ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features

Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, Duen Horng Chau

Presented by Jakub Grzywaczewski

MI2.AI Winter Seminar
at Warsaw University of Technology

November 3, 2025

Overview

1. Introduction
2. Background
3. Previous method: DAAM
4. Concept Attention
5. Results

Introduction

QR Codes

(a) Arxiv

(b) OpenReview

(c) ICML presentation

Conference paper

Spotlight Poster

ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features

Alec Helbling · Tuna Han Salih Meral · Benjamin Hoover · Pinar Yanardag · Polo Chau

East Exhibition Hall A-B #E-3001

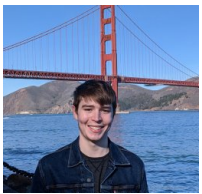
[Abstract] [Lay Summary]

[OpenReview]

Tue 15 Jul 4:30 p.m. PDT — 7 p.m. PDT
Oral presentation: Oral 2D Efficient ML
Tue 15 Jul 3:30 p.m. PDT — 4:30 p.m. PDT

ICML 2025 Oral presentation: Oral 2D Efficient ML
Best Paper Award at CVPR Workshop on Visual Concepts

Authors



Alec Helbling
PhD Student
Georgia Tech

Tuna Meral
PhD Student
Virginia Tech

Ben Hoover
PhD Student
Georgia Tech,
IBM Research

Pinar Yanardag
Asc. Professor
Virginia Tech,
MIT

Duen Chau
Professor
Georgia Tech,
Apple

Reviews

Rating

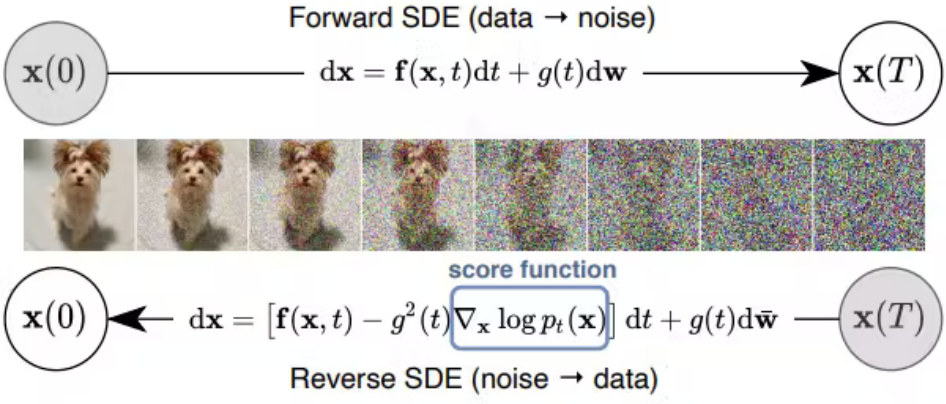
Choose an overall recommendation for this paper.

- 5: Strong accept
- 4: Accept
- 3: Weak accept (i.e., leaning towards accept, but could also be rejected)
- 2: Weak reject (i.e., leaning towards reject, but could also be accepted)
- 1: Reject

Paper ratings: 4, 4, 3, 4.
Recurrent weaknesses: Missing references (TCAV, OVAM), More evaluations needed, Lack of theoretical underpinning

