ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features

Anonymous CVPR submission

Paper ID 38

Abstract

Do the rich representations of multi-modal diffusion transformers (DiTs) exhibit unique properties that enhance their interpretability? We introduce CONCEPTATTENTION*, a novel method that leverages the expressive power of DiT attention layers to generate high-quality saliency maps that precisely locate textual concepts within images. Without requiring additional training,* CONCEPTATTENTION *repurposes the parameters of DiT attention layers to produce highly contextualized concept embeddings, contributing the major discovery that performing linear projections in the output space of DiT attention layers yields significantly sharper saliency maps compared to commonly used cross-attention mechanisms. Remarkably,* CONCEPTATTENTION *even achieves state-of-the-art performance on zero-shot image segmentation benchmarks, outperforming 15 other zero-shot interpretability methods on the ImageNet-Segmentation dataset and on a single-class subset of PascalVOC. Our work contributes the first evidence that the representations of multi-modal DiT models like Flux are highly transferable to vision tasks like segmentation, even outperforming multi-modal foundation models like CLIP.*

1. Introduction

Diffusion models have recently gained widespread popularity, emerging as the state-of-the-art approach for a variety of generative tasks, particularly text-to-image synthesis [21]. These models transform random noise into photorealistic images guided by textual descriptions, achieving unprecedented fidelity and detail. Despite the impressive generative capabilities of diffusion models, our understanding of their internal mechanisms remains limited. Diffusion models operate as black boxes, where the relationships between input prompts and generated outputs are visible, but the decision-making processes that connect them are hidden from human understanding.

Existing work on interpreting T2I models has predominantly focused on UNet-based architectures [19, 21], which utilize shallow cross-attention mechanisms between prompt embeddings and image patch representations. UNet *cross attention maps* can produce high-fidelity saliency maps that predict the location of textual concepts [27] and have found numerous applications in tasks like image editing [5, 12]. However, the interpretability of more recent multi-modal diffusion transformers (DiTs) remains underexplored. DiT-based models have recently replaced UNets [22] as the state-of-the-art architecture for image generation, with models such as Flux [13] and SD3 [8] achieving breakthroughs in text-to-image generation. The rapid advancement and enhanced capabilities of DiT-based models highlight the critical importance of methods that improve their interpretability, transparency, and safety.

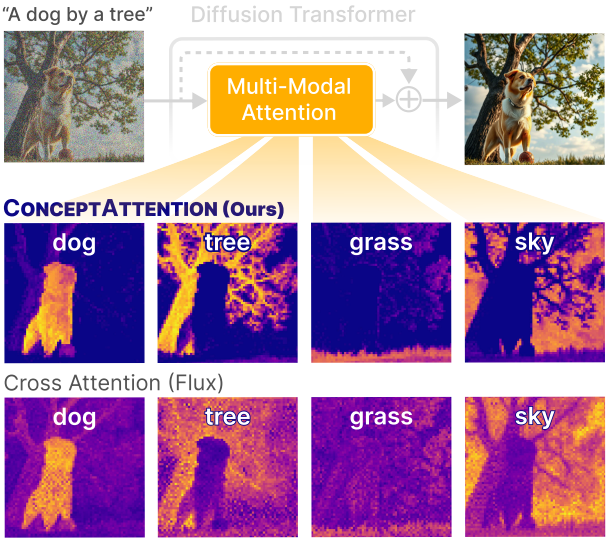

Figure 1. **CONCEPTATTENTION interprets the representations of multi-modal diffusion transformers by producing high-quality saliency maps of textual concepts.** We compare to the activations of the cross attention mechanisms of the DiT.

In this work, we propose CONCEPTATTENTION, a novel method that leverages the representations of multi-modal DiTs to produce high-fidelity saliency maps that localize textual concepts within images. Our method provides insight into the rich semantics of DiT representations. CONCEPTATTENTION is lightweight and requires no additional

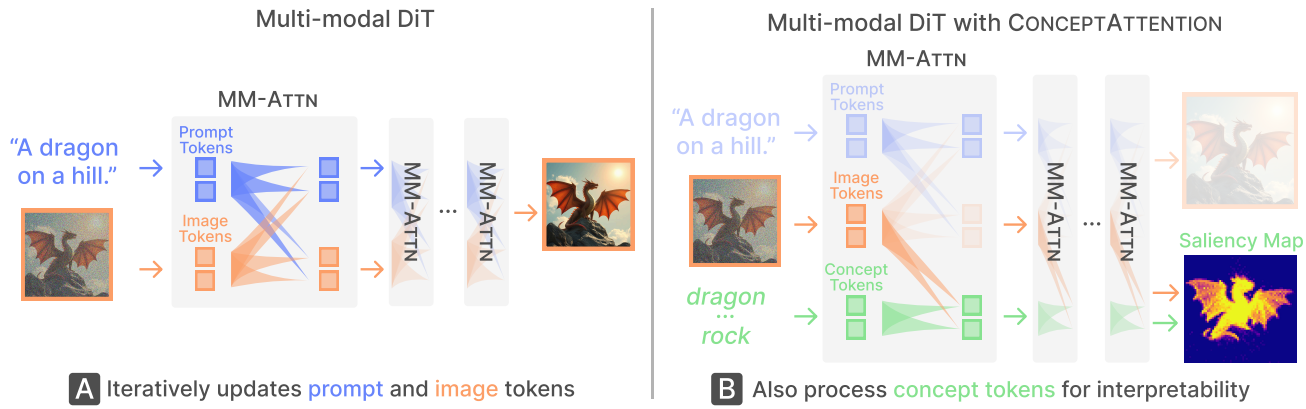

Figure 2. **ConceptAttention augments multi-modal DiTs with a sequence of concept embeddings that can be used to produce saliency maps.** (Left) An unmodified multi-modal attention (MMAttn) layer processes both **prompt** and **image** tokens. (Right) ConceptAttention augments these layers without impacting the image appearance to create a set of contextualized **concept** tokens.

training, instead it repurposes the existing parameters of DiT attention layers. ConceptAttention works by producing a set of rich contextualized text embeddings each corresponding to visual concepts (e.g. "dragon", "sun"). By linearly projecting these *concept embeddings* and the image we can produce rich saliency maps that are even higher quality than commonly used cross attention maps.

We evaluate the efficacy of ConceptAttention in a zero-shot semantic segmentation task on real world images. We compare our interpretative maps against annotated segmentations to measure the accuracy and relevance of the attributions generated by our method. Our experiments and extensive comparisons demonstrate that ConceptAttention provides valuable insights into the inner workings of these otherwise complex black-box models. By explaining the meaning of the representations of generative models our method paves the way for advancements in interpretability, controllability, and trust in generative AI systems.

