Enhancing Video Stylization with Integral Noise

Krzysztof Ostrowski
TCL Research Europe
Warsaw, Poland
krzysztof.ostrowski@tcl.com

Michał Piasecki
TCL Research Europe
Warsaw, Poland
michal.piasecki@tcl.com

Małgorzata Kudelska
TCL Research Europe
Warsaw, Poland
malgorzata.kudelska@tcl.com

Patryk Bartkowiak
TCL Research Europe
Warsaw, Poland
patryk.bartkowiak@tcl.com

Haohong Wang
TCL Research America
San Jose, USA
haohong.wang@tcl.com

Brian Bullock
RETHINK Studios
Chicago, USA
brian@rethinkstudios.tv

***Abstract*—Temporal coherence between video frames remains a persistent challenge in video stylization. Integral noise, a method designed to preserve temporal correlations during diffusion, has been selectively employed within intra-frame correspondence mechanisms to enhance frame-to-frame consistency. However, its application to intra-frame correspondence often results in artifacts such as blurring. This paper investigates the integration of integral noise into video stylization workflows to improve temporal coherence across video frames. Furthermore, the incorporation of the IP-Adapter enables reference image-based stylization, expanding the approach's versatility. The combined methodology delivers superior outcomes, producing videos that are not only temporally coherent but also stylistically consistent and of high quality. Its effectiveness is validated in professional workflows, such as creating background animations for AI-assisted movie production, underscoring its value in advanced video editing and production applications.**

***Keywords*—temporal coherence, video stylization, integral noise, diffusion models, FRESCO framework, IP-Adapter, video editing**

I. Introduction

Video stylization presents unique challenges, primarily in maintaining temporal coherence across frames while delivering high-quality, stylistically consistent outputs. While text-to-image diffusion models have seen remarkable advancements, their extension to video remains hindered by persistent issues such as frame-to-frame flickering and inconsistent texture issues. Addressing these issues necessitates frameworks that integrate temporal constraints without sacrificing the stylistic fidelity achievable in individual frames.

The FRESCO framework [1] has demonstrated notable progress in video translation by introducing spatial-temporal correspondences. By leveraging both inter-frame and intra-frame correspondence mechanisms, FRESCO ensures improved temporal consistency while maintaining visual coherence.

In parallel, Integral Noise ($\int$-noise) [2] has emerged as an approach for additional preservation of temporal correlations during diffusion. Unlike traditional noise sampling methods that introduce artifacts such as high-frequency flickering or texture-sticking, $\int$-noise reinterprets noise samples as continuous integrated fields. This reinterpretation facilitates a noise transport mechanism that aligns temporal coherence with the structural properties of video frames, enhancing motion continuity without degrading spatial quality.

This paper proposes integrating the $\int$-noise methodology into the FRESCO framework to address the persistent challenges in video stylization. Additionally, we used IP-Adapter for reference image-based stylization. This combination bridges the gap between high-quality, stylistically versatile outputs and robust temporal consistency. We further validate the proposed approach in professional workflows, highlighting its applicability in AI-assisted film production and advanced video editing tasks.

II. Related Work

A. Image diffusion models

Diffusion models have become one of the most influential advancements in generative modeling, driving innovations across domains such as text-to-image generation, image-to-image transformation, video synthesis, and even 3D rendering. Tools like DALL-E 2 [3], Stable Diffusion (based on [5]), Midjourney [4], and Imagen [6] have made these technologies accessible to a wide audience, empowering users to create diverse visual outputs from simple text prompts.

At their core, diffusion models follow an iterative denoising process. They start from random noise and progressively refine the signal to create coherent images, guided by conditioning inputs like text descriptions. This process, first formalized through approaches like DDPM [7], has been refined by various architectural and algorithmic enhancements.

The introduction of latent space diffusion models [5], exemplified by Stable Diffusion, has greatly improved efficiency by conducting the denoising process within a compressed representation space. This innovation has lowered computational requirements while enabling broader applications, such as high-resolution image generation and real-time editing workflows.

