Real-Time AttentionBender: Granular Interactive Network Bending of Video Diffusion Transformers

Adam Cole
a.cole@arts.ac.uk
University of the Arts London
London, UK

Mick Grierson
m.grierson@arts.ac.uk
University of the Arts London
London, UK

Figure 1: *Real-Time AttentionBender* Expressive Potential: This diverse set of samples was generated with the exact same video model and text prompt, but with varying attention bending settings. Each row shows modulations targeting a single DiT layer type (top to bottom: cross-attention, self-attention, feed-forward).

Abstract

Generative video models have achieved remarkable visual fidelity, yet their prompt-only interface offers thin creative agency and obscures the model's material process from the artists working with it. We present *Real-Time AttentionBender*, a tool that extends the practice of network bending across the full depth of the video diffusion transformer (DiT) and brings it into live, interactive generation. Built as a plugin within the DayDream Scope ecosystem and wrapping open-source real-time Wan pipelines, the tool exposes self-attention, cross-attention, and the feed-forward network as independently manipulable surfaces, with targeting down to individual diffusion steps, DiT layers, prompt tokens, and hidden neurons. The immediacy of live manipulation affords what we call *material intimacy* with the model: a responsive, near-mechanistic feel for how specific layers and neurons shape generated video. We position the tool as simultaneously an explainable AI probe into transformer internals and an expressive instrument for discovering aesthetics outside the model's default representational space.

Keywords

Video Diffusion Transformers, Real-Time Interaction, Network Bending, Explainable AI for the Arts (XAIxArts), Material Intimacy

1 Introduction

Generative video models have achieved striking quality and temporal consistency, pushing the boundaries of synthetic media production. However, to the working artist and creative coder, these massive architectures remain deeply opaque. While generative outputs are increasingly realistic, the ubiquitous prompt-only interface severely limits creative agency. Text prompts restrict artists' ability to build intuition for the model's material process, or to meaningfully intervene and work beyond the network's default representational

tendencies. If artists are to co-create *with* these systems rather than simply feed them prompts, it is essential we expose the black box of AI video models through direct, responsive interaction.

To address this, we build upon the lineage of Network Bending [1, 4]: the practice of directly manipulating the internal activations of generative models to discover novel visual languages and glitch aesthetics. We previously introduced *AttentionBender* [2], a tool that allowed artists to apply 2D transforms (rotation, scaling, translation) to the cross-attention maps of video diffusion transformers (DiTs). While effective for structural manipulation, its scope was limited to offline rendering and isolated to cross-attention layers, breaking the tight feedback loop necessary for instinctual creative exploration.

In this paper, we present *Real-Time AttentionBender*, significantly expanding the scope, depth, and interactivity of the original tool. Built as a plugin within the DayDream Scope ecosystem [6] and wrapping open-source real-time Wan pipelines [15], it delivers live network manipulation at interactive frame rates on a single GPU. To our knowledge, this work represents the first application of network bending to the full depth of the video transformer architecture, and the first time such manipulations can be explored in real-time. Specifically, this tool contributes:

(1) **Real-Time Interaction:** A live, responsive interface for modulating generative video diffusion pipelines.
(2) **Full DiT Block Modulation:** An expansion of network bending controls across all core components of the DiT block: self-attention, cross-attention, and feed-forward layers.
(3) **Granular Targeting:** The ability to isolate modulations by diffusion step, specific layer, prompt token, or down to the individual hidden neuron.

By enabling this level of fine-grained, responsive control, *Real-Time AttentionBender* functions simultaneously as an XAIxArts [5] probe into transformer attention mechanics, and as a highly expressive creative instrument for discovering novel aesthetics outside the model's learned representational space. The immediacy of live manipulation produces what we term *material intimacy* [2] with the model—a responsive, near-mechanistic feel for how specific layers and neurons contribute to the final generated video.

2 Background

2.1 Explainable AI Video for the Arts

While state-of-the-art AI video generation models (e.g., Veo [10], Sora [12], Wan [15], and Hunyuan [11]) continue to rapidly advance in both performance metrics and sample fidelity, they severely lack meaningful creative control. The dominant paradigm of prompt-based generation offers limited creative affordances, abstracting away the computational medium entirely. This opacity extends to the model itself: as artists, we do not yet possess a material understanding of how this vast collection of learned weights coalesce to produce such a wealth of representations.