In summary, we contribute:

- **ConceptAttention, a method for interpreting text-to-image diffusion transformers.** Our method requires no additional training, by leveraging the representations of multi-modal DiTs to generate highly interpretable saliency maps that depict the presence of arbitrary textual concepts (e.g. "dragon", "sky", etc.) in images (as shown in Figure 1).
- **The novel discovery that the output vectors of attention operations produce higher-quality saliency maps than cross attentions.** ConceptAttention repurposes the parameters of DiT attention layers to produce a set of rich textual embeddings corresponding to different concepts, something that is uniquely enabled by multi-modal DiT architectures. By performing linear projections between these *concept embeddings* and image patch representations in the attention output space we can produce high quality saliency maps.
- **ConceptAttention generalizes to achieve state-of-the-art performance in zero-shot segmentation on benchmarks like ImageNet Segmentation and Pascal VOC.** We achieve superior performance to a diverse set of zero-shot interpretability methods based on various foundation models like CLIP, DINO, and UNet-based diffusion models; this highlights the potential for the representations of DiTs to transfer to important downstream vision tasks like segmentation.

2. Preliminaries

2.1. The Anatomy of a Multi-modal DiT Layer

Multi-modal DiTs like Flux and Stable Diffusion 3 leverage *multi-modal attention layers (*MMAttn*)* that process a combination of textual tokens and image patches. There are two key classes of layers: one that keeps separate residual streams for each modality and one that uses a single stream. In this work, we take advantage of the properties of these dual stream layers, which we refer to as multi-modal attention layers (MMAttns).

The input to a given layer is a sequence of image patch representations $x \in \mathbb{R}^{h \times w \times d}$ and prompt token embeddings $p \in \mathbb{R}^{l \times d}$. The initial prompt embeddings at the beginning of the network are formed by taking the T5 [20] embeddings of the prompt tokens.

Following [18], each MMAttn layer leverages a set of adaptive layer norm *modulation layers* [28], conditioned on the time-step and global CLIP vector. An adaptive layer-norm operation is applied to the input image and text embeddings. The final modulated outputs are then residually added back to the original input. Notably, the image and text modalities are kept in separate residual streams. The exact details of this operation are omitted for brevity.

The key workhorse in MMAttn layers is the familiar multi-head self attention operation. The prompt and im-

Figure 3. **CONCEPTATTENTION generates saliency maps for multiple concepts simultaneously, even concepts not in the prompt.**

age embeddings have separate learned key, value, and query projection matrices which we refer to as K_x, Q_x, V_x for images and K_p, Q_p, V_p for text. The keys, queries, and values for both modalities are collectively denoted q_{xp}, k_{xp}, and v_{xp}, where for example $k_{xp} = [K_x x_1, \ldots, K_p p_1 \ldots]$. A self attention operation is then performed

$$o_x, o_p = \text{softmax}(q_{xp} k_{xp}^T) v_{xp} \tag{1}$$

Here we refer to o_x and o_p as the *attention output* vectors. Another linear layer is then applied to these outputs and added to a separate residual streams weighted according to the output of the modulation layer. This gives us updated embeddings x^{L+1} and p^{L+1} which are given as input to the next layer.

3. Proposed Method: CONCEPTATTENTION

We introduce CONCEPTATTENTION, a method for generating high quality saliency maps depicting the location of textual concepts in images. CONCEPTATTENTION works by creating a set of contextualized *concept embeddings* for simple textual concepts (e.g. "cat", "sky", etc.). These concept embeddings are sequentially updated alongside the text and image embeddings, so they match the structure that each MMATTN layer expects. However, unlike the text prompt, concept embeddings do not impact the appearance of the image. We can produce high-fidelity saliency maps by projecting image patch representations onto the concept embeddings. CONCEPTATTENTION requires no additional training and has minimal impact on model latency and memory footprint. A high level depiction of our methodology is shown in Figure 2.

Using CONCEPTATTENTION. The user specifies a set of r single token concepts, like "cat", "sky", etc. which are passed through a T5 encoder to produce an initial embedding c^0 for each concept. For each MMATTN layer (indexed by L) we layer-normalize the input concept embeddings c^L and repurpose the text prompt's projection matrices (i.e. K_p, Q_p, V_p), to produce a set of keys, values, and queries (i.e. $k_c = [K_p c_1, \ldots K_p c_k]$).

One-directional Attention Operation. We would like to update our concept embeddings so they are compatible with subsequent layers, but also prevent them from impacting the image tokens. Let k_x and v_x be the keys and values of the image patches x respectively. We can concatenate the image and concept keys to get

$$k_{xc} = [K_x x_1 \ldots, K_x x_n, K_p c_1 \ldots, K_p c_r] \tag{2}$$

and the image and concept values to get

$$v_{xc} = [V_x x_1 \ldots, V_x x_n, V_p c_1 \ldots, V_p c_r] \tag{3}$$

We can then perform the following attention operation

$$o_c = \text{softmax}(q_c k_{xc}^T) v_{xc} \tag{4}$$

which produces a set of *concept output embeddings*.

Notice, that instead of just performing a cross attention (i.e. $\text{softmax}(q_c k_x^T) v_x$) our approach leverages both cross attention from the image patches to the concepts and self attention among the concepts. We found that performing both operations improves performance on downstream evaluation tasks like segmentation (See Table 5). We hypothesize this is because it allows the concept embeddings to repel from each other, avoiding redundancy between concepts. Meanwhile, the image patch and prompt tokens ignore the concept tokens and attend only to each other as in

$$o_x, o_p = \text{softmax}(q_{xp} k_{xp}^T) v_{xp}. \tag{5}$$

A diagram of these operations is shown in Fig. 9 (b) in the Appendix.

A Concept Residual Stream. The above operations create a residual stream of concept embeddings distinct from the image and patch embeddings. Following the pretrained transformer's design, after the MMATTN we apply another projection matrix P and MLP, adding the result residually to c^L. We apply an adaptive layernorm at the end of the attention operation which outputs several values: a scale γ, shift β, and some gating values α_1 and α_2. The residual stream is then updated as

$$c^{L+1} \leftarrow c^L + \alpha_1 (P o_c) \tag{6}$$

$$c^{L+1} \leftarrow c^{L+1} + \alpha_2 \,\text{MLP}\left((1+\gamma)\,\text{lnorm}(c^{L+1}) + \beta\right) \tag{7}$$

where $\leftarrow$ denotes assignment. The parameters from each of our modulation, projection, and MLP layers are the same as those used to process the text prompt.

Saliency Maps in the Attention Output Space. These concept embeddings can be combined with the image patch embeddings to produce saliency maps for each layer L. Specifically, we found that taking a simple dot-product similarity between the image output vectors o_x and concept output vectors o_c produces high-quality saliency maps

Method	ImageNet-Seg Acc	ImageNet-Seg mIoU	ImageNet-Seg mAP	PascalVOC Acc	PascalVOC mIoU	PascalVOC mAP
CLIP SA [7]	67.84	46.37	80.24	68.51	44.81	83.63
TextSpan [10]	75.21	54.50	81.61	75.00	56.24	84.79
TransInterp [4]	79.70	61.95	86.03	76.90	57.08	86.74
CLIPasRNN [26]	74.05	58.80	84.80	61.76	41.48	76.57
DINOv2 SA [16]	77.39	63.12	84.19	79.65	57.61	87.26
DAAM [27]	78.47	64.56	88.79	72.76	55.95	88.34
Cross Attn (SD3.5)	77.80	63.67	83.50	80.22	61.46	86.97
Cross Attn (Flux)	74.92	59.90	87.23	80.37	54.77	89.08
Ours (SD3.5)	81.92	67.47	**90.79**	83.90	69.93	90.02
Ours (Flux)	**83.07**	**71.04**	90.45	**87.85**	**76.45**	**90.19**

Table 1. CONCEPTATTENTION outperforms a variety of Diffusion, DINO, and CLIP interpretability methods on ImageNet-Segmentation and PascalVOC (Single Class). An extended version of this table with additional baselines is shown in Table 2 of the Appendix.

$$\phi(o_x, o_c) = \text{softmax}(o_x o_c^T). \quad (8)$$

This is in contrast to cross attention maps which are between the image keys k_x and prompt queries q_p.