The ability to incorporate conditioning mechanisms has greatly expanded the functionality of diffusion models. Techniques like SDEdit [8] introduced image-based guidance by injecting noise into existing images, enabling targeted transformations. Further advancements, such as Prompt2Prompt

979-8-3315-2096-0/25/$31.00 ©2025 IEEE

[9], leveraged cross-attention mechanisms to maintain layout consistency during editing, while inversion-based methods like Null-Text Inversion [10] enabled precise manipulation of real images by embedding them into the noisy latent space. Techniques like MasaCtrl [11] further enhanced self-attention layers by incorporating mutual feature interactions, achieving finer control over generated structures.

ControlNet [12] significantly advanced structural guidance by introducing a parallel control path for input conditions such as edge maps, depth cues, and pose annotations. This innovation allowed users to direct generation with unprecedented precision, opening new avenues for layout-focused editing and design applications. Similarly, methods like Pix2Pix-Zero [13] autonomously discovered editing directions in latent spaces, eliminating reliance on predefined prompts while maintaining structural fidelity.

Recent methods have explored even more versatile conditioning paradigms. IP-Adapters [14] facilitated appearance and style transfer by adapting diffusion processes to stylistic cues from reference images. Pair Diffusion [15] conceptualized images as collections of objects, allowing attribute-specific edits like texture or structural adjustments. Additionally, instructional editing frameworks like InstructPix2Pix [16] introduced fine-grained user guidance through natural language instructions, further democratizing access to complex editing capabilities.

These advancements collectively underscore the transition from generic generation to task-specific and user-driven controls in diffusion models. By integrating diverse conditioning mechanisms, diffusion frameworks now support a broader array of creative and functional applications, demonstrating their evolution as both scalable and adaptable tools for generative AI.

B. Zero-shot video editing and the FRESCO Framework

Zero-shot methods for video editing aim to achieve stylistic and structural coherence without requiring extensive pretraining or fine-tuning on specific datasets. Inversion-based methods [17], [18], utilize techniques like DDIM inversion [19] to project video frames into noisy latent spaces, enabling cross-frame attention to ensure temporal and structural coherence. While effective, these methods are computationally demanding and lack adaptability.

Inversion-free approaches, such as Text2Video-Zero [20] and ControlVideo [21], aim to overcome these challenges by directly modifying the inference process. Leveraging mechanisms like ControlNet [12] and optical flow, they enhance temporal alignment and structural consistency. Despite their flexibility, these methods often struggle with issues like flickering and spatial artifacts.

FRESCO [1] (Frame Spatial-Temporal Correspondence) was recently proposed to address these issues. It combines spatial alignment within frames with robust temporal correspondence across video sequences. Built upon inversion-free frameworks like Stable Diffusion with SDEdit [8], FRESCO combines intra-frame spatial correspondence with inter-frame temporal alignment to enhance visual coherence.

An input frame I is encoded into a latent representation $x_t = E(I)$, and Gaussian noise is added, as defined by DDPM [7]:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I\right), \tag{1}$$

where $\bar{\alpha}_t$ controls the noise schedule at timestep t. During denoising, the U-Net ϵ_t iteratively predicts the noise to translate $x'_T = x_T$ to x'_0, guided by a text prompt c and structural information such as edges, poses, or depth maps derived from I:

$$x'_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}\beta_t}}{1-\bar{\alpha}_t}\hat{x}'_0 + \frac{(1-\bar{\alpha}_{t-1})(\sqrt{\bar{\alpha}_t}x'_t + \beta_t z_t)}{1-\bar{\alpha}_t}, \tag{2}$$

where α_t and βt are pre-defined hyperparameters, z_t is a randomly sampled standard Gaussian noise, $\hat{x}'_0$ and is the predicted x'_0 at the denoising step t.

