MODEL ARCHITECTURE

The whole model design is composed of five parts: historical image pre-processing, depth estimation, spatial structure encoding, video diffusion generation, and temporal optimization. The input images are first processed using noise removal, scratch correction and contrast correction to obtain relatively stable images, and then use single frame depth estimation to obtain scene level information, and use the spatial structure encoder to convert into conditions features. The enhanced image, depth encoding, and motion conditions are then concatenated and fed into the video diffusion generation module where the consecutive frames are generated in a step-wise manner through a denoising process [12]. The temporal consistency constraint and motion compensation modules are further combined to reduce inter-frame drift, texture flicker and local distortions, and stabilize and ensure the structural credibility of the animated video. The general structure of the model is given in Figure 2.

B. HISTORICAL IMAGE PREPROCESSING AND VISUAL ENHANCEMENT MODULE

Historical images typically suffer from issues such as noise grain, scratches and obstructions, grayscale shifts, and local texture degradation. If fed directly into the diffusion model, these issues can easily lead to unstable depth estimates and drift in the details of generated frames [13]. Therefore, the preprocessing module first performs normalization, denoising, contrast enhancement, and detail restoration on the original image I_0 to obtain the enhanced image I_e. The grayscale normalization process can be expressed as:

$$I_n = \frac{I_0 - I_{\min}}{I_{\max} - I_{\min} + \varepsilon} \tag{4}$$

where I_n is the normalized image, $I_{\min}$ and $I_{\max}$ are the minimum and maximum grayscale values of the image, respectively, and ε is a very small constant to prevent the denominator from becoming zero [14]. To address the issue of uneven brightness distribution in historical images, an enhancement function is introduced:

$$I_e = \alpha I_n + \beta \cdot H(I_n) \tag{5}$$

In the equation, $H(I_n)$ represents the histogram-equalized image, while α and β are fusion weights that satisfy $\alpha+\beta = 1$. In the experiment, setting α to 0.6 and β to 0.4 improved edge sharpness while preserving the original texture [15]. The output of this module, I_e, serves as the base input for subsequent depth estimation and video diffusion generation.

C. SINGLE-FRAME DEPTH ESTIMATION AND SPATIAL STRUCTURE ENCODING

The single-frame depth estimation and spatial structure encoding module is used to extract spatial hierarchical information from 2D historical images that can constrain video generation [16]. After the enhanced image I_e is fed into the depth estimation network, the corresponding depth map D_0 is obtained, which can be expressed as:

$$D_0 = f_d(I_e; \theta_d) \tag{6}$$

where $f_d(\cdot)$ denotes the depth estimation network, θ_d represents the network parameters, and D_0 denotes the depth distribution results of a single-frame historical image [17]. Since historical images typically lack ground-truth depth annotations, the model uses a pre-trained monocular depth estimation network for initial prediction and combines this with edge-preserving filtering to smooth and correct

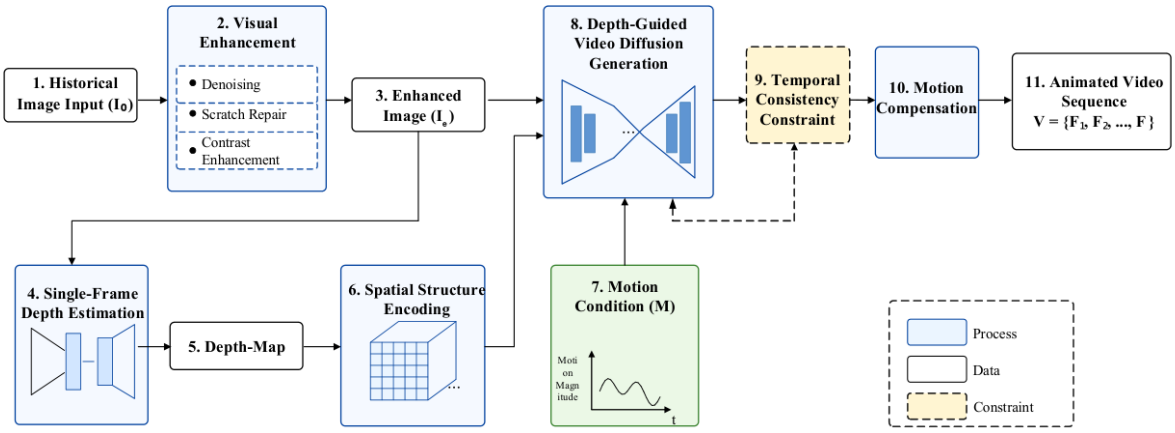

Fig. 1. Flowchart of the Historical Image Animation Conversion Task.