We can aggregate the information from multiple layers by averaging them $\frac{1}{|L|}\sum_{L=1}^{|L|}\phi(o_x^L, o_c^L)$ where $|L|$ denotes the number of MMATTN layers (Flux has $|L| = 19$). These attention output space maps are unique to MM-DiT models as they leverage *concept embeddings* corresponding to textual concepts which fundamentally can not be produced by UNet-based models.

4. Experiments

We are interested in investigating (1) the efficacy of CONCEPTATTENTION to generate highly localized and semantically meaningful saliency maps, and (2) understand the transferability of multi-modal DiT representations to important downstream vision tasks. Zero-shot image segmentation is a natural choice for achieving these goals.

Datasets. For our key evaluation we leverage two zero-shot-image segmentation datasets: ImageNet-Segmentation [11] which was introduced for this task in [4, 10]. The dataset contains 4,276 images from 455 classes. Each image depicts a single central object or subject, which makes it a good method for comparing to a variety of interpretability methods that only generate a single saliency map, not a class specific one. For the second dataset we leverage PascalVOC 2012 benchmark [9]. We investigate both a single class (930 images) and multi-class split (1,449 images) of this dataset. We also evaluate on a multi-class segmentation task in Appendix A.1.

Experimental Details. For our first task we closely follow the established evaluation framework from [10] and [4]. We perform this evaluation setup on both ImageNet-Segmentation and a subset of 930 PascalVOC images containing only a single class. For each method we assume the class present in the image is known and use simplified descriptions of each ImageNet class (e.g. "Maltese dog" → "dog) this allows the concepts to be captured by a single token. We construct a concept vocabulary for each image composed of this target class and a set of fixed background concepts for all examples (e.g. "background", "grass", "sky").

Baselines. We compare our approach to a variety of zero-shot interpretability methods which leverage several different multi-modal foundation models. We omit numerous models from Table 1 due to space constraints, but we have an extended version with additional baselines in Table 2 of the Appendix. We compare to numerous CLIP interpretability methods [1, 2, 4, 7, 10, 24, 26], the self-attentions of various DINO models [3, 6, 16], and approaches that leverage the cross attentions of UNet-based diffusion models [14, 27], and the cross attention maps of Flux and SD3.5 Turbo [23].

Implementation Details. For our experiments we implemented ConceptAttention for both the distilled Flux-Schnell model and Stable Diffusion 3.5 Turbo [23] in PyTorch [17]. We encode real images with the DiT by first mapping them to the VAE latent space and then adding varying degrees of Gaussian noise before passing them through the DiT. We then cache all of the concept output o_c and image output vectors o_x from each MMATTN layer. At the end of generation we then construct our concept saliency maps for each layer and average them over all layers of interest. In our experiments we leverage the activations from the last 10 of the 19 MMATTN layers.

Quantitative Metrics. Each method produces a set of scalar *raw scores* for each image patch which we then threshold based on the mean value to produce a binary segmentation prediction. We compare each method using standard segmentation evaluation metrics, namely: mean Intersection over Union (mIoU), patch/pixelwise accuracy (Acc), and mean Average Precision (mAP). Accuracy alone is an insufficient metric as our dataset is highly imbalanced, mIoU is significantly better, and mAP captures threshold agnostic segmentation capability. We found that CONCEPTATTENTION significantly out performs all of the baselines we tested across all three metrics (Tab. 1). This is true for diffusion, CLIP, and DINO based interpretability methods.

Qualitative Evaluation. We also show qualitative results comparing the segmentation performance to each baseline in Figures 1, 3 and in Appendix B.

Multi Object Image Segmentation. We also performed a quantitative evaluation for images with multiple objects, with details outlined in Appendix A.1.

Ablation Studies We performed various architectural ablation studies shown in Appendix A.2

References

[1] Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers, 2020. arXiv:2005.00928 [cs]. 4, 7

[2] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers, 2016. arXiv:1604.00825 [cs]. 4, 7

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, 2021. arXiv:2104.14294 [cs]. 4, 7

[4] Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability Beyond Attention Visualization, 2021. arXiv:2012.09838 [cs]. 4, 7

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Transactions on Graphics*, 42(4):148:1–148:10, 2023. 1

[6] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers, 2024. arXiv:2309.16588 [cs]. 4, 7

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. arXiv:2010.11929 [cs]. 4, 7

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, 2024. arXiv:2403.03206. 1

[9] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 4

[10] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting CLIP's Image Representation via Text-Based Decomposition, 2024. arXiv:2310.05916 [cs]. 4, 7

[11] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. ImageNet Auto-Annotation with Segmentation Propagation. *International Journal of Computer Vision*, 110(3):328–348, 2014. 4

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross Attention Control, 2022. arXiv:2208.01626 [cs]. 1

[13] Black Forest Labs. FLUX, 2023. 1

[14] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary Object Segmentation with Diffusion Models, 2023. arXiv:2301.05221 [cs]. 4

[15] Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C. SanMiguel, and Jose M. Martínez. Open-Vocabulary Attention Maps with Token Optimization for Semantic Segmentation in Diffusion Models, 2024. arXiv:2403.14291 [cs]. 7

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2024. arXiv:2304.07193 [cs]. 4, 7

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, 2019. arXiv:1912.01703 [cs]. 4

[18] William Peebles and Saining Xie. Scalable Diffusion Models with Transformers, 2023. arXiv:2212.09748. 2

[19] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, 2023. arXiv:2307.01952 [cs]. 1

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2023. arXiv:1910.10683 [cs]. 2

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, 2022. arXiv:2112.10752 [cs]. 1

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015. arXiv:1505.04597. 1

[23] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation, 2024. arXiv:2403.12015 [cs]. 4

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2019. arXiv:1610.02391. 4

[25] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. arXiv:1610.02391 [cs]. 7

[26] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. CLIP as RNN: Segment Countless Visual Concepts without Training Endeavor, 2024. arXiv:2312.07661 [cs]. 4, 7

[27] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin,

and Ferhan Ture. What the DAAM: Interpreting Stable Diffusion Using Cross Attention, 2022. arXiv:2210.04885 [cs]. 1, 4, 7

[28] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and Improving Layer Normalization, 2019. arXiv:1911.07013 [cs]. 2

[29] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer, 2025. arXiv:2408.06072 [cs]. 13, 14

