A Controllable Video Generation Method Based on Diffusion Models and ControlNet

Qi Chen, Xinyu Guo*, Yang Liu, Wenchao Luo

Harbin Institute of Information Technology, Harbin, China

*guoxinyu-0827@qq.com

Abstract. With the rapid advancement of artificial intelligence, diffusion models have shown strong potential in video generation tasks. However, existing methods still suffer from limited controllability and inconsistent frame generation. To address these issues, this paper proposes a controllable video generation method based on diffusion models and ControlNet. The proposed approach integrates text-guided latent diffusion with structural control signals, such as pose and edge information, to achieve precise and flexible control over generated content. In addition, a temporal smoothing module is introduced to enhance inter-frame consistency. Experimental results demonstrate that the proposed method improves visual quality, semantic alignment, and structural consistency compared with baseline models. The method provides an effective and practical solution for AI-driven video generation in digital media applications.

Keywords: Diffusion Models; ControlNet; Video Generation; Controllable Generation; AIGC; Deep Learning.

1 Introduction

With the rapid development of artificial intelligence, diffusion models have demonstrated remarkable capabilities in high-quality image and video generation[1]. Recently, controllable generation has become an important research focus, especially in digital media and content creation, where precise control over generated content is essential. However, existing video generation methods often suffer from limited controllability and inconsistency across frames, which restricts their practical applications[3].

To address these issues, this paper proposes a controllable video generation method based on diffusion models and ControlNet[6]. By introducing additional conditional inputs, the proposed approach enables fine-grained control over video content while maintaining visual coherence. The method is designed to improve generation quality and flexibility for various multimedia applications.

Experimental results demonstrate that the proposed method outperforms baseline approaches in both visual quality and controllability, showing its effectiveness and potential in AI-driven video production.

R. Zur et al. (eds.), *Proceedings of the 2026 5th International Conference on Art Design and Digital Technology (ADDT 2026)*, Advances in Computer Science Research 131,
https://doi.org/10.2991/978-94-6239-737-8_9

2 Literature Review

2.1 Diffusion-based Video Generation

Diffusion models have achieved significant success in image and video generation due to their high-quality outputs and stable training process [2]. Compared with traditional generative models such as GANs, they better capture complex data distributions and produce more realistic and diverse visual content. As a result, diffusion models have been widely applied in various visual generation tasks.

Recently, these models have been extended to video generation, including text-to-video and image-to-video synthesis [4][5], enabling the creation of dynamic visual content from text or reference inputs. Latent diffusion models further improve efficiency by performing generation in a compressed latent space, making large-scale video synthesis more feasible.

However, diffusion-based video generation still faces several challenges. Many methods require substantial computational resources and large-scale datasets, limiting their practical deployment. Moreover, maintaining temporal consistency remains difficult, often resulting in flickering, motion distortion, and reduced visual quality . These issues highlight the need for improved efficiency and controllability in video generation.

2.2 AIGC in Cultural and Creative Design

To enhance controllability, conditional mechanisms have been introduced into diffusion models. ControlNet is a representative approach that incorporates structural conditions such as pose, edge, and depth [6], enabling precise spatial control and improving the flexibility of generative models. It has been widely applied in image synthesis and interactive content creation.

Recent work has extended ControlNet to video generation by introducing control signals into the generation process, improving controllability and adaptability. However, many existing approaches rely on complex architectures or additional training, increasing computational cost and implementation difficulty [7]. In addition, efficiency and scalability are often overlooked.

Therefore, there remains a need for a controllable video generation framework that is both effective and lightweight [8]. This paper addresses this gap by integrating diffusion models with ControlNet in a simple and efficient framework, making it more suitable for practical digital media applications.

3 Methodology

3.1 System Overview

As illustrated in Fig. 1, the proposed method consists of six components, including the input layer, text encoding module, control signal extraction module, diffusion model core, frame generation module, and temporal module, forming a left-to-right pipeline

for controllable video generation. Specifically, the input includes a text prompt and a reference image or video frame. The text is encoded into semantic embeddings, while structural information such as pose and edges is extracted from the reference input. These conditions are jointly fed into a ControlNet-enhanced latent diffusion model to generate video frames through an iterative denoising process. A temporal smoothing module is then applied to improve inter-frame consistency, and the refined frames are finally composed into the output video.