To address this, we adopt the methodology of Explainable AI for the Arts (XAIxArts) [5]. Rather than treating the neural network as a black box that maps text to pixels, XAIxArts advocates for interrogating the internal mechanics of the system. This project embodies this approach by presenting an expressive and controllable interface that exists beyond the prompt box. By allowing artists to probe the system down to the level of individual neurons, we facilitate the development of a richer material intuition for AI video.

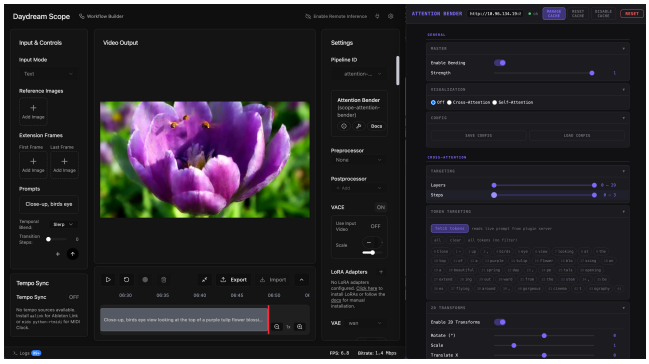

Figure 2: Real-Time *AttentionBender* Interface: The left panel displays the live generative video output, while the right panel houses interactive attention bending controls organized by self-attention, cross-attention, and the feed-forward layers

2.2 Video DiTs and Real-Time Architectures

Recent advancements in generative content are driven primarily by video diffusion models [3, 8], and increasingly, the Video Diffusion Transformer (DiT) architecture [12, 13]. For this project, we build upon Wan [15], an open-source latent [14] video DiT that has become the foundation for a diverse ecosystem of research. The Wan architecture is elegantly structured as a sequence of DiT blocks, each containing three primary components: a self-attention layer, a cross-attention layer, and a feed-forward network (FFN) with a hidden layer—30 blocks in the 1.3B model, 40 in the 14B.

Historically, diffusion models have been inherently limited in real-time settings because their bidirectional attention mechanism requires processing the entire sequence altogether, compounded by a long iterative denoising process. However, recent algorithmic breakthroughs—such as causal attention re-orientation [18], few-step diffusion distillation [17], and self-forcing [9]—have dramatically accelerated these systems. Consequently, Wan now serves as the underlying engine for numerous real-time and near real-time AI video pipelines, including CausVid [18], Self-Forcing [9], LongLive [16], and Krea [7]. This shared backbone gives us a consistent target architecture across the real-time video ecosystem.

Our work addresses a critical gap in the existing creative AI literature. While network bending has been fruitfully applied to older architectures like GANs [4] and U-Nets [1], and while the original *AttentionBender* explored offline manipulations of DiT cross-attention [2], there currently exists no framework for exploring the *entire* DiT block interactively. By wrapping these newly performant open-source real-time pipelines, our tool makes live interaction possible across every layer of the transformer, building the instinctual understanding required for an embodied XAIxArts practice.

3 Real-Time AttentionBender

3.1 System Architecture

Real-Time AttentionBender is built as a custom plugin within the DayDream Scope ecosystem. The plugin operates via a dynamic wrapper around existing real-time diffusion pipelines (currently supporting LongLive and Krea). During initialization, the plugin monkeypatches key classes within the Wan DiT blocks—specifically `WanSelfAttention`, `WanT2VCrossAttention`, and the Feed-Forward Sequential module. These patched classes are injected with global listeners that fetch user-defined modulation parameters at inference time. This architecture allows us to alter network behavior on the fly without interrupting the generative loop. In our current configuration, wrapping the LongLive pipeline at a resolution of 320×576, the system achieves a responsive ~15 frames per second on a single NVIDIA RTX A6000 Pro GPU. To further aid material intuition, the interface includes built-in live visualizers for attention maps and neuronal activations, allowing artists to directly monitor the internal state of the network as they manipulate it.

3.2 Levers of Control

The interface exposes a suite of modulations, organized around the three components of the DiT block: self-attention, cross-attention, and the feed-forward network (Fig. 2).