[30] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer, 2022. arXiv:2111.07832 [cs]. 7

Method	Architecture	ImageNet-Segmentation Acc ↑	mIoU↑	mAP↑	PascalVOC (Single Class) Acc ↑	mIoU↑	mAP↑
LRP [2]	CLIP ViT	51.09	32.89	55.68	48.77	31.44	52.89
Partial-LRP [2]	CLIP ViT	76.31	57.94	84.67	71.52	51.39	84.86
Rollout [1]	CLIP ViT	73.54	55.42	84.76	69.81	51.26	85.34
ViT Attention [7]	CLIP ViT	67.84	46.37	80.24	68.51	44.81	83.63
GradCAM [25]	CLIP ViT	64.44	40.82	71.60	70.44	44.90	76.80
TextSpan [10]	CLIP ViT	75.21	54.50	81.61	75.00	56.24	84.79
TransInterp [4]	CLIP ViT	79.70	61.95	86.03	76.90	57.08	86.74
CLIPasRNN [26]	CLIP ViT	74.05	58.80	84.80	61.76	41.48	76.57
OVAM [15]	SDXL UNet	79.41	65.02	88.12	73.50	58.12	87.91
DINO SA [3]	DINO ViT	81.97	69.44	86.12	80.71	64.33	88.90
DINOv2 SA [16]	DINOv2 ViT	77.39	63.12	84.19	79.65	57.61	87.26
DINOv2 Reg SA [6]	DINOv2 Reg	72.04	56.31	80.83	77.16	56.60	86.35
iBOT SA [30]	iBOT ViT	76.34	61.73	82.04	74.96	55.80	85.26
DAAM [27]	SDXL UNet	78.47	64.56	88.79	72.76	55.95	88.34
DAAM [27]	SD2 UNet	64.52	47.62	78.01	64.28	45.01	83.04
Cross Attention	Flux DiT	74.92	59.90	87.23	80.37	54.77	89.08
Cross Attention	SD3.5 DiT	77.80	63.67	83.50	80.22	61.46	86.97
CONCEPTATTENTION	SD3.5 DiT	81.92	67.47	**90.79**	83.90	69.93	90.02
CONCEPTATTENTION	Flux DiT	**83.07**	**71.04**	90.45	**87.85**	**76.45**	**90.19**

Table 2. CONCEPTATTENTION outperforms a variety of Diffusion, DINO, and CLIP ViT interpretability methods on ImageNet-Segmentation and PascalVOC (Single Class). An extended version of this table with additional baselines is shown in Appendix **??**.

A. Additional Quantitative Results

A.1. Multi-object Semantic Segmentation

We also wanted to evaluate the capabilities of our method at differentiating between multiple classes in an image. However, only a subset of methods produce distinct saliency maps for open ended classes. For this experiment we compare to DAAM using a SDXL backbone, TextSpan using a CLIP backbone, and the raw cross attentions of Flux. Instead of binarizing the image to produce segmentations, for each patch we predict the concept with the highest score. We used pixelwise accuracy and mIoU as our evaluation metrics and found that our method significantly outperformed the baselines (See Table 3). We also show qualitative results of our approach differentiating between multiple concepts in a single image in Figures 1, **??** and we show more results in Appendix B.

A.2. Ablation Studies

We perform several ablation studies to investigate the impact of various architectural choices and hyperparameters on the performance of CONCEPTATTENTION.

Impact of Layer Depth on Segmentation We hypothesized that deeper MMATTN layers in the DiT would have more refined representations that better transfer to segmentation. This was confirmed by our evaluation (see Figure 4). We pull features from each diffusion layer and evaluated the segmentation performance of these features on ImageNet Segmentation. We also compare the performance of combining all layers simultaneously, which we found performs better than any individual layer.

Impact of Diffusion Timestep on Segmentation We add Gaussian noise to encoded images before passing them to the DiTs, this conforms with the expectations of the models. Intuitively one might expect the later timesteps (less noise) to have much higher segmentation performance as less information about the original image is corrupted. However, we found that the middle diffusion timesteps best (See **??**). Throughout the rest of our experiments we use timestep 500 out of 1000 following this result.

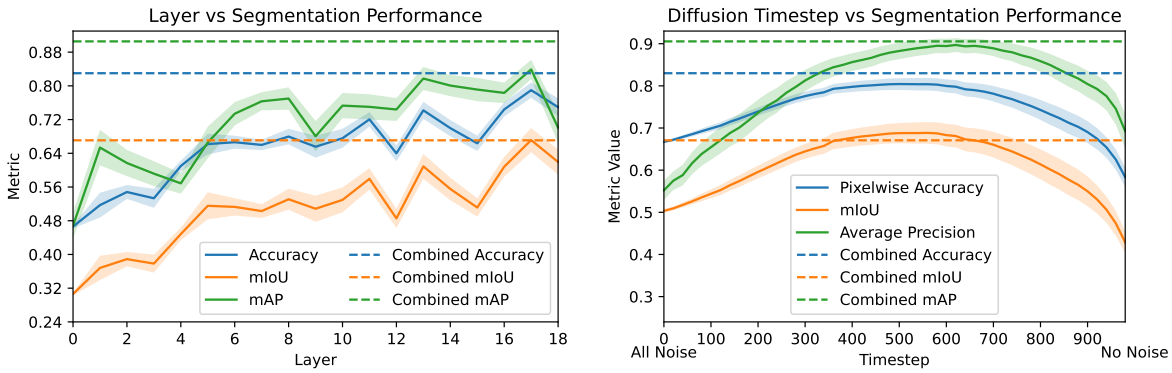

Figure 4. **(Left) Later MMAttn layers encode richer features for zero-shot segmentation.** We investigated the impact of using features from various MMAttn layers and found that deeper layers lead to better performance on segmentation metrics like pixelwise accuracy, mIoU, and mAP. We also found that combining the information from all layers further improves performance. **(Right) Optimal segmentation performance requires some noise to be present in the image.** We evaluated the performance of ConceptAttention by encoding samples from a variety of timesteps (determines the amount of noise). Interestingly, we found that the optimal amount of noise was not zero, but in the middle to later stages of the noise schedule.

Method	**Acc↑**	**mIoU↑**
TextSpan	73.84	38.10
DAAM	62.89	10.97
Flux Cross Attention	79.52	27.04
ConceptAttention	**86.99**	**51.39**

Table 3. **ConceptAttention outperforms alternative methods on images with multiple classes from PascalVOC.** Notably, the margin between ConceptAttention and other methods is even higher for this task than when a single class is in each image.

Space	**Softmax**	**Acc↑**	**mIoU↑**	**mAP↑**
CA		66.59	49.91	73.17
CA	✓	74.92	59.90	87.23
Value		45.93	29.81	65.79
Value	✓	45.78	29.68	39.61
Output		78.75	64.95	88.39
Output	✓	**83.07**	**71.04**	**90.45**

Table 4. **The output space of DiT attention layers produces more transferable representations than cross attentions.** We explore the transferability of several representation spaces of a DiT: the cross attentions (CA), the value space, and the output space. We performed linear projections of the image patches and concept vectors in each of these spaces and evaluated their performance on ImageNet-Segmentation.