FRESCO enhances this process through two primary mechanisms. First, it optimizes decoder features $\mathbf{f} = \{f_i\}_{i=1}^N$ in the U-Net architecture by introducing spatial ($\mathcal{L}_{\text{spat}}$) and temporal ($\mathcal{L}_{\text{temp}}$) consistency losses:

$$\mathcal{L}_{\text{temp}}(\mathbf{f}) = \sum_i \left\| M_i^{i+1}\left(f_{i+1} - w_i^{i+1}(f_i)\right)\right\|_1, \tag{3}$$

where w_i^{i+1} denotes optical flow between adjacent frames and M_i^{i+1} the occlusion mask. For spatial consistency:

$$\mathcal{L}_{\text{spat}}(\mathbf{f}) = \lambda_{\text{spat}} \sum_i \left\| \tilde{f}_i \tilde{f}_i^T - \tilde{f}_i^r \left(\tilde{f}_i^r\right)^T \right\|_2^2, \tag{4}$$

where $\tilde{f}_i$ and $\tilde{f}_i^r$ are normalized feature matrices from the original and translated frames.

Second, FRESCO improves temporal and spatial coherence in video generation by utilizing advanced attention mechanisms. The FRESCO-guided attention mechanism consists of three consecutive components: spatial-guided attention, efficient cross-frame attention, and temporal-guided attention. Spatial-guided attention aggregates patches based on pre-translation similarity, maintaining spatial coherence within individual frames. Efficient cross-frame attention ensures global style consistency by referencing only the occlusion regions in adjacent frames, reducing computational redundancy while preserving newly emerged objects. Finally, temporal-guided attention aggregates features along the optical flow paths, ensuring that consistent temporal information is maintained across the video sequence.

Together, these mechanisms allow FRESCO to achieve high-quality, temporally coherent video translations while maintaining computational efficiency.

C. Integral Noise ($\int$-noise)

Maintaining temporal coherence is critical in video synthesis, where noise inconsistencies can cause flickering or texture sticking. Integral noise ($\int$-noise) [2] addresses this by aligning noise fields with motion vectors, ensuring temporal stability

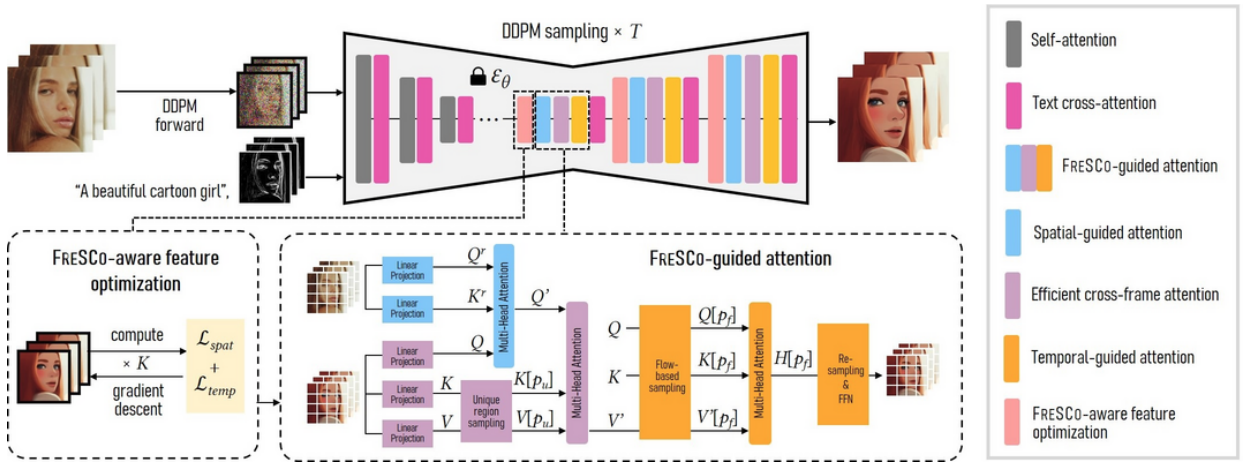

Fig. 1. FRESCO Framework (image from [1])

