Scale-Separated Feature Conditioning for Temporally Consistent Stylized Video Generation

Zach Eustace
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
zache@binghamton.edu

Cohen McCallum
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
cohen.work@ucf.edu

Abstract

Stylized video generation sits at the intersection of generative modeling, temporal representation learning, and large-scale media infrastructure. Although diffusion-based image synthesis has achieved remarkable visual fidelity and controllability, extending these capabilities to video introduces persistent challenges in temporal coherence, computational cost, and governance. This paper examines scale-separated feature conditioning as a system-level strategy for temporally consistent stylized video generation. The central argument is that conditioning style and motion at distinct spatial and temporal scales reduces the interference between content preservation, stylistic transformation, and frame-to-frame stability. Rather than treating style transfer as a single monolithic operation, a scale-separated approach assigns coarse scales to global aesthetic properties, intermediate scales to structural and semantic cues, and fine scales to texture and high-frequency detail. When coupled with temporal attention, latent propagation, and selective keyframe anchoring, this separation can improve consistency while keeping memory and latency within deployable bounds. The paper also analyzes the structural trade-offs among fidelity, controllability, throughput, and robustness in cloud, edge, and hybrid deployments. It further considers the socio-technical implications of stylized video pipelines, including provenance, content moderation, fairness, energy consumption, and lifecycle governance. By connecting architectural choices to infrastructure constraints and policy requirements, the discussion positions scale-separated feature conditioning not merely as a model design pattern but as an organizing principle for accountable, scalable, and sustainable video generation systems.

Keywords

stylized video generation; temporal consistency; scale-separated conditioning; diffusion models; socio-technical systems; governance; robustness; infrastructure.

1. Introduction

Generative video systems are rapidly moving from research demonstrations to production infrastructure, where they must satisfy visual quality, temporal stability, latency, and policy constraints simultaneously. The success of generative adversarial networks and convolutional style transfer established that neural systems can separate content from style in static images, enabling artistic transformation without explicit procedural rendering [1], [2], [3]. Subsequent work on instance normalization and adaptive normalization improved the speed and flexibility of style transfer, allowing arbitrary styles to be applied in real time [4], [5]. Image-to-image translation further generalized these ideas by learning mappings between visual domains,

including paired and unpaired settings [6], [7]. These advances provide the conceptual foundation for stylized video, but they do not resolve the temporal dimension. A style that appears coherent in a single frame may flicker, drift, or dissolve when applied independently across a sequence.

Diffusion models have since become a dominant paradigm for high-fidelity generation. Denoising diffusion probabilistic models and score-based generative modeling produce images by reversing a gradual noise process, yielding stable training and strong sample diversity [8], [9]. Latent diffusion models reduce computational cost by operating in a compressed latent space, which makes high-resolution synthesis more practical [10]. Extending diffusion to video requires the model to learn not only spatial appearance but also temporal dependencies. Video diffusion models, latent video diffusion, and structure-guided video synthesis have demonstrated that temporal attention and hierarchical generation can produce coherent motion [11], [12], [13]. Yet stylization adds a further burden: the system must preserve motion and identity while transforming appearance according to an external style signal. Methods such as one-shot tuning and zero-shot video generation have explored this space, but they often face tension between stylistic strength, temporal consistency, and computational practicality [14], [16].

This paper focuses on scale-separated feature conditioning as a way to manage that tension. The term refers to conditioning a generative process at multiple spatial and temporal scales, with different scales responsible for different aspects of style and motion. Coarse scales can carry global color, lighting, and composition; intermediate scales can carry object structure and semantic regions; fine scales can carry texture, brushwork, and high-frequency detail. When this separation is embedded in a temporally aware diffusion pipeline, it becomes possible to apply style at the scale where it is most stable and to avoid forcing fine-scale noise to satisfy long-range temporal constraints. The required reference at [15] is relevant here because it frames scale-separated conditioning as a route to style-encoder-free diffusion stylization, but the present discussion treats that work as one point in a broader system design space rather than as the sole center of analysis.

The paper contributes a system-level synthesis. It examines architecture, temporal consistency, infrastructure, deployment, robustness, fairness, sustainability, and governance. The guiding question is not only whether scale-separated feature conditioning improves perceptual metrics, but whether it produces a pipeline that can be operated responsibly at scale. That question requires attention to memory bandwidth, distributed inference, failure modes, content provenance, and the policy environments in which stylized video is generated and distributed.