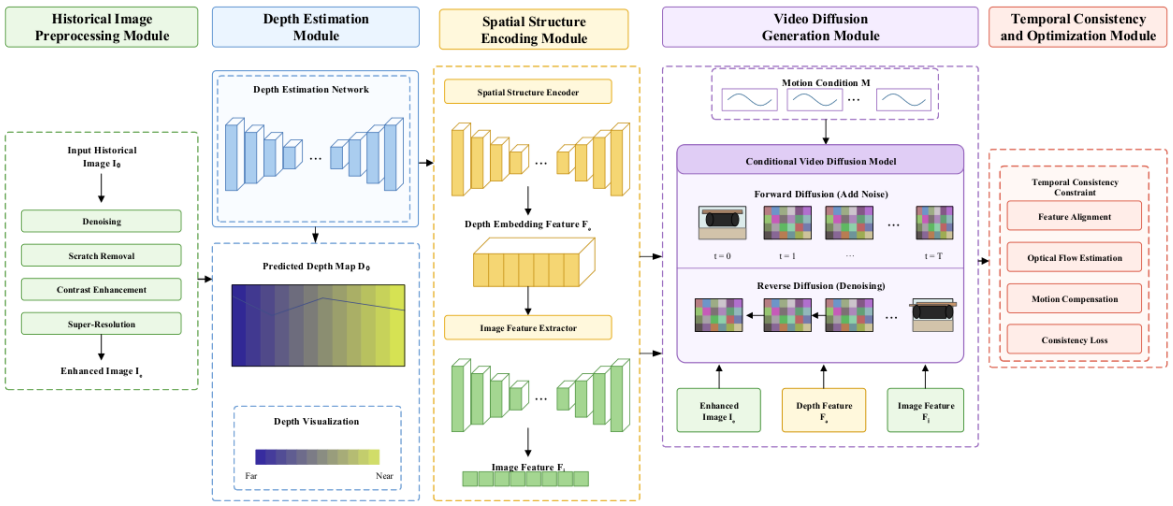

Fig. 2. Overall Model Architecture Diagram.

the depth map, thereby reducing depth discontinuities in building outlines, person boundaries, and background areas.

To enable depth information to participate in the diffusion generation process, D_0 must be converted into spatial structural features. The spatial structure encoding process can be expressed as:

$$F_d = E_d(D_0) \tag{7}$$

where $E_d(\cdot)$ is the spatial structure encoder, and F_d is the depth embedding feature [18]. The features mainly include the relationship between the foreground and background, scene depth, subject boundaries and occlusion hierarchy, and are input into the video diffusion model together with the image texture features. The generative model can use depth constraints to preserve a fairly consistent shape of the geometry from one frame to the next, which can help minimize character drift, building distortion and background jitter. The single-frame depth estimation process and spatial structure encoding process are shown in Figure 3.

D. DEPTH-GUIDED DIFFUSION GENERATION MECHANISM

The depth-guided diffusion generation mechanism uses enhanced image features F_i, depth-embedded features F_d, and motion conditions M as joint constraints to generate consecutive animation frames through a conditional diffusion process. This mechanism can also integrate the depth information into the denoising network, which can consider the image texture information and spatial hierarchical relationship in historical images during the whole generation process [19]. In the forward diffusion process, the features of the real frames are gradually corrupted by Gaussian noise to get latent representations with different noise levels; in the backward diffusion process, the noise is gradually estimated and removed based on the condition of the depth encoding, and finally obtain animation frames with temporal consistency. The noise prediction process can be expressed as:

$$\widehat{\varepsilon} = \varepsilon_\theta(z_t, t, F_i, F_d, M) \tag{8}$$

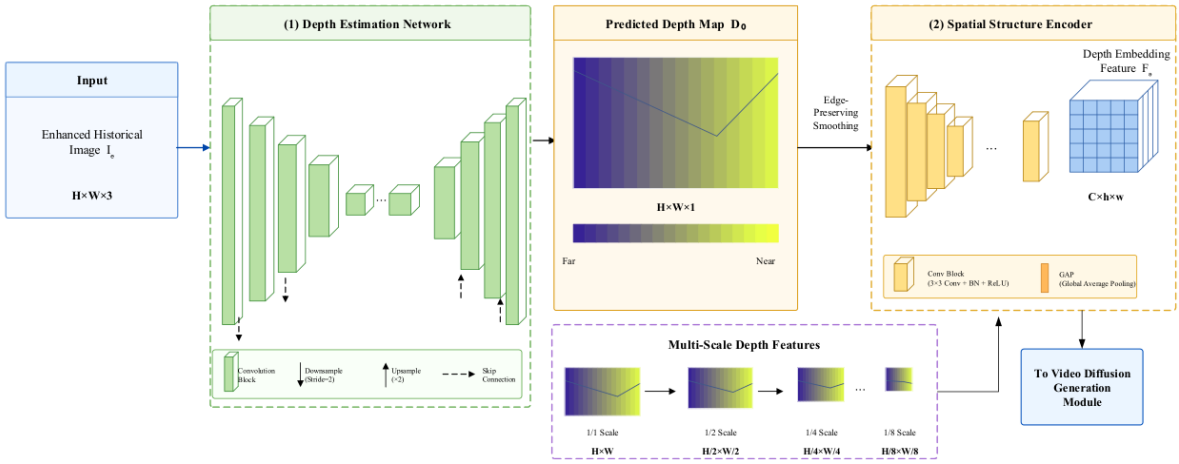

Fig. 3. Schematic diagram of single-frame depth estimation and spatial structure encoding.

where $\hat{\varepsilon}$ is the noise predicted by the model, ε_θ denotes the conditional denoising network with parameters θ, z_t is the noisy latent variable at step t, F_i represents image texture features, F_d denotes deep structural features, and M represents motion conditions [20]. To enhance the effect of depth constraints, F_d is injected into the diffusion network via a cross-attention mechanism, ensuring that the foreground subject, background buildings, road depth, and occlusion boundaries remain stable across consecutive frames.

During the training phase, a noise prediction loss is used to constrain the generation process:

$$L_\varepsilon = \|\varepsilon - \varepsilon_\theta(z_t, t, F_i, F_d, M)\|_2^2 \tag{9}$$

where ε represents the actual added noise, and L_ε represents the noise estimation error [21]. This mechanism reduces structural drift, background jitter, and local distortion in the animation of historical images, thereby improving the spatial credibility of the generated video.

E. TEMPORAL CONSISTENCY CONSTRAINTS AND MOTION COMPENSATION DESIGN

The temporal consistency constraint and motion compensation design are the two methods that are mainly used to solve the problems of flickering between the two consecutive frames, the subject drift problem, the background jitter problem, and the local distortion problem when animating historical images. This module involves processing the adjacent generated frames: Firstly, correspondences between people, building edges, and road structure are found between two successive frames by performing the feature alignment; secondly, the direction and the magnitude of the movement are obtained by combining the results of the feature alignment and the optical flow estimation. In case of misalignment of these areas, the motion compensation unit applies local corrections to the current frame, based on the structure of the previous one; this results in smoother motion of the subjects and stable background textures [22]. The temporal consistency constraint module is used to ensure temporal consistency of the colors, boundaries and depth levels between successive frames, thereby rejecting random disturbances from diffusion-based generation. The overall structure is shown in Figure 4.

IV. MODEL TRAINING AND OPTIMIZATION METHODS

The strategy of model training is to adopt a two-phase optimization: first, the historical image enhancement module and the depth estimation module are trained; and then the depth encoding module, the video diffusion generation module, and the temporal compensation module are trained together [23]. The training data comprises of improved images, depth maps and sequential video frames. The input resolution is fixed to 256×256 and the length of the video clip is fixed to 16 frames. The diffusion training phase is a forward noise addition and backwards denoising process, the forward diffusion can be written as:

$$q(z_t \mid z_0) = \mathcal{N}\left(\sqrt{\overline{\alpha}_t}\, z_0, (1 - \overline{\alpha}_t) I\right) \tag{10}$$

