Attention of a Kiss: Exploring Attention Maps in Video Diffusion for XAIxArts

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

Abstract

This paper presents an artistic and technical investigation into the attention mechanisms of video diffusion transformers. Inspired by early video artists who manipulated analog video signals to create new visual aesthetics, this study proposes a method for extracting and visualizing cross-attention maps in generative video models. Built on the open-source Wan model, our tool provides an interpretable window into the temporal and spatial behavior of attention in text-to-video generation. Through exploratory probes and an artistic case study, we examine the potential of attention maps as both analytical tools and raw artistic material. This work contributes to the growing field of Explainable AI for the Arts (XAIxArts), inviting artists to reclaim the inner workings of AI as a creative medium.

Keywords

video art, explainable AI, generative video, attention maps, video diffusion models

ACM Reference Format:
Adam Cole and Mick Grierson. 2025. *Attention of a Kiss*: Exploring Attention Maps in Video Diffusion for XAIxArts. In *Proceedings of Explainable AI for the Arts Workshop 2025 (XAIxArts 2025).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Reviewing the history of early video art, a common trend emerges: a drive to produce imagery that disrupted the aesthetic conventions of broadcast television and cinema [9]. To achieve this, artists like Nam June Paik and the Vasulkas cultivated a deep technical understanding of video systems, enabling them to construct bespoke tools capable of manipulating analog signals in expressive, often radical ways. These interventions yielded profoundly original video works that redefined the medium. With the rise of AI video models, we ask: can a similar strategy be applied today—one that harnesses technical insight to subvert and expand the generative possibilities of these new systems?

Video diffusion models have recently demonstrated remarkable fidelity in generating realistic moving images [7]. However, while common metrics like Fréchet Video Distance (FVD) [13] focus on output quality, less focus has been given to the internal mechanics of these models. The field of machine learning interpretability seeks to address such gaps [1, 11], with the subfield of Explainable AI for the Arts (XAIxArts) [3] focusing specifically on the relevance of interpretability for artists engaging with generative systems. One established interpretability method is the visualization of attention maps in transformer based models which can be visualized for human comprehension [4, 5]. In text-to-image models like Stable Diffusion [10], cross-attention maps highlight how text tokens in the prompt correspond to regions of a visual output [6, 12]. While these methods have been explored in image models, their application to video diffusion has only recently begun from a technical perspective [8, 16], and remains particularly underexplored from an artistic one.

This project introduces a tool for visualizing cross-attention in video diffusion transformers using the open-source video model Wan2.1 [15]. The tool enables artists to inspect attention behavior across heads (parallel attention sub-units), blocks (layered model stages), and diffusion steps (iterations of gradual video refinement during generation). Such transparency offers a new vector for artistic experimentation, exposing how prompts shape generated videos and suggesting new strategies for creative intervention that go beyond the traditional outputs of AI systems.

2 Attention Maps: High Level Overview

Attention maps are a central interpretability tool for understanding how transformer models [14] operate across modalities such as language and vision [4, 5]. In the context of text-to-video transformers, attention maps are computed through the equation $\text{softmax}\left(QK^T\right)$, where Q (queries) correspond to the prompt tokens and K (keys) correspond to the embedded video representation. These attention weights indicate how much influence a given token has on specific regions of the generated output.

This mechanism allows us to trace how specific words from the prompt direct the generation of certain visual elements, effectively offering a window into the model's internal reasoning process. By extracting and visualizing these maps for every prompt token across attention heads, attention layers, and diffusion steps, we can begin to understand the generative decision-making structure and employ it as raw material for artistic exploration.

3 Method

Our method comprises two primary components: extraction and visualization of attention maps.

Extraction: We define a Python wrapper around the cross-attention layers of the Wan model. During generation, the wrapper

XAIxArts 2025, Online

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

intercepts each cross-attention computation and stores them in local memory. The final shape of these stored cross-attention maps is: [Diffusion Steps × Attn Blocks × Attn Heads × Prompt Tokens × Video Embedding].