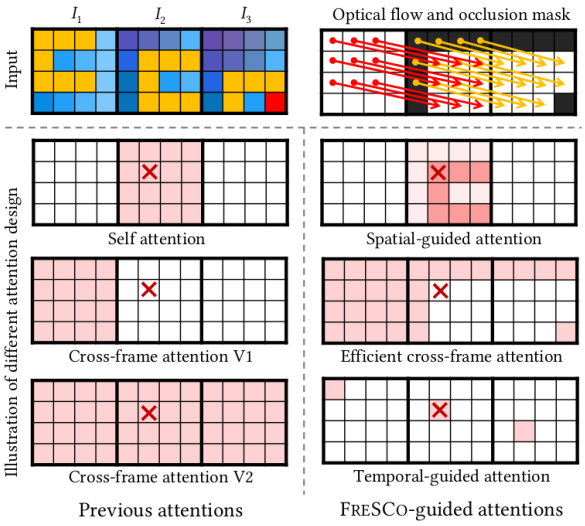

Fig. 2. Illustration of attention mechanisms from the FRESCO paper [1]

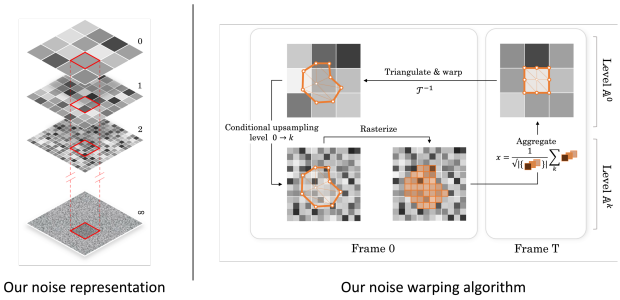

Fig. 3. llustration of Integral Noise calculation from the paper [2]

while preserving high-frequency details. The approach leverages the noise transport equation:

$$T(W)(A) = \int_{x \in A} \frac{1}{\left|\nabla T\left(T^{-1}(x)\right)\right|^{1/2}} W\left(T^{-1}(x)\right)\, dx, \quad (5)$$

where T represents a deformation field (e.g., optical flow), and $|\nabla T|$ rescales noise variance. By discretizing and warping noise via sub-pixel partitioning, $\int$-noise achieves precise noise transport, enhancing video restoration, editing, and generation by preventing artifacts like flickering and texture misalignment.

III. Methodology

To advance the capabilities of video stylization and address challenges in temporal coherence, resolution, and detail consistency, we extended the FRESCO framework by integrating $\int$-noise and additional techniques. Below, we detail the methodological enhancements and design choices.

A. Input Footage Generation

The input video frames serve as the foundational structure for the stylization process. For greater control over camera angles, lighting, and environmental conditions, the input footage was generated in Maya [22], a professional 3D animation and rendering software. This approach ensured a consistent and low-noise draft suitable for further processing.

This synthetic approach also ensured that motion vectors and spatial layout remained highly predictable, facilitating the integration of optical flow and correspondence mechanisms in subsequent processing.

B. Video Resolution

Resolution is a critical factor in achieving high-quality stylized videos. The native resolution of FRESCO's baseline model (SD 1.5) was 910x512 pixels, which fell short of industry standards for cinematic production (1920x1080). Although AI-powered up-scaling algorithms can interpolate higher resolutions, they often fail to introduce sufficient high-frequency details, such as skin textures or fine patterns.

To address this, we transitioned to SDXL, which natively operates at 1360x768 pixels for a 16:9 aspect ratio. This improved resolution provided a more suitable baseline but still fell short of full HD. During experiments, we observed

that increasing native resolution directly led to artifacts in the stylized output.

To overcome these limitations, we leveraged multiple ControlNets [12] to provide strong structural conditioning during generation. This allowed us to produce stylized outputs at higher-than-native resolutions while minimizing artifacts. By integrating edge maps and depth cues through ControlNet, the model was guided to retain structural fidelity even at larger resolutions.

C. High Frequency Details Coherence

The preservation of high-frequency details, such as textures and sharp edges, is a persistent challenge in video stylization. While FRESCO's attention mechanisms addressed global and low-frequency coherence, high-frequency details often exhibited temporal inconsistencies, particularly in areas with fine textures.

We incorporated $\int$-noise [2], a temporally correlated noise prior, into the denoising process. This method reinterprets noise samples as continuous integrated fields, ensuring consistency across frames without degrading high-frequency details. The $\int$-noise approach mitigated issues like flickering and texture-sticking by aligning noise transport with motion vectors derived from the input video.