Self-Attention. The self-attention section exposes levers that directly influence the core attention calculation. Users can apply amplitude scaling or inject noise into the attention maps (*where* the model attends) or into the attention values (*what* information is aggregated). Additional controls allow for amplification or noising of the final attention outputs before they pass to the next layer.

Cross-Attention. Cross-attention supports the same levers as self-attention, but with two important additions. First, because cross-attention maps connect the prompt to the latent, their amplitude and noise controls can be targeted at specific prompt tokens (Section 3.3). Second, building on the methodology of the original *AttentionBender* [2], cross-attention maps can be reshaped from a flat sequence back into 3D video latents, enabling a suite of spatial-temporal modulations (rotate, scale, translate, flip, blur, and sharpen) that impose geometric and compositional constraints on the generated video, independent of the prompt.

Feed-Forward Network (FFN). We introduce novel controls for all stages of the sequential FFN: inputs, hidden layers, and outputs. For the hidden layers specifically, users can modulate the activation distributions via gain (amplification), thresholding (clamping low activations), and noise injection.

3.3 Granular Network Targeting

An essential design philosophy of this tool is enabling manipulation down to the most granular components of the network. Global modifications often destroy the latent outright; targeted intervention is what makes specific layer regions and their influence on the output legible—building material intimacy with the model.

For all levers described above, users can constrain their modulations by **diffusion step** (allowing them to target coarse structural formation early in the diffusion process versus fine detail refinement later) and by **DiT layer** (allowing them to target specific depths of the transformer). Crucially, these targeting dimensions compose: any lever in Section 3.2 can be scoped to a specific step range, layer range, and, as described below, token or neuron band.

Furthermore, we implement domain-specific targeting. In the cross-attention layers, users can target modulations by specific **prompt tokens**. For example, an artist can isolate the tokens "purple tulip" in the prompt "a purple tulip flower" and selectively modulate its amplitude in real-time to shift the color and shape of the flower without altering the surrounding composition (Fig. 3). In the feed-forward layers, we enable targeting down to **individual hidden neurons** (indices 0–8,960), allowing for localized structural and textural disruptions (Fig. 4).

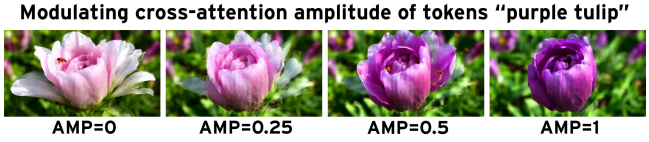

Figure 3: Prompt Token Targeting: Granular control over output features is achieved by targeting specific tokens. Here, the cross-attention amplitude for the tokens "purple tulip" within the prompt "a purple tulip flower" is modulated between 0 to 1, altering the flower's color in real-time while preserving the global composition.

4 Reflections & Future Work

4.1 Expressive Potential

Real-Time AttentionBender has proven to be a highly expressive instrument in our early explorations. As illustrated in Figure 1, manipulating the DiT block yields an expansive aesthetic territory, arguably moving beyond the model's default representational tendencies. The exact same prompt and generation settings can serve as the starting point for a surprisingly diverse range of outputs. Crucially, these results feel directly rooted in the native materiality of the neural network itself, rather than mimicking the analog or digital glitch aesthetics of previous media technologies. The images presented in this paper represent only the tip of the iceberg; the full expressive range of these combinatory levers opens rich research pathways for artists and technologists alike.

4.2 Material Intimacy as XAIxArts

Beyond its utility as a generative tool, *Real-Time AttentionBender* successfully functions as an XAIxArts probe. While prior offline network bending approaches are often limited by a cycle of parameter tweaking and waiting, bridging the gap to real-time interaction shifts the relationship we have with the tool. The ability to instantly see the impact of shifting a slider, modulating a specific layer, or targeting a token allows the artist to develop what we call a *material intimacy* with the system. It transforms the abstract "black box" of the transformer into a visible, tactile, responsive medium. The artist may begin to "feel" the structure of the network, building an instinctual, embodied understanding of how a sequence of attention blocks can develop varied visual representations.