In the equation, z_0 is the original latent variable, z_t is the noisy latent variable at step t, $\overline{\alpha}_t$ is the cumulative noise retention coefficient, and I is the identity matrix [24]. To stabilize the training process, a cosine learning rate decay strategy is adopted:

$$\eta_e = \eta_{\min} + \frac{1}{2}(\eta_0 - \eta_{\min})\left(1 + \cos\frac{\pi e}{E}\right) \tag{11}$$

where η_e is the learning rate at the eth iteration, η_0 is the initial learning rate, $\eta_{\min}$ is the minimum learning rate, and E is the total number of training iterations [25]. Parameter updates are performed using the AdamW optimizer, with the update process as follows:

$$\theta_{e+1} = \theta_e - \eta_e \frac{\widehat{m}_e}{\sqrt{\widehat{v}_e} + \varepsilon} \tag{12}$$

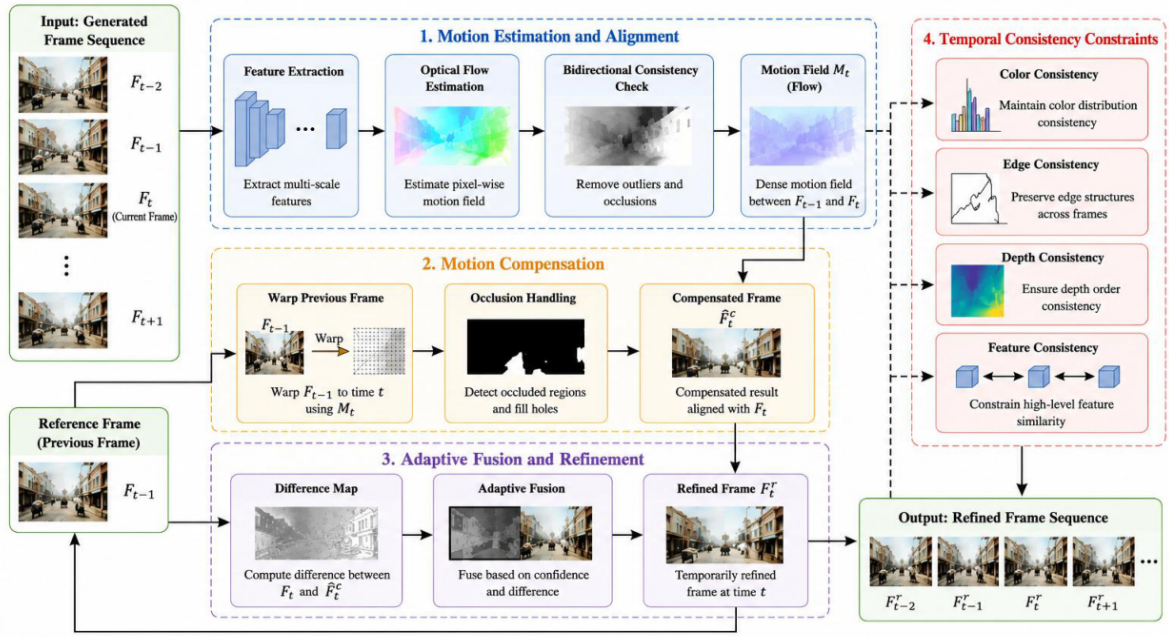

Fig. 4. Structural Diagram of Temporal Consistency Constraints and Motion Compensation.

where θ_e represents the model parameters at the eth iteration, $\widehat{m}_e$ and $\widehat{v}_e$ are the first- and second-order moment estimates, respectively, and ε is the stability term to prevent the denominator from becoming zero. The training parameter settings are shown in Table 1.

TABLE 1. Model Training Parameter Settings

Parameter Category	Parameter Name	Setting
Input Setting	Input Resolution	256×256
Video Setting	Frame Sequence Length	16 frames
Diffusion Setting	Diffusion Steps	1000
Optimizer	Optimization Method	AdamW
Learning Rate	Initial Learning Rate	10×10^{-4}
Batch Setting	Batch Size	8
Training Epochs	Epoch	200
Loss Weight	Depth / Temporal Weight	0.3 / 0.4
Hardware Environment	GPU	RTX 3090

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT CONFIGURATION

All experiments were run in the same hardware and software environment to guarantee consistency in the model training process and the comparison of results. The hardware platform was composed of an Intel i9 processor, 64 GB of RAM and NVIDIA RTX 3090 GPU with 24 GB of video memory, and the software environment was based on the Ubuntu 20.04 operating system, the PyTorch 2.0 deep learning framework and CUDA version 11.8. The experimental data included historical images, restored images and historical scene videos which were created by themselves. The data set was split into a training set (70%), a validation set (20%) and a test set (10%). All input images were uniformly scaled to 256×256 and the video clips were uniformly made of 16 frames. The configuration used in the experimental environment is given in Table 2.