2. Background and Related Work

The pre-diffusion era of neural style transfer demonstrated that deep feature statistics can encode artistic style. Gatys and colleagues showed that convolutional features from a pretrained network can be used to separate content and style, enabling artistic rendering through optimization [2]. Johnson and colleagues introduced perceptual losses for fast style transfer, making real-time applications more feasible [3]. Ulyanov and colleagues emphasized instance normalization as a key ingredient for fast stylization, while Huang and Belongie introduced adaptive instance normalization for arbitrary style transfer [4], [5]. These methods are important because they establish a multi-layer view of style: early layers capture texture and color, while deeper layers capture semantic and structural information. That multi-layer perspective is a precursor to scale-separated conditioning.

Image-to-image translation extended stylization into broader domain mapping. Conditional adversarial networks learned mappings from labels or sketches to images, while CycleGAN enabled unpaired translation between domains [6], [7]. These approaches demonstrated that generative systems can transform appearance while preserving underlying structure. However, they also revealed failure modes such as geometric distortion, semantic hallucination, and mode collapse. In video, these failures become more visible because temporal continuity exposes inconsistencies that a single frame can hide.

Diffusion models changed the quality ceiling for generative media. Denoising diffusion probabilistic models and score-based methods provide a principled way to learn complex distributions [8], [9]. Latent diffusion models then reduced the cost of high-resolution synthesis by moving the diffusion process into a learned latent space [10]. Video diffusion models extended this paradigm by adding temporal layers, factorized attention, or 3D attention to capture motion [11], [12]. Align your latents, for example, introduced high-resolution video synthesis with latent diffusion, while structure and content-guided video synthesis explored stronger control over scene composition [11], [13]. These systems are computationally intensive, but they establish the architectural vocabulary for temporally coherent generation.

Stylized video has been approached through one-shot tuning, zero-shot adaptation, and open diffusion models. Tune-a-video adapts an image diffusion model to a specific video using one-shot tuning, while Text2Video-Zero uses pretrained image diffusion models for zero-shot video generation [14], [16]. VideoCrafter1 and AnimateDiff further expand the ecosystem by providing open video generation models and animation-friendly pipelines [17], [18]. Neural radiance fields and DreamFusion are not video stylization systems per se, but they illustrate how multi-view and differentiable rendering can enforce consistency across viewpoints and time-like transformations [19], [20]. Across this literature, a common lesson emerges: temporal consistency is not a single module but an emergent property of architecture, conditioning, training data, inference strategy, and deployment constraints.

3. Scale-Separated Feature Conditioning as a System Abstraction

Scale-separated feature conditioning organizes style control across spatial and temporal resolutions. At the coarsest spatial scale, the system conditions on global attributes such as palette, lighting, and overall aesthetic mood. These attributes change slowly and can be constrained strongly across frames without suppressing motion. At intermediate scales, the system conditions on object-level structure, region boundaries, and semantic layout. Here, the risk is that aggressive style transfer may warp faces, limbs, or architectural lines. At fine scales, conditioning governs texture, edges, and high-frequency detail, which are the most likely to flicker when generated independently for each frame. By separating these responsibilities, the pipeline can apply different temporal constraints at different scales. Coarse features may be anchored over long windows, intermediate features may use motion-aware propagation, and fine features may be synthesized with shorter temporal memory but stronger spatial priors.

The required reference at [15] describes scale-separated conditioning for style-encoder-free diffusion stylization, but the architectural idea can be understood independently. In a style-encoder-free setup, the system does not rely on a separate network to embed a style image into a single global vector. Instead, style information is injected at selected scales through conditioning signals, cross-attention, or modulation layers. This reduces the representational bottleneck of a single style embedding and allows different scales to carry different aspects of

the style. It also creates a governance advantage: if style is decomposed into scales, operators can audit which scales are responsible for identity alteration, texture synthesis, or semantic drift. That auditability is valuable when stylized video is used in advertising, entertainment, journalism, or political communication.

From a systems perspective, scale separation interacts with memory hierarchy. Coarse features are small and can be cached cheaply across many frames. Intermediate features are larger and may require keyframe storage or compressed temporal summaries. Fine features are expensive and may need to be recomputed or refined only in regions where motion and occlusion demand it. This suggests a tiered inference strategy: maintain stable coarse conditioning globally, propagate intermediate structure with optical flow or feature matching, and regenerate fine texture with local temporal attention. Such a strategy can reduce redundant computation while preserving perceived consistency. It also creates new failure modes, because errors at coarse scales can propagate downward and become difficult to correct at fine scales. Governance therefore requires monitoring at each scale, not only at the output frame.