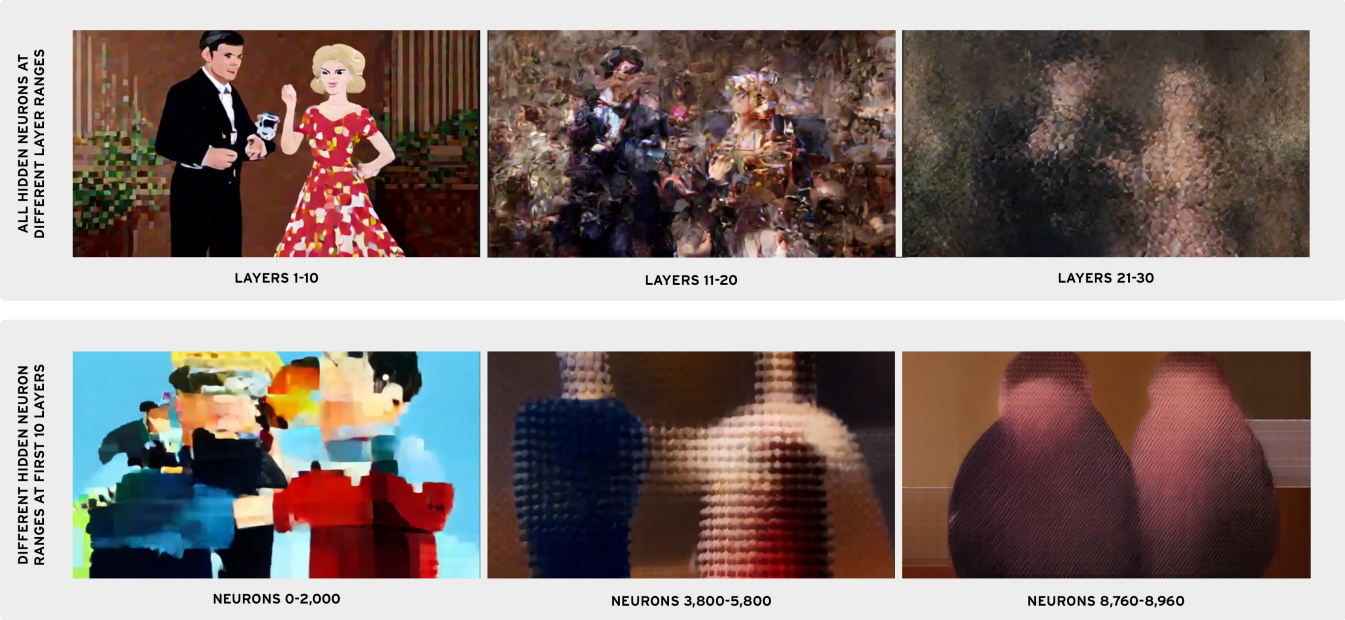

Figure 4: Granular FFN Modulation: Diverse outputs generated from identical prompts using targeted hidden neuron modulation. Top row: Injecting the same amount of noise into all hidden neurons at early (0—10), middle (10—20), and late (20—30) layers produces varied textural impacts, ranging from coarse geometry to fine, high-frequency details. Bottom row: Amplifying distinct ranges of hidden neurons (3×) within the first 10 DiT layers produces varied structural interpretations. These interventions hint at the underlying functional roles of these specific network regions.

4.3 Limitations & Future Work

The pursuit of real-time video diffusion introduces specific technical limitations that must be addressed in future iterations. Most notably, real-time architectures rely heavily on Key-Value (KV) caching to reduce redundant computation per step. However, because the KV cache stores the activation states of previously generated frames, it inherently resists sudden structural changes. If a bending parameter is altered mid-generation, the cached visual history can mute or override the new modulation, reducing responsiveness. Currently, we mitigate this by implementing manual "reset cache" and "disable cache" controls within the interface, which improves immediate explainability but artificially limits the length of coherent video generation. Developing a more sensitive, context-aware cache management system is a primary goal for future technical work.