TABLE 2. Experimental Environment Configuration

Configuration Category	Configuration Item	Setting
Hardware Environment	CPU	Intel Core i9
Hardware Environment	Memory	64 GB
Hardware Environment	GPU	NVIDIA RTX 3090
GPU Memory	GPU Memory	24 GB
Operating System	OS	Ubuntu 20.04
Deep Learning Framework	Framework	PyTorch 2.0
Acceleration Environment	CUDA	CUDA 11.8
Input Resolution	Image Size	256×256
Video Length	Sequence Length	16 frames
Data Split	Train/Validation/Test	7:2:1

B. DESIGN OF COMPARATIVE EXPERIMENTS

Multiple groups of comparison models were set up, such as traditional frame interpolation methods, optical flow-based video generation methods, depth-unguided video diffusion models, and diffusion generation models with edge constraints. To maintain the same experimental conditions, the same test set, the same input resolution and the same video length were used in all methods. Three directions were evaluated: generation quality, structural preservation and temporal stability. The metrics used for the assessment of generation quality were PSNR, SSIM, LPIPS and FVD, whereas the metrics related to the preservation of the structure were subject contours, background hierarchy and occlusion relationships, while the metrics related to temporal stability were flicker analysis in consecutive frames, texture drift and motion continuity. Side-by-side comparisons can be used to reveal the significance of the depth estimation guidance mechanism for enhancing the generation of high-

quality animation.

C. ANALYSIS OF ANIMATION GENERATION QUALITY

Overall performance of different generation methods in the task of animating historical images. PSNR, SSIM, LPIPS, and FVD were selected as quantitative metrics to comprehensively compare the clarity of generated frames, structural similarity, perceived quality, and consistency with the video distribution. The results are shown in Table 3.

TABLE 3. Quantitative Comparison of Generation Quality Among Different Methods

Method	PSNR↑	SSIM↑	LPIPS↓	FVD↓
Frame Interpolation	25.84	0.812	0.183	356.42
Optical Flow Generation	26.31	0.827	0.171	328.76
Video Diffusion Model	27.48	0.852	0.142	274.63
Edge-Guided Diffusion	27.91	0.861	0.134	252.38
Proposed Method	28.76	0.883	0.108	219.54

The results demonstrate that the proposed method obtains a PSNR of 28.76, which is 1.28 higher than that of the standard video diffusion model of 27.48, and the SSIM value of the generated frames is 0.883, which is also 0.031 higher than 0.852. Meanwhile, LPIPS dropped from 0.142 to 0.108 and FVD dropped from 274.63 to 219.54, suggesting that perceptual distortion and video distribution bias were reduced with depth guidance, with more continuous and natural character boundaries, architectural lines and background textures.

D. VALIDATION OF DEPTH ESTIMATION GUIDANCE EFFECTS

To examine the constraining effect of depth information on the animation generation process, four schemes were set up: no guidance, edge guidance, semantic guidance, and depth guidance. These were compared based on metrics such as depth consistency, boundary error, LPIPS, and FVD, as shown in Table 4.

TABLE 4. Comparison of Depth Estimation Guidance Effects

Guidance Type	Depth Consistency↑	Boundary Error↓	LPIPS↓	FVD↓
No Guidance	0.741	6.83	0.142	274.63
Edge Guidance	0.768	5.92	0.134	252.38
Semantic Guidance	0.781	5.64	0.129	241.76
Depth Guidance	0.824	4.37	0.108	219.54

Table 4 shows that the Depth Consistency of the depth guidance scheme reaches 0.824, which is higher than the 0.741 of the non-guided scheme; the Boundary Error is reduced to 4.37, a decrease of approximately 36.0% compared to the 6.83 of the non-guided scheme. Meanwhile, LPIPS decreases from 0.142 to 0.108, and FVD decreases from 274.63 to 219.54. The results indicate that depth estimation not only enhances the hierarchical representation of the foreground and background but also reduces subject boundary misalignment and spatial structural distortion.