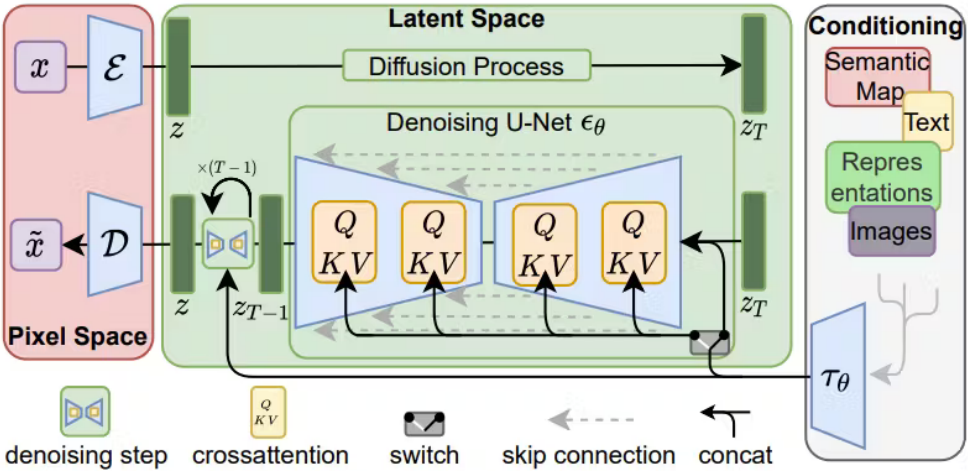
Background

Diffusion models primer



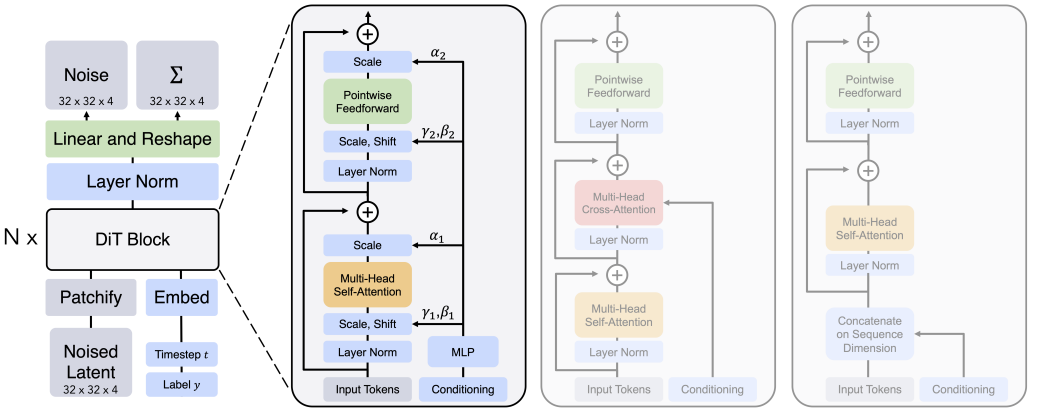
"Score-based generative modeling through stochastic differential equations." [Song et al., 2020]

Latent diffusion



"High-resolution image synthesis with latent diffusion models." [Rombach et al., 2022]

Diffusion Transformer

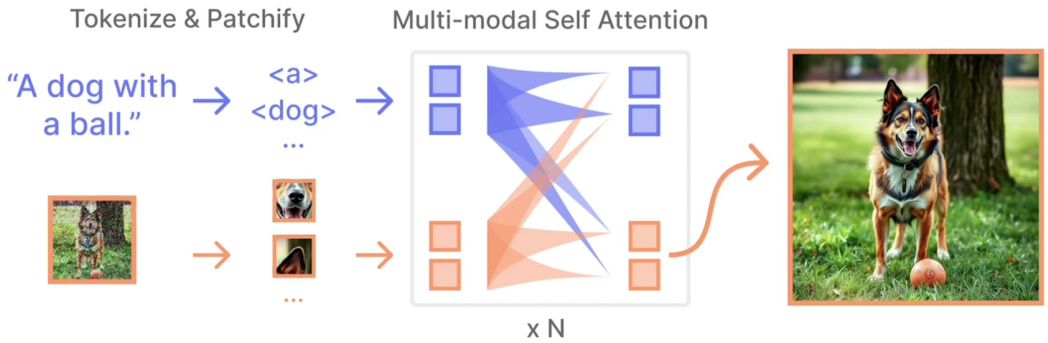


"Scalable diffusion models with transformers." [Peebles and Xie, 2023]

Multi-modal Diffusion Transformer (MMDiT)

A transformer architecture for text-to-image generation.

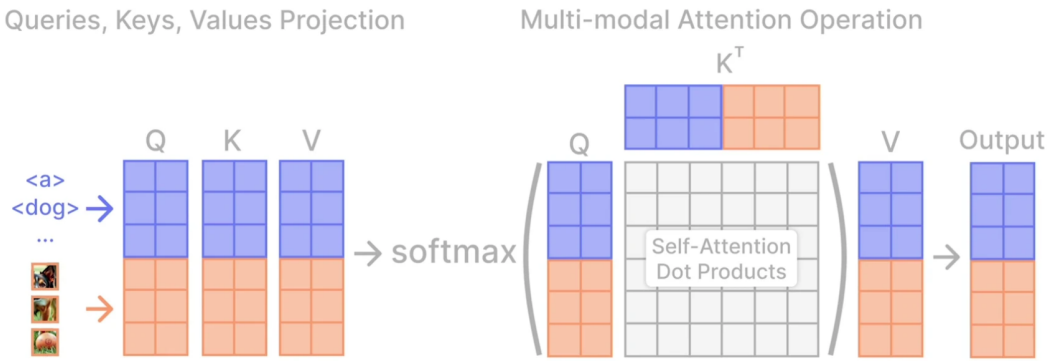
Jointly process text and image tokens with attention layers.