Visualization: To visualize a stored attention map for a given prompt token, we reshape the attention map from the flat latent embedding size to a latent 3D video tensor (a lower-dimensional representation of the video across time and space, structured as [Frames × Height × Width]). These tensors are then upscaled spatially and temporally to match the shape of the output video and visualized as heatmaps. In these visual outputs, brighter colors correspond to greater attention values. Users can view maps per attention head, block, or diffusion step, or average across dimensions for higher-level overviews. The tool enables both fine-grained and global views of how individual tokens influence generation over time.

4 Results

4.1 Exploratory Probes

Exploratory probes were used to confirm the effectiveness of visualizing cross-attention maps in video diffusion models. Presented here are a selection of experimental strategies with additional results in Appendix A.

Figure 1: Comparison of attention maps from left-to-right for the tokens: "cat", "ball", and "Eiffel". The focus of attention maps neatly onto the relevant object in the scene.

Single Object: Using the prompt "a cat," we confirmed that attention maps coherently align with object regions over time. The cat token produced high attention over the visual region occupied by the cat in the generated video, confirming the interpretability of Wan's attention maps.

Multiple Objects: In a probe combining "cat," "soccer ball," and "Eiffel Tower," we observed each token's attention localized on the corresponding object, demonstrating multi-object semantic separation.

Complex Actions and Abstract Concepts: We prompted the model with more complex actions and abstract concepts like romantic scenes centered on a "classic Hollywood kiss." For the token "kiss", the attention maps were less spatially precise but still exhibited meaningful clustering around the kissing subject's lips. This revealed how more abstract concepts manifest within the model's latent space.

4.2 *Attention of a Kiss*: An Artistic Exploration

The video study *Attention of a Kiss* visualizes the evolving attention map of the "kiss" token across the generation timeline. The video begins in abstraction and gradually gains structure, paralleling both the diffusion process and the development of emotional intimacy. This metaphorical alignment—between the model's construction of meaning and human interpretive processes—suggests new narrative forms grounded in AI mechanics.

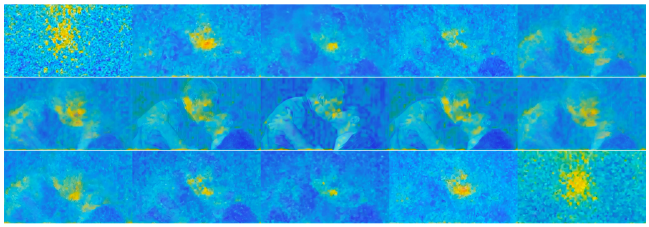

Figure 2: *Attention of a Kiss*: Video art study made from the attention maps of the token "kiss" in the open source video model Wan.

5 Discussion

5.1 Usefulness of Attention Maps for Artists

Visualizing attention maps offers artists a valuable means of understanding how their textual prompts influence the visual outputs of generative video models. By revealing which regions of an image or video frame a given token attends to over time, attention maps allow artists to "see what the model sees," making aspects of the model's internal generative process more tangible. This can help artists cultivate a more intuitive grasp of how prompts are parsed and interpreted.

As patterns emerge in how certain tokens correspond to specific visual traits, artists may begin to identify recurring motifs or dominant forms that shape the model's output. For instance, an artist might investigate how the token "woman" translates into conventional visual markers of gender. This kind of feedback loop—between creative intent, expectation, and model behavior—can deepen the artist's understanding of how language steers visual outcomes. In turn, it opens new possibilities for crafting prompts with greater intentionality and for more deliberate experimentation with the interplay between language and vision.