E. TEMPORAL STABILITY ANALYSIS

The experiments were evaluated using optical flow error, registration error, flicker index, and frame consistency scores, with a focus on texture discontinuities, subject drift, and background jitter between consecutive frames. The results are shown in Table 5.

TABLE 5. Comparison of Temporal Stability Metrics

Method	Optical Flow Error↓	Warping Error↓	Flicker Index↓	Frame Consistency↑
Frame Interpolation	0.086	0.071	0.124	0.802
Video Diffusion Model	0.073	0.058	0.097	0.836
Depth-Guided Diffusion	0.061	0.046	0.079	0.872
Proposed Method	0.047	0.035	0.052	0.914

As shown in Table 5, the Optical Flow Error of the proposed method is 0.047, which is lower than the 0.061 of the depth-guided diffusion model; the Warping Error decreases from 0.046 to 0.035, and the Flicker Index decreases from 0.079 to 0.052. At the same time, Frame Consistency improved to 0.914, an increase of 0.078 compared to the 0.836 achieved by a standard video diffusion model. The results indicate that temporal consistency constraints and motion compensation design can effectively improve inter-frame coherence and reduce local jitter and texture flickering.

F. ABLATION EXPERIMENT ANALYSIS

To assess the contribution of each module to model performance, we sequentially removed the historical image preprocessing, depth guidance, temporal consistency constraint, and motion compensation modules and compared the results with those of the full model to validate the effectiveness of the overall architectural design, as shown in Table 6.

TABLE 6. Comparison of Ablation Experiment Results

Model Setting	PSNR↑	SSIM↑	LPIPS↓	FVD↓
Without Preprocessing	27.94	0.861	0.128	250.41
Without Depth Guidance	27.48	0.852	0.142	274.63
Without Temporal Constraint	28.11	0.866	0.124	246.82
Without Motion Compensation	28.28	0.872	0.119	238.57
Full Model	28.76	0.883	0.108	219.54

Table 6 shows that the full model achieves a PSNR of 28.76, an SSIM of 0.883, an LPIPS of 0.108, and an FVD of 219.54, demonstrating the best overall performance. After removing the depth guidance, the FVD rose to 274.63 and the LPIPS increased to 0.142, representing the most significant performance decline, indicating that depth structure constraints are central to improving generation quality. After removing the temporal constraints, the FVD rose to 246.82; after removing motion compensation, the FVD rose to 238.57, indicating that both play a crucial role in the stability of consecutive frames.

VI. CONCLUSIONS

Historical images are static images, and the conversion of historical images into animation is a solution to the problem of dynamic display and digital preservation of static images. Considering the problems of low-resolution, no spatial hierarchy, and frame instability, we propose a video diffusion generation method with the guidance of depth

estimation. The results generated demonstrate enhanced subject contour preservation, background structure stability, and frame-to-frame coherence, achieved through collaborative design of historical image preprocessing, single-frame depth estimation, spatial structure encoding, conditional diffusion generation, and temporal consistency constraints. Experimental results demonstrate the effectiveness of this solution compared to comparison models according to PSNR, SSIM, LPIPS, FVD and frame consistency metrics, which confirms the effective constraining role of the depth information in the animation of historical images. The innovation is mainly in introducing single-frame depth structure to the video diffusion process and using motion compensation with it to minimise drift and flickering during video generation. However, there are still some limitations for detail recovery in extremely degraded images due to the quality of the historical images, the lack of ground-truth motion annotations and the occlusion scenario. Future work may consider integrating more semantic information of multiple modalities, controllable motion priors, and high-resolution temporal generation methods, to further improve the quality of animations and their adaptability to applications in complex historical scenarios.

REFERENCES

[1] W. Ma, X. Yang, L. Jiao, et al., "Video diffusion generation: comprehensive review and open problems," Artificial Intelligence Review, vol. 58, no. 11, p. 338, 2025.

[2] S. Cai, D. Ceylan, M. Gadelha, et al., "Generative rendering: Controllable 4d-guided video generation with 2d diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 7611–7620.