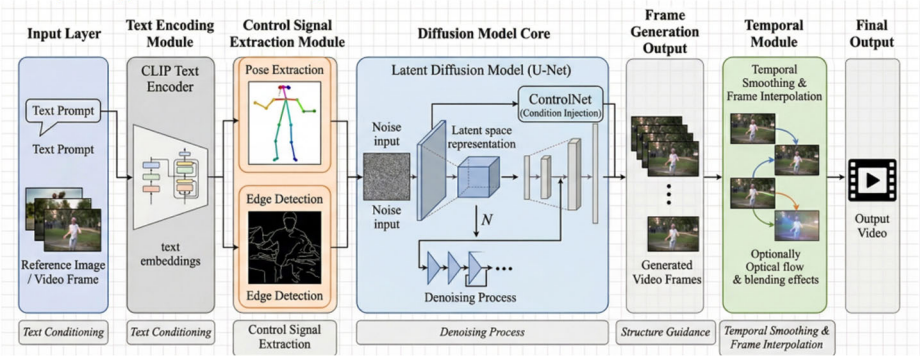

Fig. 1. Framework of the Proposed Method.

3.2 Text Encoding and Control Signal Extraction

To ensure semantic consistency, the input text prompt is encoded using a pre-trained CLIP text encoder. The encoder transforms the input text into high-dimensional embeddings, which are injected into the cross-attention layers of the diffusion model for conditional generation.

Meanwhile, structural control signals are extracted from the reference image or video frame. In this work, we employ two representative types of structural control signals:

- Pose Extraction: Human pose is detected using OpenPose, generating skeleton maps that describe body structure.
- Edge Detection: The Canny operator is applied to extract edge maps, preserving spatial contours.

These control signals provide explicit structural guidance and are used as inputs to the ControlNet branch. By combining semantic and structural information, the model achieves more precise and controllable generation.

Although this work focuses on pose and edge maps, the proposed framework is not limited to these modalities. Since ControlNet supports diverse forms of spatial conditioning, the method can be naturally extended to other structural priors, such as depth maps, semantic segmentation, and optical flow, without requiring modifications to the core architecture. This demonstrates the flexibility of the proposed approach under different control conditions.

3.3 Diffusion Model with ControlNet Integration

The core of the proposed method is a latent diffusion model based on a U-Net architecture. To improve efficiency, the generation process is conducted in the latent space rather than the pixel space. Given a latent variable x_0, Gaussian noise is gradually added during the forward process:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0,1) \tag{1}$$

The reverse process aims to predict the noise and recover the clean latent representation step by step.

To introduce controllability, ControlNet is integrated into the diffusion model as an auxiliary branch. The extracted control signals c are injected into intermediate layers of the U-Net, enabling the model to follow structural constraints during generation. The training objective is defined as:

$$\mathcal{L} = E_{x_0,\epsilon,t,c}[\|\epsilon - \epsilon_\theta(x_t, t, y, c)\|_2^2] \tag{2}$$

where y denotes the text condition and c represents the control signals. This design allows the model to simultaneously preserve semantic consistency and structural fidelity.

3.4 Frame Generation and Temporal Enhancement

After the denoising process, the model generates a sequence of video frames under consistent text and control conditions. These frames form the initial video output.

To further improve visual continuity, a temporal module is applied as a lightweight post-processing step, rather than a trainable component within the diffusion model. Specifically, adjacent frames are refined using temporal smoothing and frame interpolation techniques. In this work, we adopt an optical-flow-guided frame blending strategy to enhance inter-frame consistency:

$$I'_t = \alpha I_t + (1-\alpha) \cdot \mathrm{Warp}(I_{t-1}, F_{t-1\to t}) \tag{3}$$

where $F_{t-1\to t}$ denotes the estimated optical flow between consecutive frames, and Warp(*) represents motion alignment. This operation helps reduce temporal flickering and ensures smoother transitions across frames.

The temporal module is applied after frame generation and does not introduce additional training complexity. Its computational overhead is minimal, adding only a small constant increase to inference time.

Through this process, the generated video achieves both high visual quality and improved temporal coherence. The final frames are then composed into a complete output video sequence.

4 Experiments and Comparative Analysis

4.1 Experimental Setup

To evaluate the effectiveness of the proposed method, experiments are conducted on publicly available datasets, including UCF101 and MSR-VTT, which provide diverse scenes and motion patterns.

The method is implemented based on a pre-trained latent diffusion model and ControlNet, and all experiments are conducted on an NVIDIA RTX 3090 GPU. Unless otherwise specified, the model generates video sequences of 16 frames at a resolution of 512×512.