5.2 Limitations and Future Work

While attention maps provide a useful lens into the internal mechanics of a generative model, they also have several limitations. In some test cases, the maps were noisy, inconsistent, or visually unintelligible—especially for tokens that were abstract or ambiguous. With longer prompts, multiple tokens often attend to overlapping regions, making it difficult to isolate the influence of each one. These issues highlight the importance of interpreting attention maps with care and point to opportunities for future work to improve their utility for artists.

More broadly, attention maps reveal only a narrow aspect of the model's behavior. While they capture token-to-region relationships, they do not account for broader elements such as compositional structure, representational logic, or temporal dynamics. As such, they should be treated as interpretive tools rather than comprehensive explanations of how generative models function. On a technical level, generating and analyzing attention maps remains resource-intensive, requiring significant GPU memory and producing large

volumes of data that can be cumbersome to navigate. Future work will aim to streamline this process by improving performance and developing higher-level visualizations that preserve interpretive value without requiring inspection of low-level detail.

5.3 Conclusion: From Early Video Art Toward Network Bending

Just as early video artists built their own tools to understand and subvert the signal-based logic of analog video, artists today can gain creative leverage by exploring the inner mechanics of AI video models. Attention maps offer one such entry point—revealing how specific language tokens modulate the generation process over space and time. However, this is only the beginning.

A deeper understanding of a model's internal architecture can inspire a new generation of media practices that operate not just on the outputs of models but within their internal logic. This opens the door to network bending [2], where the structure and flow of computation within generative models are reimagined as artistic parameters.

By treating the neural network itself as a malleable medium, artists can step beyond prompt engineering to creatively intervene in the generation process, producing outputs that go beyond the intended domain of the model. These explorations extend the lineage of experimental video art into the realm of generative AI, where the artwork emerges not only from what is seen, but from how the network sees.

Acknowledgments

Adam Cole's research is supported by the UKRI Techné Studentship, AHRC Grant reference number AH/R01275X/1.

References

[1] Samira Abnar and Willem Zuidema. 2020. Quantifying Attention Flow in Transformers. doi:10.48550/arXiv.2005.00928 arXiv:2005.00928 [cs]

[2] Terence Broad, Frederic Fol Leymarie, and Mick Grierson. 2021. Network Bending: Expressive Manipulation of Deep Generative Models. doi:10.48550/arXiv.2005.12420 arXiv:2005.12420 [cs]

[3] Nick Bryan-Kinns, Shuoyang Jasper Zheng, Francisco Castro, Makayla Lewis, Jia-Rey Chang, Gabriel Viglensoni, Terence Broad, Michael Clemens, and Elizabeth Wilson. 2025. XAIxArts Manifesto: Explainable AI for the Arts. doi:10.1145/3706599.3716227 arXiv:2502.21220 [cs]

[4] Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer Interpretability Beyond Attention Visualization. doi:10.48550/arXiv.2012.09838 arXiv:2012.09838 [cs]

[5] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. doi:10.48550/arXiv.1906.04341 arXiv:1906.04341 [cs]

[6] Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. 2025. ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features. doi:10.48550/arXiv.2502.04320 arXiv:2502.04320 [cs]

[7] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022. Video Diffusion Models. doi:10.48550/arXiv.2204.03458 arXiv:2204.03458 [cs]

[8] Bingyan Liu, Chengyu Wang, Tongtong Su, Huan Ten, Jun Huang, Kailing Guo, and Kui Jia. 2025. Understanding Attention Mechanism in Video Diffusion Models. doi:10.48550/arXiv.2504.12027 arXiv:2504.12027 [cs]

[9] Chris Meigh-Andrews. 2013. *A History of Video Art.* A&C Black.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models (Stable Diffusion). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[11] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2021. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. doi:10.48550/arXiv.2103.11251 arXiv:2103.11251 [cs]

[12] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2022. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. doi:10.48550/arXiv.2210.04885 arXiv:2210.04885 [cs]

[13] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2019. Towards Accurate Generative Models of Video: A New Metric & Challenges. doi:10.48550/arXiv.1812.01717 arXiv:1812.01717 [cs]