Concept Attention Operation Ablations We compared the performance on the ImageNet Segmentation benchmark of performing (a) just cross attention from the image patches to the concept vectors, (b) just self attention, (c) no attention operations, and (d) both cross and self attention. Our results seen in Table 5 indicate that using a combination of both cross and self attention achieves the best performance.

We also investigated the impact of applying a pixelwise softmax operation over our predicted segmentation coefficients. We found that it slightly improves segmentation performance in the attention output space and significantly improves performance for the cross attention maps (see Table 4.

B. Additional Qualitative Results

Here we show a variety of qualitative results for our method that we could not fit into the original paper.

CA	SA	Acc↑	mIoU↑	mAP↑
		52.63	35.72	70.21
	✓	51.68	34.85	69.36
✓		76.51	61.96	86.73
✓	✓	**83.07**	**71.04**	**90.45**

Table 5. **CONCEPTATTENTION performs best when we utilize both cross and self attention.** We tested the effectiveness of performing just a cross attention operation between the concepts and image tokens, just a self attention among the concepts, both cross and self attention, and neither. We found that doing both operations leads to the best results. Metrics are computed on the ImageNet Segmentation benchmark.

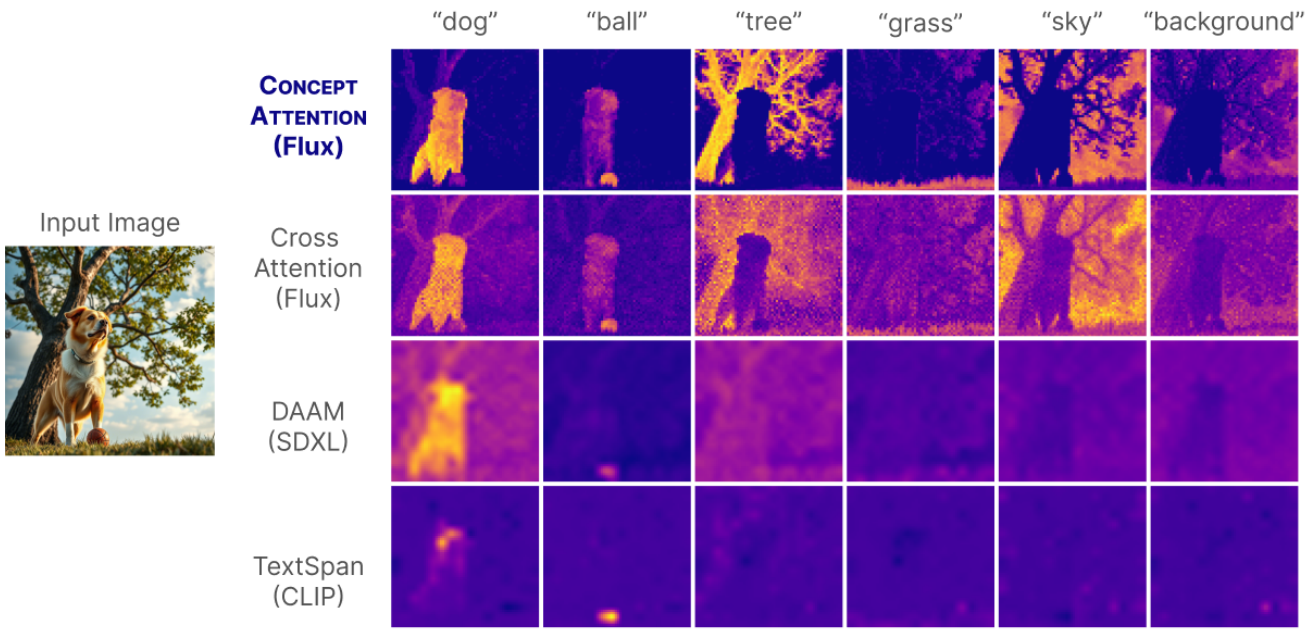

Figure 5. A qualitative comparison between our method and several others.

C. Pseudo-code for the CONCEPTATTENTION Algorithm

We show pseudo-code depicting the difference between a vanilla multi-modal attention mechanism and a multi-modal attention mechanism with concept attention added to it.

D. Concept Attention on Video Generation Models

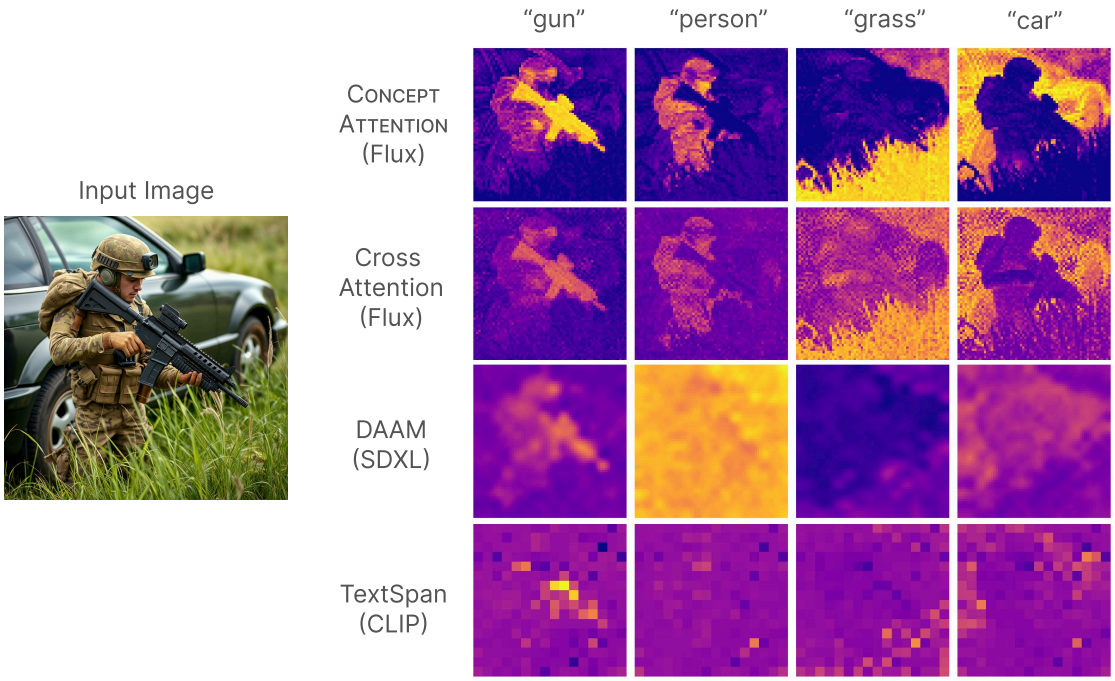

Figure 6. A qualitative comparison between our method and several others.