Furthermore, while this paper establishes the technical architecture and initial expressive potential of the system, rigorous evaluation remains. In our immediate future work, we intend to conduct formal user studies with working media artists to evaluate the interface's expressive potential and observe how real-time network bending integrates into experimental creative workflows. Additionally, we plan to leverage this real-time environment to conduct quantitative mechanistic interpretability studies, systematically mapping the functional roles of specific DiT layers and neuronal bands to their corresponding visual outputs. As the barrier to entry for real-time models continues to drop, tools like *Real-Time AttentionBender* will become increasingly vital in opening up the internal mechanics of AI to a larger audience, allowing artists to actively shape the medium rather than passively being shaped by it.

References

[1] Ahmed M. Abuzuraiq and Philippe Pasquier. 2025. Explainability-in-Action: Enabling Expressive Manipulation and Tacit Understanding by Bending Diffusion Models in ComfyUI. (Aug. 10, 2025). arXiv: 2508.07183 [cs]. Pre-published.

[2] Adam Cole and Mick Grierson. 2026. AttentionBender: Manipulating Cross-Attention in Video Diffusion Transformers as a Creative Probe. In *Proceedings of the 2026 ACM Conference on Creativity and Cognition*. ACM Creativity and Cognition 2026. London, United Kingdom, (July 13, 2026). arXiv: 2604.20936 [cs]. doi:10.1145/3803784.3807565.

[3] Andreas Blattmann, Robin Rombach, et al. 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, 22563–22575.

[4] Terence Broad, Frederic Fol Leymarie, and Mick Grierson. 2021. Network Bending: Expressive Manipulation of Deep Generative Models. In *Artificial Intelligence in Music, Sound, Art and Design: 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings*. Springer-Verlag, Berlin, Heidelberg, (Apr. 7, 2021), 20–36. doi:10.1007/978-3-030-72914-1_2.

[5] Nick Bryan-Kinns, Shuoyang Jasper Zheng, et al. 2025. XAIxArts Manifesto: Explainable AI for the Arts. (Feb. 28, 2025). arXiv: 2502.21220 [cs]. Pre-published.

[6] [SW] Daydream, Scope 2026. URL: https://github.com/daydreamlive/scope.

[7] Erwann Millon. 2025. Krea Realtime 14B: Real-Time, Long-Form AI Video Generation. Krea Blog. (Oct. 15, 2025). https://www.krea.ai/blog/krea-realtime-14b.

[8] Jonathan Ho, Tim Salimans, et al. 2022. Video Diffusion Models. (June 22, 2022). arXiv: 2204.03458 [cs]. Pre-published.

[9] Xun Huang, Zhengqi Li, et al. 2025. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. (Nov. 10, 2025). arXiv: 2506.08009 [cs]. Pre-published.

[10] Tom Hume. 2025. Meet Flow: AI-powered filmmaking with Veo 3. Google: The Keyword. (May 20, 2025). https://blog.google/innovation-and-ai/products/google-flow-veo-ai-filmmaking-tool/.

[11] Weijie Kong, Qi Tian, et al. 2025. HunyuanVideo: A Systematic Framework For Large Video Generative Models. (Jan. 17, 2025). arXiv: 2412.03603 [cs]. Pre-published.
[12] OpenAI. 2024. Sora (Blogpost). (Feb. 15, 2024). https://openai.com/index/sora/.
[13] William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers (preprint). In Proceedings of the IEEE/CVF International Conference on Computer Vision. (Mar. 2, 2023), 4195–4205.
[14] Robin Rombach, Andreas Blattmann, et al. 2022. High-Resolution Image Synthesis With Latent Diffusion Models (Stable Diffusion). In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (Apr. 13, 2022), 10684–10695.
[15] WanTeam, Ang Wang, et al. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. (Mar. 26, 2025). arXiv: 2503.20314 [cs]. Pre-published.
[16] Shuai Yang, Wei Huang, et al. 2025. LongLive: Real-time Interactive Long Video Generation. (Sept. 26, 2025). arXiv: 2509.22622 [cs]. Pre-published.
[17] Tianwei Yin, Michaël Gharbi, et al. 2024. One-step Diffusion with Distribution Matching Distillation. (Oct. 4, 2024). arXiv: 2311.18828 [cs]. Pre-published.
[18] Tianwei Yin, Qiang Zhang, et al. 2025. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (July 11, 2025). arXiv: 2412.07772 [cs]. doi:10.48550/arXiv.2412.07772.