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[15] WanTeam, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. doi:10.48550/arXiv.2503.20314 arXiv:2503.20314 [cs]

[16] Yuxin Wen, Jim Wu, Ajay Jain, Tom Goldstein, and Ashwinee Panda. 2025. Analysis of Attention in Video Diffusion Transformers. doi:10.48550/arXiv.2504.10317 arXiv:2504.10317 [cs]

A Media Files

(1) ***Attention of a Kiss***: https://youtu.be/dFay2ko8dmk

B Supplementary Results

The following tests provide a deeper look into the multi-object scene study, generated with the following settings:

```
prompt = "cinematic video of a cat playing with a soccer ball in front of the Eiffel Tower, realistic, 8k, high quality,
        masterpiece, best quality"
negative_prompt = "Bright tones, overexposed, static, blurred details, subtitles, style, works, paintings, illustration,
        images, overall gray, worst quality, low quality, JPEG compression residue, ugly, incomplete"
seed = 58
guidance_scale = 6
height = 480
width  = 832
num_frames = 61
num_inference_steps = 25
```

B.1 Attention Developing Across Diffusion Steps

The image grids in Figure 3 show composites of the cross-attention maps for the prompt token `"cat"`. Within each grid, every cell represents a different transformer block in the model (30 in total for Wan2.1 1.3B). The three grids represent, from left to right: the first diffusion step, the middle diffusion step, and the final diffusion step.

These visualizations offer some exploratory qualitative insights:

(1) Over successive diffusion steps, attention for prompt-referenced objects becomes more focused and distinct.
(2) Within a single diffusion step, attention appears to begin diffusely, tighten around object regions, and then broaden slightly again.

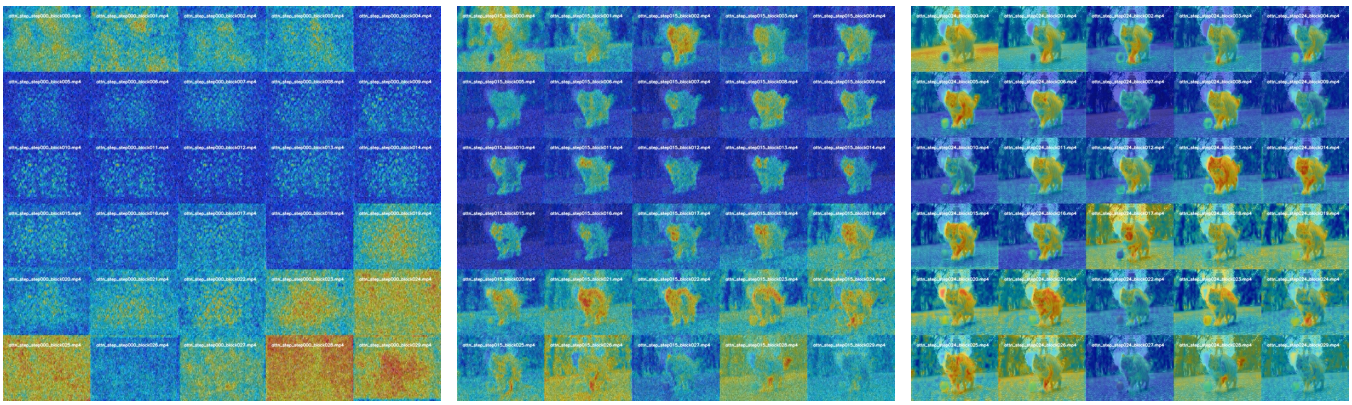

Figure 3: Cross-attention maps for the token `"cat"` across all 30 transformer blocks. Grids from left to right correspond to the first, middle, and final diffusion steps.