Official ICML presentation slides

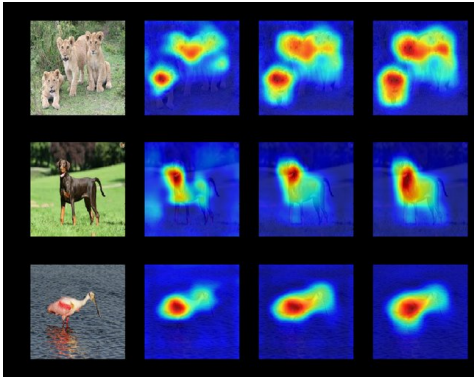
Multi-modal Self-attention

Self-attention on concatenated text and image embeddings.

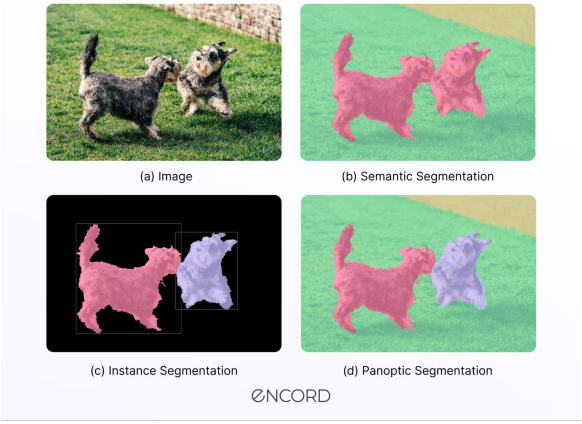


Official ICML presentation slides

Interpreting models with saliency maps



(a) Example of Saliency maps
[GeeksforGeeks, 2021]



(b) Different types of segmentation

Previous method: DAAM

What a DAAM

ACL Anthology

What the DAAM: Interpreting Stable Diffusion Using Cross Attention

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, Ferhan Ture

Abstract

Diffusion models are a milestone in text-to-image generation, but they remain poorly understood, lacking interpretability analyses. In this paper, we perform a text-image attribution analysis on Stable Diffusion, a recently open-sourced model. To produce attribution maps, we upscale and aggregate cross-attention maps in the denoising module, naming our method DAAM. We validate it by testing its segmentation ability on nouns, as well as its generalized attribution quality on all parts of speech, rated by humans. On two generated datasets, we attain a competitive 58.8-64.8 mIoU on noun segmentation and fair to good mean opinion scores (3.4-4.2) on generalized attribution. Then, we apply DAAM to study the role of syntax in the pixel space across head–dependent heat map interaction patterns for ten common dependency relations. We show that, for some relations, the head map consistently subsumes the dependent, while the opposite is true for others. Finally, we study several semantic phenomena, focusing on feature entanglement; we find that the presence of cohyponyms worsens generation quality by 9%, and descriptive adjectives attend too broadly. We are the first to interpret large diffusion models from a visuolinguistic perspective, which enables future research. Our code is at https://github.com/castorini/daam.

PDF

Cite

Search

Fix data

Diffusion Attentive Attribution Maps (DAAM)

Given a 2D latent $\ell_t \in \mathbb{R}^{w \times h}$, the downsampling blocks output a series of vectors $\{h^{\downarrow}_{i,t}\}_{i=1}^{K}$, where $h^{\downarrow}_{i,t} \in \mathbb{R}^{\lceil \frac{w}{c^i} \rceil \times \lceil \frac{h}{c^i} \rceil}$ for some $c > 1$. The upsampling blocks then iteratively upscale $h^{\downarrow}_{K,t}$ to $\{h^{\uparrow}_{i,t}\}_{i=K-1}^{0} \in \mathbb{R}^{\lceil \frac{w}{c^i} \rceil \times \lceil \frac{h}{c^i} \rceil}$:

$$h^{\downarrow}_{i,t} := F^{(i)}_t(\hat{h}^{\downarrow}_{i,t}, X) \cdot (W^{(i)}_v X), \tag{1}$$

$$F^{(i)}_t(\hat{h}^{\downarrow}_{i,t}, X) := \text{softmax}\left((W^{(i)}_q \hat{h}^{\downarrow}_{i,t})(W^{(i)}_k X)^T / \sqrt{d}\right), \tag{2}$$