The inclusion of $\int$-noise significantly enhanced the coherence of high-frequency details, particularly in challenging scenarios such as stylized skin textures, intricate patterns, and dynamically moving regions. This improvement was particularly evident in applications involving close-up shots or complex environmental elements.

D. Batch Length and Processing Strategy

Stylizing long videos efficiently while maintaining temporal consistency is a central capability of the FRESCO framework, achieved through its batching strategy. This approach involves overlapping frames within batches to ensure coherent style and motion propagation across sequences.

To manage memory constraints for high-resolution inputs, we limited each batch to three keyframes. The features from the first and last frames of each batch were shared with the subsequent batch, ensuring smooth transitions and temporal coherence. However, this approach inherently limited the framework to producing a single stylized frame per batch, necessitating iterative processing for longer sequences.

In line with FRESCO's design, we stylized only the selected keyframes while leveraging Ebsynth to interpolate the intermediate frames. By propagating stylistic features and structural details derived from the stylized keyframes, Ebsynth ensured temporal alignment and stylistic consistency across non-keyframes. This strategy capitalized on optical flow to accurately guide the interpolation process, maintaining fidelity to the original motion dynamics while significantly reducing the computational load.

This combination of selective keyframe stylization and efficient interpolation preserved the framework's scalability for extended sequences without compromising quality or coherence.

E. Over-blurring On Semantically Similar Parts in the Input Video

During testing, we identified over-blurring in regions where the input video contained semantically similar structures, such as repeated patterns or uniform textures. This issue arose due to FRESCO's intra-frame consistency mechanisms, which overly constrained transformations in these areas.

To address this, we disabled FRESCO's intra-frame optimization mechanism. By allowing the model greater freedom to adapt stylization dynamically, we achieved richer detail generation without compromising overall coherence.

F. Denoise Strength and Stylization Control

Stylization intensity in FRESCO [1] is controlled through SDEdit [8], a method that allows users to adjust the strength of transformations by varying the initial noise level in the denoising process.

For significant stylistic transformations, we experimented with high denoise values. With extensive conditioning from two ControlNets, we discovered that starting generation from pure noise provided better results. This approach not only maintained coherence, but also allowed for greater stylistic flexibility, particularly in scenarios requiring large deviations from the input frame appearance.

G. IP-Adapter for reference image-based stylization

Achieving precise control over the style of the output sequence is crucial to aligning artistic intent with production requirements. To this end, we integrated the Image Prompt Adapter (IP-Adapter), a powerful tool designed for reference image-based stylization. The IP-Adapter enables the model to extract stylistic cues from a reference image and apply them consistently across the generated sequence, ensuring a coherent visual aesthetic.

The use of the IP-Adapter greatly enhanced the creative flexibility of our workflow. For example, it enabled seamless integration of artistic styles such as oil painting, watercolors, or photorealism, tailored to specific production needs. Additionally, this method ensured consistency in style across complex motion sequences, avoiding common pitfalls such as style drift or localized mismatches. This capability proved particularly valuable in scenarios that require high-quality stylized outputs for cinematic sequences or narrative-driven content.

IV. Experiments and results

This section evaluates the performance of the proposed method through a series of carefully designed experiments. The experiments are structured to systematically address key aspects of the method's performance, including baseline analysis, the impact of noise initialization strategies, stylization strength, the use of multiple ControlNet mechanisms, and intra-frame feature optimization.

Section IV-A establishes a baseline by assessing the temporal coherence achieved using Integral Noise. Section IV-B explores the effects of different noise initialization strategies on maintaining temporal consistency. Section IV-C investigates the role of stylization strength in balancing temporal coherence and stylistic flexibility. Section IV-D compares the benefits of using multiple versus single ControlNet mechanisms, while Section IV-E examines the impact of intra-frame feature optimization on surface textures and detail fidelity. Together, these experiments provide a comprehensive evaluation of the method and its components.

A. Results with Integral Noise

To establish a foundation for comparative analysis, we first examine the performance of FRESCO combined with Integral Noise ($\int$-noise) in maintaining temporal coherence during stylization.

Fig. 4. Stylization pipeline on three keyframes. Top row: input frames, bottom row: stylization with Fresco and Integral noise.