B.2 Attention in a Specific Transformer Block

As noted above, attention often consolidates around object regions in the middle transformer blocks. In Figure 4, we visualize cross-attention for the token `"cat"` at block 15 (the middle of 30 blocks) across all 25 diffusion steps. The grid on the left shows the first frame of each step; the grid on the right shows the final frame. The image on the far left displays the intermediate diffusion output at step 6 (corresponding to row 2, column 1).

Exploratory qualitative inferences include:

(1) Cross-attention for the token `"cat"` is remarkably sharp across frames at this middle block.
(2) The composition appears to be established early in the diffusion process—though the cat is not clearly visible to human eyes in the noisy output at step 6, it is already distinguishable in the corresponding attention maps.

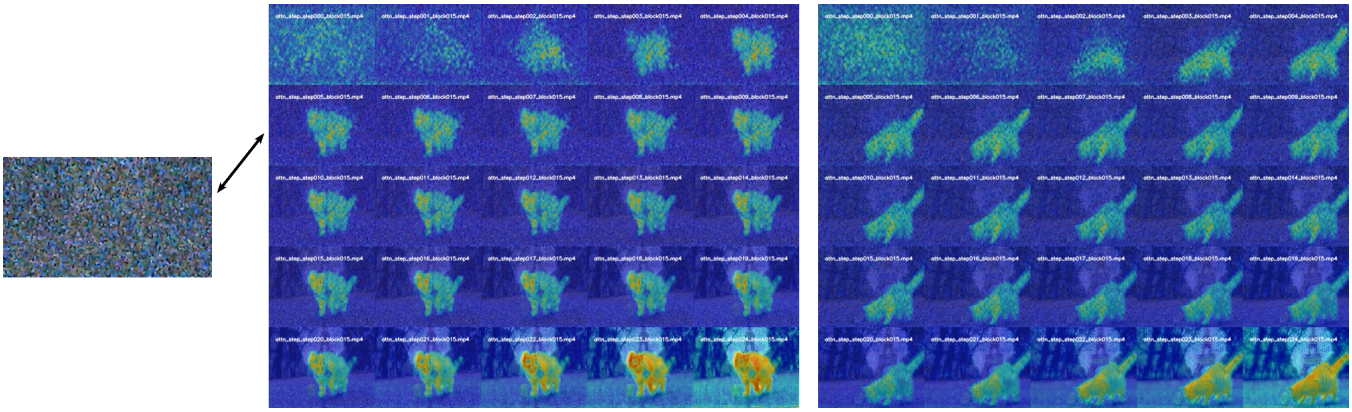

Figure 4: Cross-attention for the token "cat" in transformer block 15 across all diffusion steps. Left-grid corresponds to the first video frame; Right-grid corresponds to the last video frame. The far-left image is the intermediate output at diffusion step 6. While visually indecipherable for human eyes, the attention maps show that the system infers quite a lot of detail by this step.

B.3 Attention in Individual Attention Heads

Figure 5 visualizes every attention head within transformer block 15 across all 25 diffusion steps. The Wan2.1 model has 12 attention heads per block. This is the finest-grain view of cross-attention behavior available during generation.

While this visualization may be too dense for casual inspection, it reveals several notable patterns:

(1) Individual attention heads show consistent behaviors across steps. For example, head 7 consistently attends to the outline of the cat, while head 10 exhibits a more global pattern.
(2) Some heads appear significantly more responsive or discriminative than others.

B.4 Averaged Attention Across Blocks and Steps

Figure 6 shows a high-level overview of average cross-attention for the token "cat", aggregated across all heads, blocks, and diffusion steps. While this provides a broad sense of where the token attends, it lacks the temporal and architectural nuance shown in previous visualizations.