where $F^{(i)\downarrow}_t \in \mathbb{R}^{\lceil \frac{w}{c^i} \rceil \times \lceil \frac{h}{c^i} \rceil \times l_H \times l_W}$ and W_k, W_q, and W_v are projection matrices with l_H attention heads. The same mechanism applies when upsampling $h^{\uparrow}_i$.
The DAAM is computed as:

$$D^{\mathbb{R}}_k[x, y] := \sum_{i,j,\ell} \tilde{F}^{(i)\downarrow}_{t_j,k,\ell}[x, y] + \tilde{F}^{(i)\uparrow}_{t_j,k,\ell}[x, y], \tag{3}$$

where k is the k^{th} word and $\tilde{F}^{(i)\downarrow}_{t_j,k,\ell}[x, y]$ is shorthand for $F^{(i)\downarrow}_t[x, y, \ell, k]$, bicubically upscaled to fixed size (w, h).

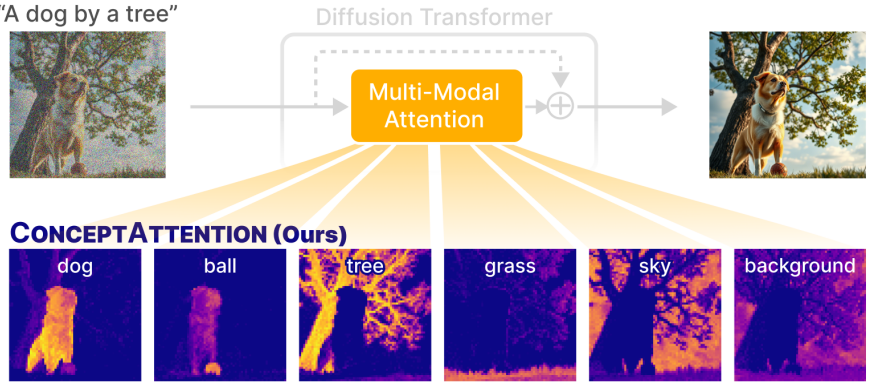
DAAM Results

"What the daam: Interpreting stable diffusion using cross attention." [Tang et al., 2022]

Concept Attention

Concept Attention: Overview

CONCEPTATTENTION interprets the representations of diffusion transformers by producing saliency maps of text concepts.



Official GitHub ConceptAttention repository

Concept Attention: Promises

Concept Attention [Helbling et al., 2025]:

- Better quality of saliency maps (ex. zero-shot segmentation quality),
- No training required,
- Works for concepts not in the prompt,
- Minimal overhead,
- Generalization to video generation.