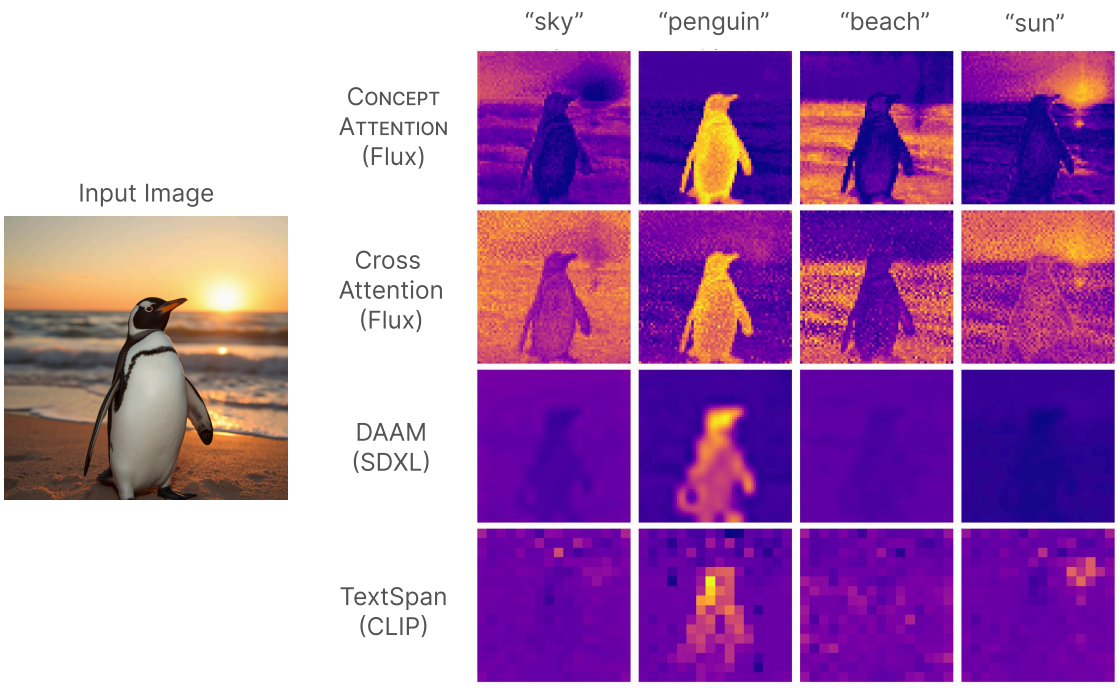

Figure 7. A qualitative comparison between our method and several others.

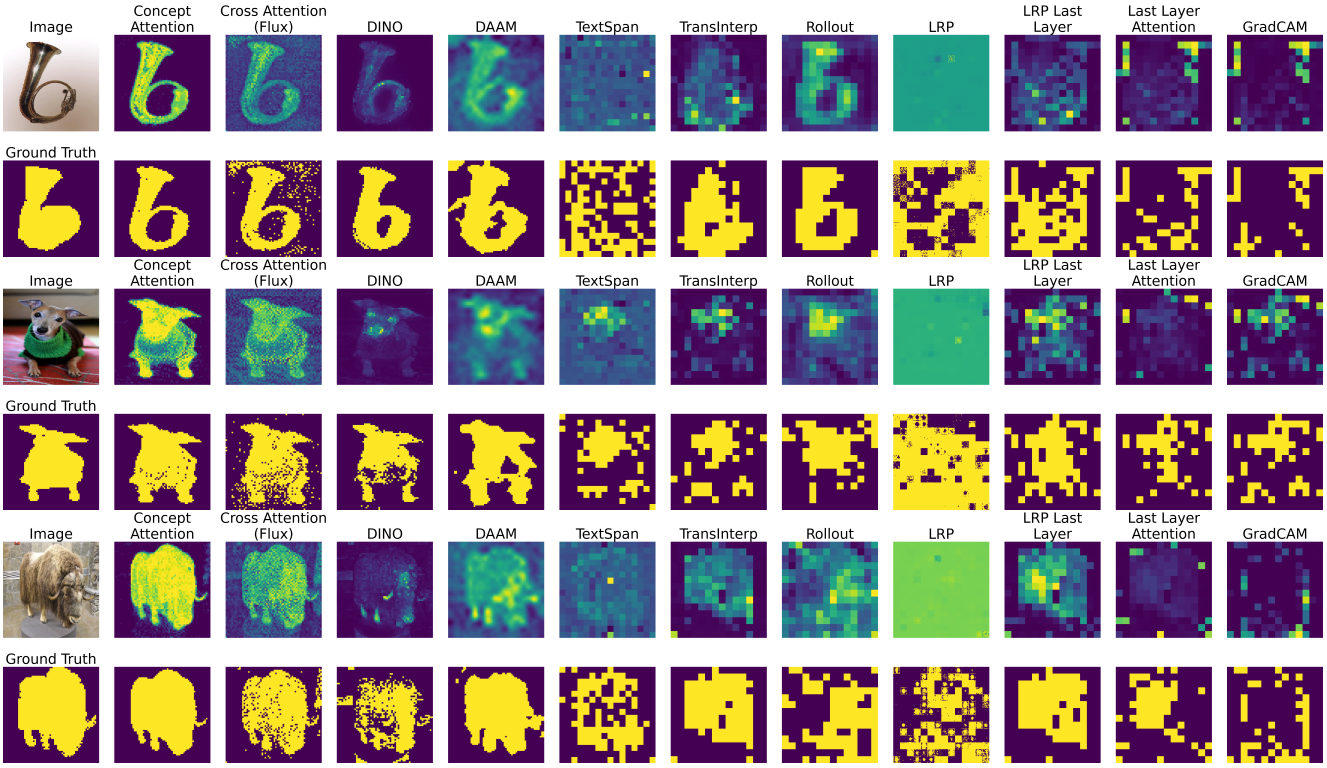

Figure 8. Qualitative comparisons between numerous baselines on ImageNet Segmentation Images. We show the soft predictions for each method and binarized segmentation predictions.

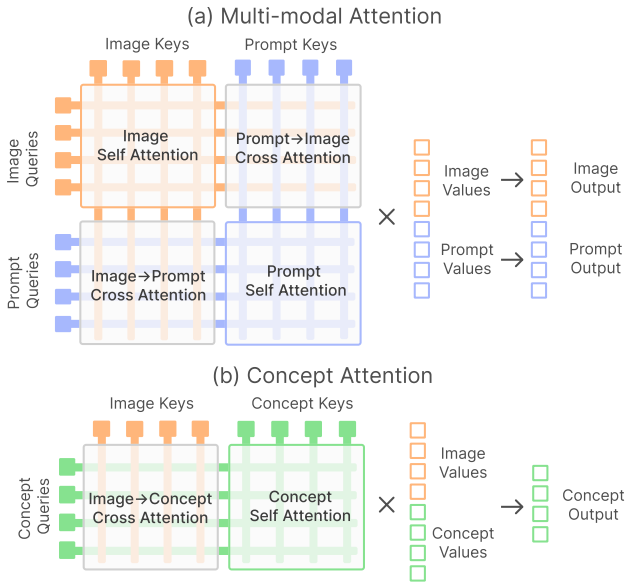

Figure 9. (a) MMATTN combines cross and self attention operations between the prompt and image tokens. (b) Our CONCEPTATTENTION allows the concept tokens to incorporate information from other concept tokens and the image tokens, but not the other way around.