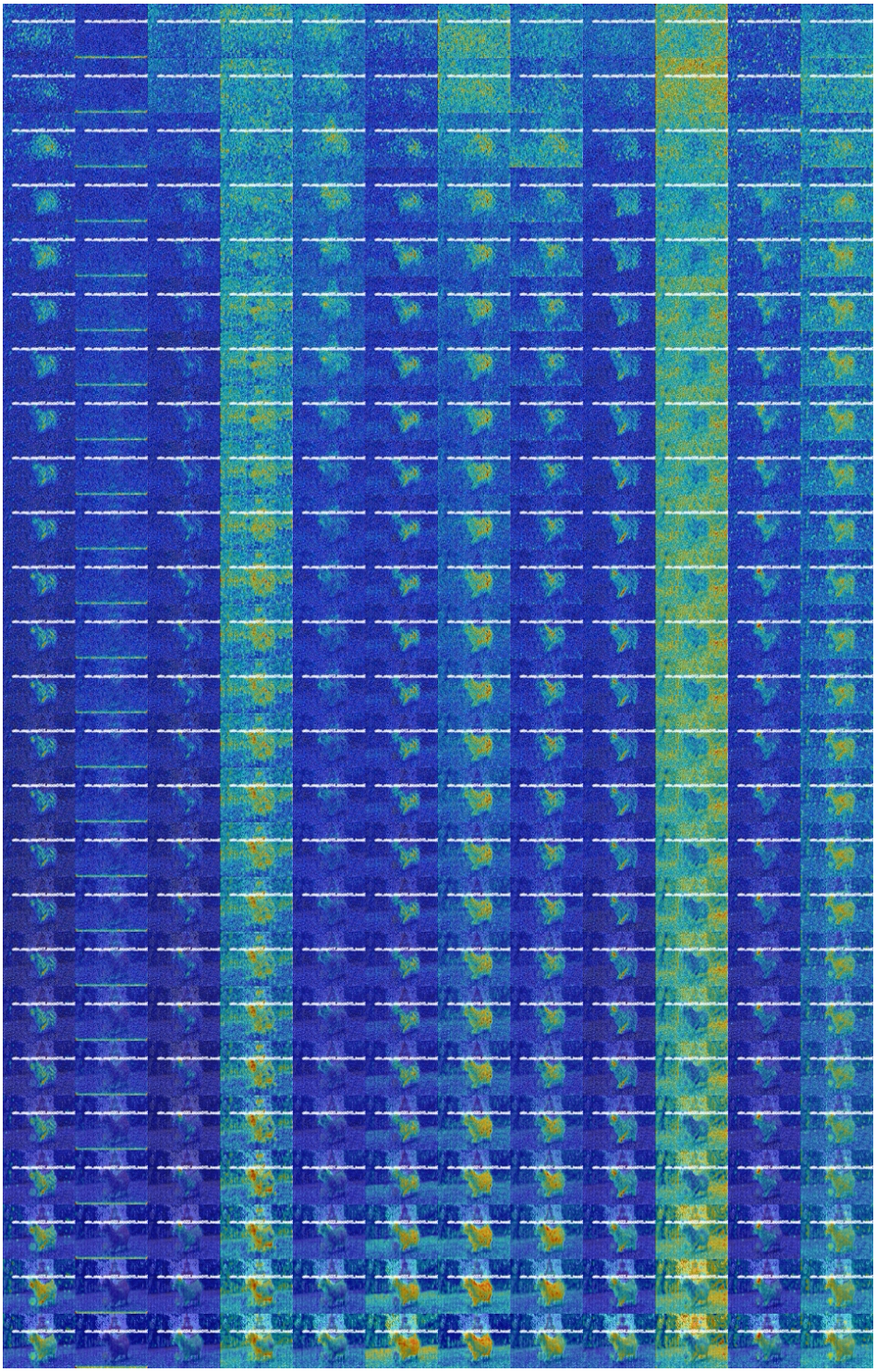

Figure 5: Cross-attention for the token "cat" across all 12 attention heads in block 15, visualized over 25 diffusion steps. Columns represent attention heads; rows represent diffusion steps.

Figure 6: Cross-attention map for "`cat`" averaged across all heads, blocks, and diffusion steps.