4. Temporal Consistency Mechanisms and Structural Trade-offs

Temporal consistency in stylized video depends on how the system represents motion, appearance, and uncertainty. One common approach is to add temporal attention to a diffusion model, allowing each frame to attend to neighboring frames in the latent space [11], [12]. This can produce smooth motion but increases memory consumption and can blur fast or complex movement. Another approach is latent propagation, where features from a previous frame are warped forward using estimated optical flow and then fused with newly generated features. This preserves detail but is sensitive to occlusion, disocclusion, and flow estimation errors. A third approach is keyframe anchoring, where a small number of frames are generated with high fidelity and then used as references for intermediate frames. This reduces drift but can introduce stutter if the interpolation between anchors is not carefully controlled.

Scale-separated conditioning changes the trade-offs among these mechanisms. Coarse-scale features can be propagated over long windows because they are low-dimensional and stable. Intermediate features may use a combination of attention and flow-based warping, with confidence weighting that downplays uncertain regions. Fine features can be generated with shorter temporal windows or with local patch attention that focuses on texture continuity. The result is not a single temporal consistency mechanism but a hierarchy of mechanisms matched to the stability of each scale. This hierarchy is attractive for deployment because it allows operators to trade quality for latency by reducing the temporal window at fine scales while preserving global stability at coarse scales.

However, the hierarchy also introduces coordination costs. If coarse, intermediate, and fine scales disagree about object boundaries, the output may exhibit halos, ghosting, or texture swimming. If the temporal windows are mismatched, a slowly changing coarse style may conflict with a rapidly changing fine texture, producing a perceptual uncanny effect. These failures are not merely aesthetic; they can affect trust, accessibility, and legal interpretation. For example, a stylized news video that subtly alters facial expressions or scene boundaries may raise concerns about manipulation even if no explicit false claim is made. Thus, temporal consistency must be evaluated alongside semantic fidelity and provenance.

5. Architecture, Infrastructure, and Deployment Considerations

A practical scale-separated stylized video system must be designed for heterogeneous infrastructure. In cloud deployments, the dominant constraints are GPU memory, interconnect bandwidth, storage throughput, and cost per generated second. Latent diffusion reduces the resolution at which the core generative process operates, but video still multiplies the number of tokens or latent frames that must be processed [10], [11]. A scale-separated architecture can exploit this by keeping coarse conditioning on a small set of latent frames, while distributing fine-scale refinement across multiple workers. This creates opportunities for pipeline parallelism, where different stages handle different scales, and for model parallelism, where large temporal attention modules are sharded across devices.

Edge deployment presents a different profile. On mobile devices or embedded systems, memory and energy are scarce, and network connectivity may be intermittent. A scale-separated system can degrade gracefully by running coarse and intermediate conditioning locally while offloading fine-scale diffusion to a cloud service when available. This hybrid model raises synchronization challenges: if the cloud refines texture with a different style strength or temporal window than the local coarse model, the result may flicker when the two streams are recombined. Governance must therefore include versioning of scale-specific models and conditioning parameters, so that a deployed client and server agree on the stylistic contract.

Infrastructure choices also affect robustness. Distributed inference introduces partial failures, stragglers, and inconsistent clocks. A scale-separated pipeline can tolerate some failures by falling back to cached coarse features or by skipping fine-scale refinement for a subset of frames. However, fallback behavior must be observable. If the system silently switches from high-quality to low-quality mode, users may interpret artifacts as intentional style choices or as evidence of tampering. Monitoring should track per-scale latency, cache hit rates, temporal error metrics, and style drift. These signals can feed into service-level objectives that balance user experience against cost and energy.

The broader ecosystem includes storage and content delivery networks. Stylized video may be generated once and viewed many times, so the cost of generation must be amortized against distribution. If the style is applied at coarse scales, the system may store a compact style-conditioned representation and reconstruct fine details at playback, but this increases client-side computation. If the style is baked into the video, storage and bandwidth costs increase, but playback is simpler. These trade-offs are architectural and economic, not merely algorithmic.

6. Robustness, Fairness, and Governance

Stylized video generation can amplify social risks because it transforms appearance while preserving recognizable identity and motion. Fairness concerns arise when style transfer performs unevenly across skin tones, facial features, clothing, or cultural symbols. A style that preserves detail for one demographic group may distort another, especially if training data are imbalanced. Scale-separated conditioning can help if fairness audits are conducted at each scale. For example, intermediate-scale semantic preservation can be evaluated across identity groups, while fine-scale texture synthesis can be checked for artifacts that disproportionately affect certain appearances. However, scale separation does not automatically produce fairness; it simply provides more granular points of intervention.

Content moderation and provenance are equally important. Stylized video may be used for satire, art, or entertainment, but it can also be used to obscure the origin of manipulated

content. Watermarking, signed metadata, and model fingerprinting are common technical responses, but they are not sufficient without policy and platform governance. Operators should maintain logs that record the input video, style reference, scale-specific conditioning parameters, model versions, and post-processing steps. These logs support incident response and independent audit. They also create privacy risks, because detailed logs may contain sensitive source material. Governance must therefore balance transparency with data minimization.