Fig. 4 illustrates the stylization applied to three keyframes using Fresco and Integral Noise. The results highlight the preserved coherence across the frames, as seen in the consistent structure of the sidewalk and the shading on the wall, which will serve as a baseline for subsequent experiments.

B. Comparison of Random, Repeated Noise Priors, and Integral Noise

To evaluate the impact of different noise initialization strategies on temporal coherence, we compare the performance of random noise, repeated noise priors, and Integral Noise ($\int$-noise).

TABLE I
QUALITY METRICS FOR DIFFERENT NOISE PRIORS

Noise type	SSIM (↑)	PSNR (↑)	MSE (↓)
Random Noise	0.264	12.404	99.196
Repeated Noise	0.256	14.131	95.260
Integral Noise	**0.312**	**16.810**	**89.777**

As shown in Fig. 5, random noise introduces significant flickering and fails to maintain consistency in high-frequency details across frames. While repeated noise partially enhances temporal coherence, it falls short of achieving the desired level of stability and introduces an additional artifact: texture ”sticking” to viewport coordinates. In contrast, Integral Noise ($\int$-noise) demonstrates a substantial improvement in temporal coherence, effectively preserving high-frequency details without introducing artifacts or compromising visual quality. The quantitative metrics in Table I further highlight its strengths.

C. Stylization Strength

To assess the impact of stylization strength on temporal coherence and stylistic flexibility, we conduct an experiment varying the stylization parameter.

As illustrated in Fig. 6, using a stylization strength of 1, equivalent to generating from pure noise, does not compromise temporal coherence. Instead, this approach significantly enhances stylistic flexibility, allowing for more dramatic and diverse transformations while maintaining consistency across frames.

Depending on how much of the reference style we want to apply, a higher strength should be used. However, in practice, the optimal value will depend on the specific requirements of the film production.

D. Multiple vs. Single ControlNet

To explore the benefits of combining multiple ControlNet mechanisms, we perform an experiment comparing single and combined ControlNet usage.

Fig. 7 demonstrates that combining Depth and Mistoline ControlNets significantly enhances adherence to the input video compared to using either ControlNet independently. The complementary nature of these conditioning mechanisms ensures improved structural and stylistic alignment with the input frames.

E. Intra-Frame Feature Optimization

To evaluate the impact of intra-frame feature optimization on stylization quality, we analyze its effects on maintaining surface textures and detail fidelity.

As detailed in the Methodology, intra-frame feature optimization can lead to overblurring of originally flat surfaces. This effect, illustrated in Fig. 8, highlights how excessive optimization can unintentionally smooth areas that should retain finer details or flat textures.

V. LIMITATIONS AND FUTURE WORK

The primary limitation of the proposed approach is inherited from the vanilla FRESCO framework: it struggles to maintain temporal coherence in scenarios involving a high degree of movement. This remains a challenge, particularly in sequences with complex, rapid motion or significant scene transitions.

Looking ahead, we aim to explore the potential application of the proposed mechanisms to DiT (Diffusion Transformer) [23] based text-to-image models, such as FLUX [24], as well as emerging text-to-video models. However, text-to-video models currently face several challenges, including limited community adoption, constrained resolution, and insufficient controllability, which restrict their practical application in professional workflows at this stage.

Another promising avenue for future research involves extending Integral Noise to incorporate a 3D noise field that

Fig. 5. From left to right: comparison of Random, Repeated Noise Priors and Integral Noise. The rows show every 10th video frame, highlighting improved coherence, especially in the enlarged wall area, with Integral Noise.

Fig. 6. Comparison of different stylization strengths. From top to bottom strength: 0.5, 0.8, 1.0. The higher the value, the more style from the reference image is transferred.

spans the entire scene. This would enable temporal coherence across the entire frame sequence, rather than being limited to relationships between the first frame, previous keyframes, and the current frame. Such advancements could significantly enhance consistency and realism in video stylization, particularly for dynamic or complex scenes.