For quantitative evaluation, we adopt Fréchet Inception Distance (FID), CLIP Score, and Structural Similarity Index (SSIM) to measure visual quality, semantic alignment, and structural consistency, respectively. In addition, we introduce Fréchet Video Distance (FVD) to evaluate temporal consistency at the video level, where lower values indicate better video coherence. A user study is also conducted for subjective evaluation.

4.2 Main Experimental Results

To validate the effectiveness of the proposed method, comparative experiments are conducted against baseline models, including a vanilla diffusion model and a ControlNet-based image-to-video approach. The quantitative results are summarized in Table 1.

As shown in Table 1, the proposed method achieves the best performance across all evaluation metrics. It obtains the lowest FID score, indicating higher visual quality, and the highest CLIP Score, demonstrating improved semantic alignment with the input text. The SSIM value is also increased, reflecting better structural consistency across frames.

Moreover, the proposed method achieves the lowest FVD score, which indicates significantly improved temporal coherence at the video level. This demonstrates that the integration of ControlNet and the temporal smoothing module effectively reduces inter-frame inconsistency and visual flickering.

Overall, these results confirm that the proposed method improves controllability, visual quality, and temporal consistency compared with existing approaches.

Table 1. Quantitative Comparison of Different Methods for Video Generation.

Method	FID ↓	SSIM ↑	CLIP Score ↑	FVD ↓
Diffusion Model (Baseline)	32.5	0.71	0.261	312.5
ControlNet (Image-based)	28.3	0.74	0.278	278.4
Proposed Method	**24.7**	**0.79**	**0.305**	**231.7**

4.3 Discussion

The experimental results demonstrate that the proposed method improves both controllability and visual quality in video generation. By integrating ControlNet into the latent diffusion framework, the model effectively incorporates structural guidance while maintaining semantic consistency with text inputs. Compared with baseline methods, it achieves better performance across multiple evaluation metrics.

The use of control signals such as pose and edge maps provides flexible guidance and can be extended to other structural conditions, making the method suitable for diverse generation scenarios.

From an efficiency perspective, the method remains computationally intensive. The current implementation supports moderate-length videos (e.g., 16 frames at 512×512), while longer sequences can be generated using a sliding-window strategy. Higher resolutions are also feasible but incur additional computational cost, indicating a trade-off between quality, length, and efficiency.

However, the method still depends on the quality of control signals, and artifacts may appear in complex motion scenarios. Future work will focus on improving temporal modeling and enhancing efficiency for longer and higher-resolution video generation.

5 Conclusion

This paper proposes a controllable video generation method based on diffusion models and ControlNet. By integrating text-guided latent diffusion with structural control signals such as pose and edge information, the proposed approach enables precise and flexible control over generated video content. In addition, a temporal smoothing module is introduced to improve inter-frame consistency and enhance visual continuity.

Experimental results on public datasets demonstrate that the proposed method outperforms baseline approaches in terms of visual quality, semantic alignment, and structural consistency. These results verify the effectiveness and robustness of the proposed framework for AI-driven video generation tasks.

Although the method achieves promising performance, there are still limitations related to control signal accuracy and temporal consistency in complex scenarios. Future work will focus on improving temporal modeling and exploring more advanced control strategies to further enhance generation quality and efficiency.

References

1. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020. arXiv: 2006.11239
2. R. Rombach, A. Blattmann, D. Lorenz, et al., "High-Resolution Image Synthesis with Latent Diffusion Models," Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR), 2022, pp. 10684–10695. doi: 10.1109/CVPR52688.2022.01042
3. C. Saharia, W. Chan, S. Saxena, et al., "Imagen Video: High Definition Video Generation with Diffusion Models," arXiv preprint arXiv:2210.02303, 2022.
4. U. Singer, A. Polyak, T. Hayes, et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv preprint arXiv:2209.14792, 2022.
5. A. Blattmann, T. Dockhorn, S. Kulal, et al., "Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22563–22575. doi: 10.1109/CVPR52729.2023.02159
6. L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3836–3847. doi: 10.1109/ICCV51070.2023.00354
7. X. Wang, Y. Yuan, J. Zhang, et al., "VideoComposer: Compositional Video Synthesis with Motion Controllability," arXiv preprint arXiv:2306.02018, 2023.
8. D. Guo, Y. Li, H. Wang, et al., "AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning," in The Twelfth International Conference on Learning Representations (ICLR), 2024. arXiv: 2307.04725