Policy implications extend to intellectual property, personality rights, and election integrity. Style transfer may imitate a living artist or a copyrighted visual identity, raising questions about consent and compensation. Scale-separated conditioning can make imitation more controllable, but it can also make it easier to isolate and reproduce distinctive stylistic signatures. Legal frameworks are still adapting to generative media, and technical systems should avoid assuming that a disclaimer or watermark resolves all concerns. Instead, deployment should include risk assessments that consider context of use, audience, and potential for harm.

7. Evaluation, Sustainability, and Lifecycle Governance

Evaluating stylized video requires metrics that capture spatial fidelity, temporal consistency, style strength, and semantic preservation. Per-frame perceptual metrics are insufficient because they miss flicker, drift, and texture swimming. Temporal metrics may compare optical flow between generated and reference videos, measure feature stability across frames, or use human studies to assess continuity. Scale-separated systems should be evaluated per scale, because a model may perform well globally but poorly at fine texture, or vice versa. Evaluation should also include adversarial cases such as fast motion, occlusion, transparent objects, and rapid lighting changes, where temporal consistency is most fragile.

Sustainability is a first-order concern. Video diffusion is computationally expensive, and stylization adds additional inference passes. Scale separation can reduce cost if it avoids recomputing stable coarse features, but it can also increase overhead if multiple models must be loaded and synchronized. Lifecycle governance should consider training cost, inference energy, storage, and distribution. Carbon-aware scheduling may shift non-urgent rendering to times and regions with cleaner energy, but this introduces latency and may conflict with user expectations. Efficiency improvements should be measured against quality and consistency, not in isolation, because a cheaper model that produces unstable output may generate more retries and waste.

Forward-looking perspectives suggest that scale-separated conditioning may become part of a broader trend toward modular, auditable generative systems. Instead of monolithic models, future pipelines may combine specialized modules for motion, semantics, texture, and style, each with its own governance regime. This modularity can improve robustness and fairness if modules are tested independently and composed transparently. It can also fragment accountability if no single actor understands the full pipeline. The challenge for researchers and policymakers is to design governance that matches the modularity of the technology without creating gaps in responsibility.

8. Conclusion

Scale-separated feature conditioning offers a promising systems approach to temporally consistent stylized video generation. By assigning different aspects of style and motion to different spatial and temporal scales, it reduces interference among content preservation,

stylistic transformation, and frame-to-frame stability. This separation aligns with the multi-layer nature of neural style transfer and the hierarchical structure of video diffusion models, and it creates opportunities for tiered caching, selective propagation, and graceful degradation across cloud, edge, and hybrid deployments. At the same time, scale separation is not a complete solution. It introduces coordination costs, new failure modes, and governance questions about auditability, fairness, and provenance. A responsible research agenda should therefore evaluate scale-separated systems not only by perceptual quality but also by robustness, energy use, transparency, and policy fit. The required reference at [15] provides one recent articulation of scale-separated conditioning, but the broader lesson is that temporally consistent stylized video is a socio-technical infrastructure problem. Its success will depend on architectures that are modular enough to govern, efficient enough to sustain, and robust enough to operate across diverse cultural and legal contexts.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672–2680.

2. Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2414–2423). IEEE.

3. Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision (pp. 694–711). Springer.

4. Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022.

5. Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1501–1510). IEEE.

6. Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1125–1134). IEEE.

7. Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2223–2232). IEEE.

8. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851.

9. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations.

10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684–10695). IEEE.

11. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., & Kreis, K. (2023). Align your latents: High-resolution video synthesis with latent diffusion models.

In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22563–22575). IEEE.

12. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022). Video diffusion models. Advances in Neural Information Processing Systems, 35, 8633–8646.
13. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., & Germanidis, A. (2023). Structure and content-guided video synthesis with diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7346–7356). IEEE.
14. Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., & Shou, M. Z. (2023). Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7623–7633). IEEE.
15. Zhang, Jingtao, et al. "Scale-separated conditioning for style-encoder-free diffusion stylization." arXiv preprint arXiv:2608.19719 (2026).
16. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., & Shi, H. (2023). Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 15954–15964). IEEE.
17. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., & Shan, Y. (2023). VideoCrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512.
18. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., & Dai, B. (2023). AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725.
19. Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2023). DreamFusion: Text-to-3D using 2D diffusion. In International Conference on Learning Representations.
20. Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In European Conference on Computer Vision (pp. 405–421). Springer.