VI. Conclusions

This paper presents a novel integration of Integral Noise ($\int$-noise) with the FRESCO framework, further augmented by IP-Adapters, to address longstanding challenges in video stylization. By leveraging $\int$-noise for preserving temporal correlations during diffusion, we mitigated issues such as flickering and texture inconsistency, resulting in videos with enhanced temporal coherence. The incorporation of IP-Adapters enabled precise reference-based stylization, offering unprecedented

Fig. 7. Effect of using Depth and Mistoline Controlnets alone vs combined. From top to bottom: original frames, stylization with Depth, stylization with Mistoline, stylization with both Controlnets. Combining Depth and Mistoline ControlNets significantly enhances adherence to the input video.

Fig. 8. On the left, both intra-frame and inter-frame optimization are applied, while on the right, only inter-frame optimization is used. The left side shows more blur due to the additional intra-frame optimization.

control over stylistic elements while maintaining the structural integrity of the video frames.

Our methodology represents a significant step forward in the domain of video stylization. The combined framework achieves a balance between stylistic fidelity, temporal stability, and computational efficiency, making it highly applicable to professional workflows. This includes tasks such as creating stylized animations, cinematic background sequences, and visually consistent edits for AI-assisted movie production. Additionally, the use of advanced resolution strategies and high-frequency detail preservation techniques ensures that outputs meet the quality standards of modern filmmaking and video editing.

The practical contributions of this work extend beyond technical integration. By addressing common bottlenecks, such as intra-frame blurring, over-reliance on upscaling artifacts, and coherence in high-frequency details, this approach provides a robust solution for creative professionals seeking to stylize videos while maintaining narrative consistency.

REFERENCES

[1] S. Yang, Y. Zhou, Z. Liu and C. C. Loy, "FRESCO : Spatial-Temporal Correspondence for Zero-Shot Video Translation," *Proceedings of CVPR*, 2024.

[2] P. Chang, J. Tang, M. Gross, V. C. Azevedo, "How I Warped Your Noise: a Temporally-Correlated Noise Prior for Diffusion Models," *ICLR*, 2024.

[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125*, 2022.

[4] Midjourney, 2022. https://www.midjourney.com

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *arXiv preprint arXiv:2112.10752*, 2022.

[6] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, K. Ghasemipour, R. Lopes, B. Karagol Ayan, T. Salimans, et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," *Advances in Neural Information Processing Systems*, 2022.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, 2020.

[8] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations," *Proc. International Conference on Learning Representations*, 2021.

[9] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-Prompt Image Editing with Cross Attention Control," *Proc. International Conference on Learning Representations*, 2022.

[10] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2023.

[11] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," *arXiv preprint arXiv:2304.08465*, 2023.

[12] L. Zhang, A. Rao and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," arXiv preprint arXiv: 2302.05543, 2023.

[13] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-shot image-to-image translation," in *ACM SIGGRAPH*, pp. 1–11, 2023.

[14] H. Ye, J. Zhang, S. Liu, X. Han and W. Yang, "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models," *arXiv preprint arXiv: 2308.06721*, 2023.

[15] V. Goel, E. Peruzzo, Y. Jiang, D. Xu, N. Sebe, T. Darrell, Z. Wang, and H. Shi, "Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models," *arXiv preprint arXiv:2303.17546*, 2023.

[16] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, pp. 18392–18402, 2023.

[17] C. Qi, Xiaodong Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen, "FateZero: Fusing attentions for zero-shot text-based video editing," *In Proc. Int'l Conf. Computer Vision*, 2023.

[18] D. Ceylan, C. P. Huang, and N. J. Mitra, "Pix2video: Video editing using image diffusion," In *Proc. Int'l Conf. Computer Vision*, pages 23206–23217, 2023.

[19] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021.

[20] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators," *In Proc. Int'l Conf. Computer Vision*, 2023.

[21] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "ControlVideo: Training-free controllable text-to-video generation," *In Proc. Int'l Conf. Learning Representations*, 2024.

[22] Autodesk Maya

[23] W.S. Peebles, S. Xie, "Scalable Diffusion Models with Transformers," *IEEE/CVF International Conference on Computer Vision (ICCV)*, 4172-4182, 2022.

[24] FLUX.1 [dev], Black Forest Labs, 2024. https://blackforestlabs.ai/announcing-black-forest-labs/