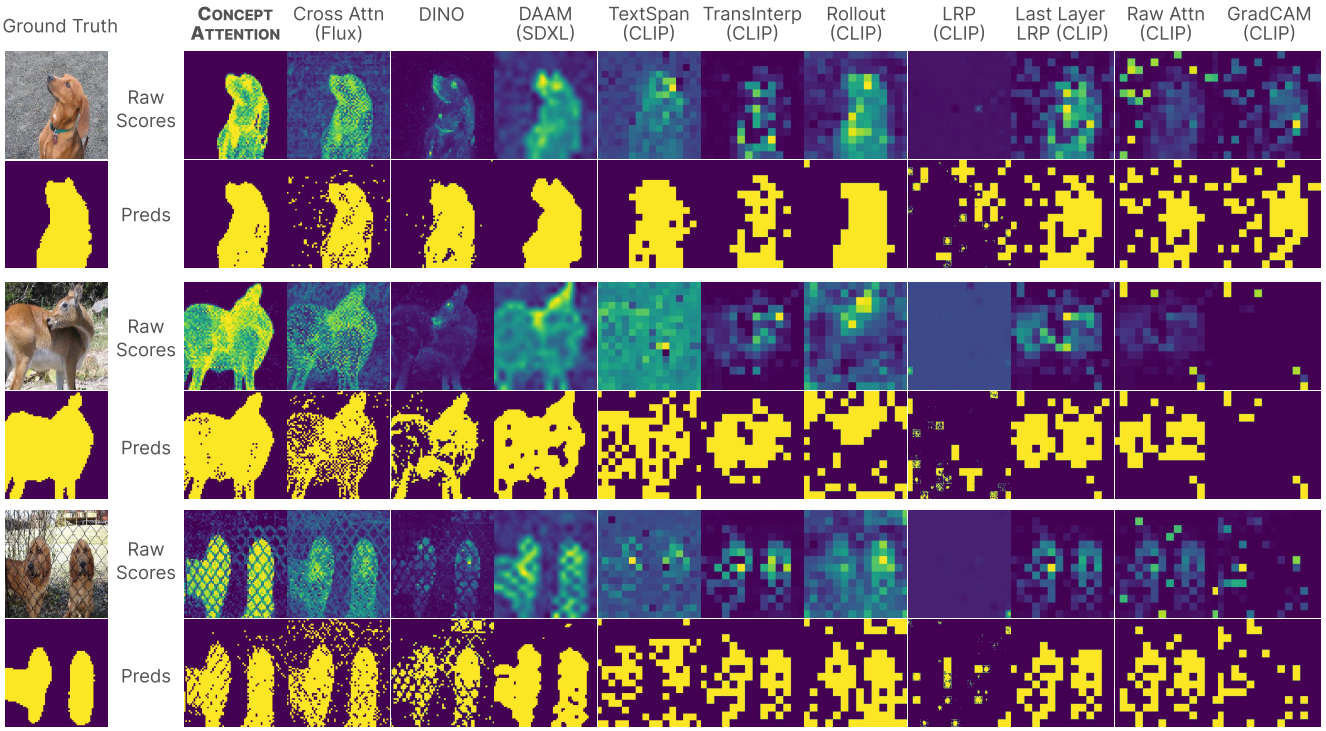

Figure 10. CONCEPTATTENTION produces higher fidelity raw scores and saliency maps than alternative methods, sometimes surpassing in quality even the ground truth saliency map provided by the ImageNet-Segmentation task. Top row shows the soft predictions of each method and the bottom shows the binarized predictions.

(a) Multi-Modal Attention

```
def multi_modal_attn(img, txt):
    # Compute the keys, queries, and values
    img_k, img_q, img_v = img_projection(img)
    txt_k, txt_q, txt_v = txt_projection(txt)

    # Concat the image and text keys, queries, and vals
    img_txt_k = concat([img_k, txt_k])
    img_txt_q = concat([img_q, txt_q])
    img_txt_v = concat([img_v, txt_v])
    # Perform self attention on combined sequence
    attn_out = self_attention(img_txt_k, img_txt_q, img_txt_v)
    # Unpack the attention outputs
    img = attn_out[:img.shape[0]], attn_out[img.shape[0]:]

    return img, txt
```

(b) Multi-modal Attention with Concept Attention

```
+ def multi_modal_attn_with_concept_attn(img, txt, concepts):
      # Compute the keys, queries, and values
      img_k, img_q, img_v = img_projection(img)
      txt_k, txt_q, txt_v = txt_projection(txt)
+     concept_k, concept_q, concept_v = txt_projection(concepts)
      # Concat the image and text keys, queries, and vals
      img_txt_k = concat([img_k, txt_k])
      img_txt_q = concat([img_q, txt_q])
      img_txt_v = concat([img_v, txt_v])
      # Perform self attention on combined sequence
      attn_out = self_attention(img_txt_k, img_txt_q, img_txt_v)
      # Unpack the attention outputs
      img, txt = attn_out[:img.shape[0]], attn_out[img.shape[0]:]
+     # Concatenate the image and concept keys and values
+     img_concept_k = concat([img_k, concept_k])
+     img_concept_v = concat([img_v, concept_v])
+     # Perform the concept attention
+     concept_attn_map = matmul(concept_q, img_concept_k.T)
+     concept_attn_map = softmax(concept_attn_map, axis=-1) * scale
+     concepts = matmul(concept_attn_map, img_concept_v)
+
+     return img, txt, concepts
```

Figure 11. **Pseudo-code depicting the (a) multi-modal attention operation used by Flux DiTs and (b) our CONCEPTATTENTION operation.** We leverage the parameters of a multi-modal attention layer to construct a set of contextualized concept embeddings. The concepts query the image tokens (cross-attention) and other concept tokens (self-attention) in an attention operation. The updated concept embeddings are returned in addition to the image and text embeddings.