Concept Attention: Main method 1/2

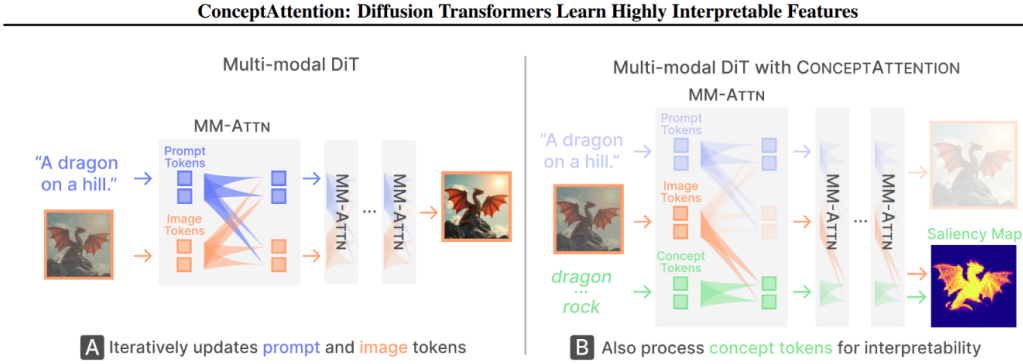


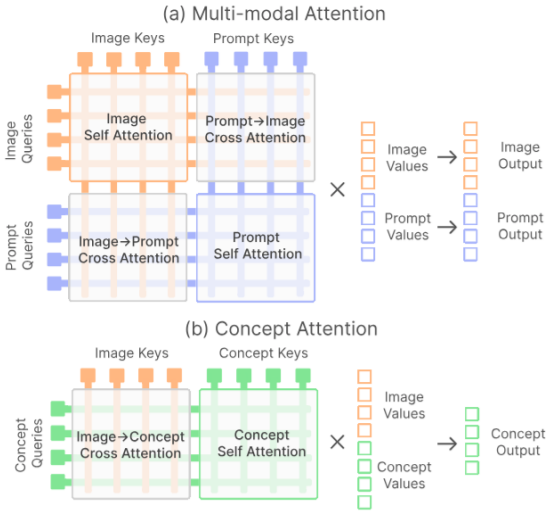
Figure 2. **CONCEPTATTENTION augments multi-modal DiTs with a sequence of concept embeddings that can be used to produce saliency maps.** (Left) An unmodified multi-modal attention (MMATTN) layer processes both **prompt** and **image** tokens. (Right) CONCEPTATTENTION augments these layers without impacting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

Figure from the paper

Concept Attention: Main method 2/2

Concept attention works by using:

- One-directional attention from the image tokens to the concept set tokens,
- Self-attention within the concept set,
- A concept residual stream,
- Dot-product similarity between image output vectors and concept output vectors



Concept Attention: Algorithm

(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 9. **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.

Algorithm from the paper

Results

Qualitative Results

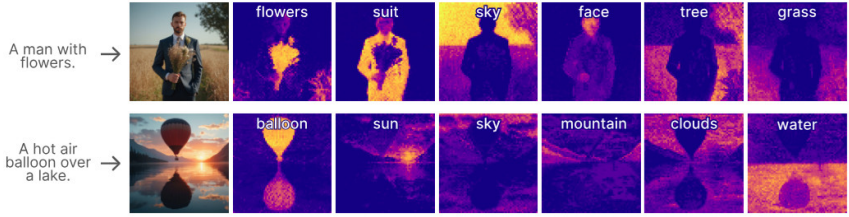


Figure 3. **CONCEPTATTENTION can generate high-quality saliency maps for multiple concepts simultaneously.**

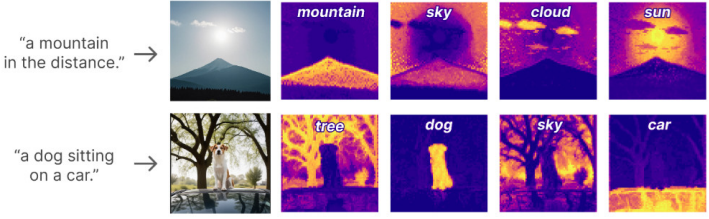


Figure 6. **CONCEPTATTENTION is capable of generating high quality saliency maps with Stable Diffusion 3.5 Turbo.**

Figures from the paper

Qualitative Comparison

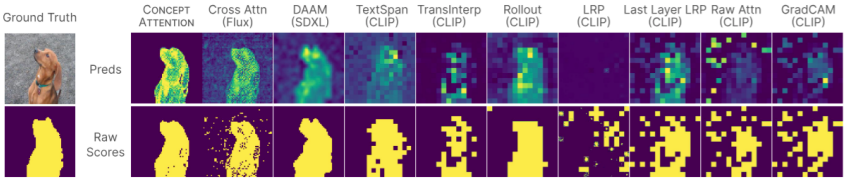
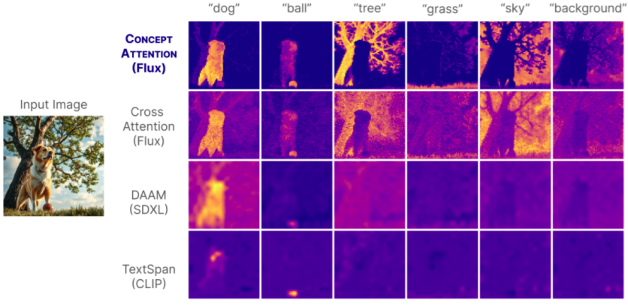


Figure 4. **ConceptAttention produces higher fidelity raw scores and saliency maps than alternative methods**



Figures from the paper

Quantitative Results (Zero-shot Segmentation)