[3] G. Rayo and R. Tous, "Temporally Coherent Video Cartoonization for Animation Scenery Generation," Electronics, vol. 13, no. 17, p. 3462, 2024.

[4] Z. Xu, J. Zhang, J. H. Liew, et al., "Magicanimate: Temporally consistent human image animation using diffusion model," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1481–1490.

[5] M. Niu, X. Cun, X. Wang, et al., "Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model," in European conference on computer vision. Cham: Springer Nature Switzerland, 2024, pp. 111–128.

[6] A. Gupta, L. Yu, K. Sohn, et al., "Photorealistic video generation with diffusion models," in European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024, pp. 393–411.

[7] Z. Gu, R. Yan, J. Lu, et al., "Diffusion as shader: 3d-aware video diffusion for versatile video generation control," in Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers, 2025, pp. 1–12.

[8] G. Lei, C. Wang, R. Zhang, et al., "Animateanything: Consistent and controllable animation for video generation," in Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 27946–27956.

[9] Z. Kuang, S. Cai, H. He, et al., "Collaborative video diffusion: Consistent multi-video generation with camera control," Advances in Neural Information Processing Systems, vol. 37, pp. 16240–16271, 2024.

[10] S. Chen, M. Xu, J. Ren, et al., "Gentron: Diffusion transformers for image and video generation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6441–6451.

[11] J. Xing, M. Xia, Y. Zhang, et al., "Dynamicrafter: Animating open-domain images with video diffusion priors," in European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024, pp. 399–417.

[12] R. Po, W. Yifan, V. Golyanik, et al., "State of the art on diffusion models for visual computing," in Computer graphics forum, vol. 43, no. 2, 2024, Art. no. e15063.

[13] X. Chen, Z. Liu, M. Chen, et al., "Livephoto: Real image animation with text-guided motion control," in European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024, pp. 475–491.

[14] S. Shen, W. Zhao, Z. Meng, et al., "Difftalk: Crafting diffusion models for generalized audio-driven portraits animation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 1982–1991.

[15] Y. Wang, X. Chen, X. Ma, et al., "Lavie: High-quality video generation with cascaded latent diffusion models," International Journal of Computer Vision, vol. 133, no. 5, pp. 3059–3078, 2025.

[16] Y. Zhou, D. Zhou, M. M. Cheng, et al., "Storydiffusion: Consistent self-attention for long-range image and video generation," Advances in Neural Information Processing Systems, vol. 37, pp. 110315–110340, 2024.

[17] W. Xie, A. Hu, Q. Xie, et al., "Bibliometric analysis and review of AI-based video generation: research dynamics and application trends (2020–2025)," Discover Computing, vol. 28, no. 1, p. 130, 2025.

[18] Y. Wu, X. Deng, H. Shao, et al., "Anime Generation through Diffusion and Language Models: A Comprehensive Survey of Techniques and Trends," Computer Modeling in Engineering & Sciences, vol. 144, no. 3, p. 2709, 2025.

[19] Y. Jiang, C. Yu, C. Cao, et al., "Animate3d: Animating any 3d model with multi-view video diffusion," Advances in Neural Information Processing Systems, vol. 37, pp. 125879–125906, 2024.

[20] B. Li, C. Zheng, W. Zhu, et al., "Vivid-zoo: Multi-view video generation with diffusion model," Advances in Neural Information Processing Systems, vol. 37, pp. 62189–62222, 2024.

[21] W. Song, X. Wang, Y. Jiang, et al., "Expressive 3d facial animation generation based on local-to-global latent diffusion," IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 11, pp. 7397–7407, 2024.

[22] C. Ding and R. Bhowmik, "Artificial intelligence in multimedia content generation: a review of audio and video synthesis techniques," Journal of the society for information display, vol. 34, no. 2, pp. 49–67, 2026.

[23] H. Kandala, J. Gao, and J. Yang, "Pix2gif: Motion-guided diffusion for gif generation," in European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024, pp. 35–51.

[24] M. Elmoghany, R. Rossi, S. Yoon, et al., "A survey on long-video storytelling generation: architectures, consistency, and cinematic quality," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 7023–7035.

[25] Z. Fang, K. Zhu, Z. Liu, et al., "Panoramic video generation with pre-trained diffusion models," Advances in Neural Information Processing Systems, vol. 38, pp. 12486–12508, 2026.