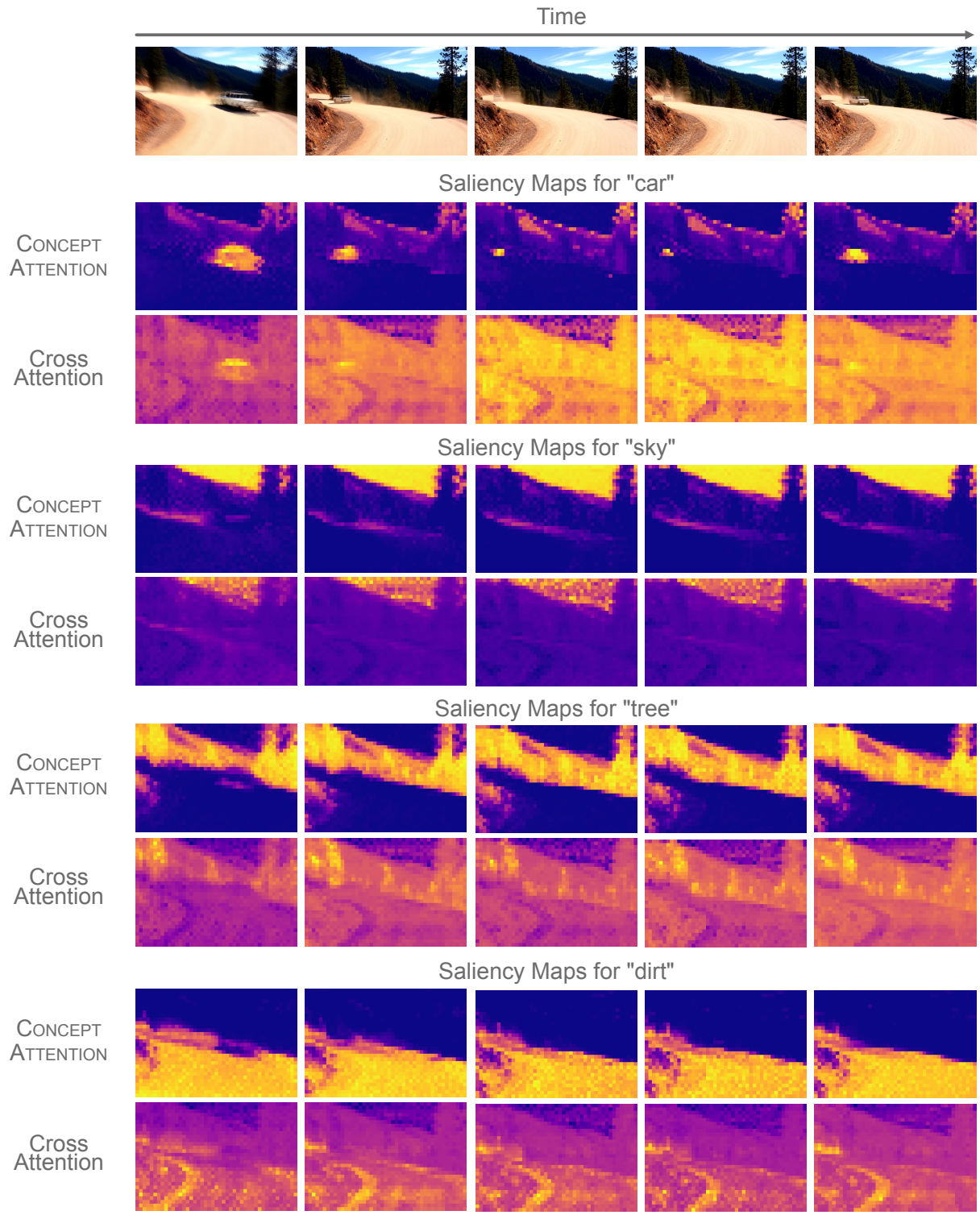

Figure 12. **CONCEPTATTENTION generalizes seamlessly to video generation MMDiT models like CogVideoX.** We apply CONCEPTATTENTION to a CogVideoX [29] video generation model. We take several frames from the video and compare the saliency maps generated by CONCEPTATTENTION to the model's internal cross attention maps.

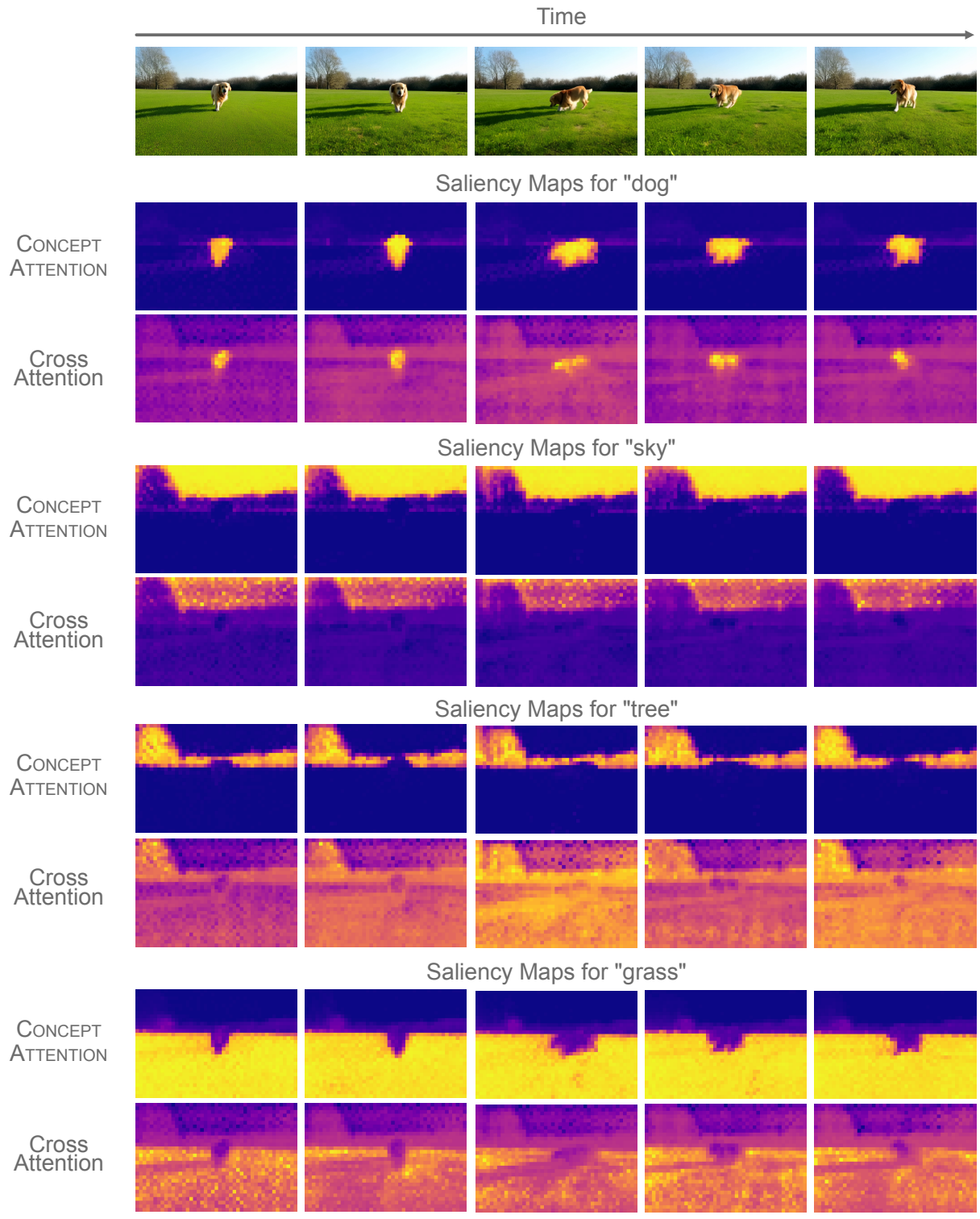

Figure 13. **ConceptAttention generalizes seamlessly to video generation MMDiT models like CogVideoX.** We apply ConceptAttention to a CogVideoX [29] video generation model. We take several frames from the video and compare the saliency maps generated by ConceptAttention to the model's internal cross attention maps.