Method	Architecture	ImageNet-Segmentation			PascalVOC (Single Class)		
		Acc ↑	mIoU↑	mAP↑	Acc ↑	mIoU↑	mAP↑
LRP (Binder et al., 2016)	CLIP ViT	51.09	32.89	55.68	48.77	31.44	52.89
Partial-LRP (Binder et al., 2016)	CLIP ViT	76.31	57.94	84.67	71.52	51.39	84.86
Rollout (Abnar & Zuidema, 2020)	CLIP ViT	73.54	55.42	84.76	69.81	51.26	85.34
ViT Attention (Dosovitskiy et al., 2021)	CLIP ViT	67.84	46.37	80.24	68.51	44.81	83.63
GradCAM (Selvaraju et al., 2020)	CLIP ViT	64.44	40.82	71.60	70.44	44.90	76.80
TextSpan (Gandelsman et al., 2024)	CLIP ViT	75.21	54.50	81.61	75.00	56.24	84.79
TransInterp (Chefer et al., 2021)	CLIP ViT	79.70	61.95	86.03	76.90	57.08	86.74
CLIPasRNN (Sun et al., 2024)	CLIP ViT	74.05	58.80	84.80	61.76	41.48	76.57
OVAM (Marcos-Manchón et al., 2024)	SDXL UNet	79.41	65.02	88.12	73.50	58.12	87.91
DINO SA (Caron et al., 2021)	DINO ViT	81.97	69.44	86.12	80.71	64.33	88.90
DINOv2 SA (Oquab et al., 2024)	DINOv2 ViT	77.39	63.12	84.19	79.65	57.61	87.26
DINOv2 Reg SA (Darcet et al., 2024)	DINOv2 Reg	72.04	56.31	80.83	77.16	56.60	86.35
iBOT SA (Zhou et al., 2022)	iBOT ViT	76.34	61.73	82.04	74.96	55.80	85.26
DAAM (Tang et al., 2022)	SDXL UNet	78.47	64.56	88.79	72.76	55.95	88.34
DAAM (Tang et al., 2022)	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 1. CONCEPTATTENTION outperforms a variety of Diffusion, DINO, and CLIP ViT interpretability methods on ImageNet-Segmentation and PascalVOC (Single Class).

Figures from the paper

Ablations 1/2

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 3. **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.

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 4. **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 from the paper

Ablations 2/2

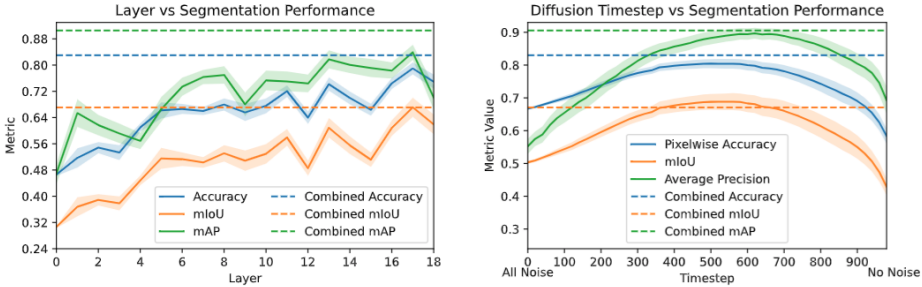


Figure 7. **(Left) Later MMATTN layers encode richer features for zero-shot segmentation.** We investigated the impact of using features from various MMATTN layers and found that deeper layers lead to better performance on segmentation metrics like pixelwise accuracy, mIoU, and mAP. We also found that 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.

Figure from the paper

What if no relevant concepts are present?

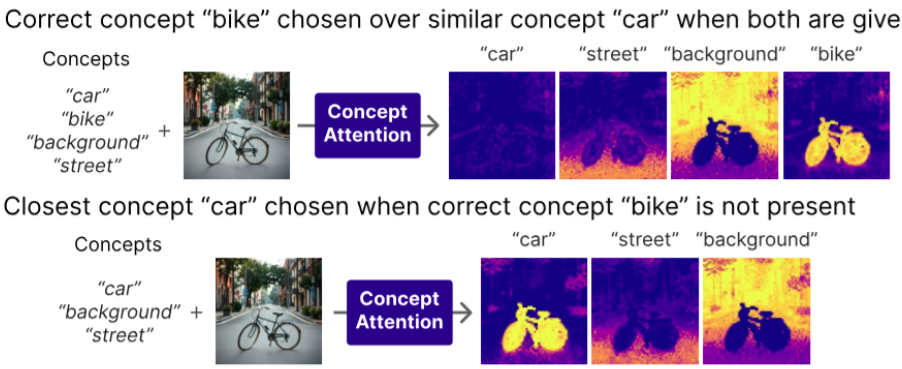


Figure 15. **The behavior of CONCEPTATTENTION when multiple relevant concepts are present and when no relevant one is.** When multiple similar concepts are given, like “car” and “bike”, the most similar one will be chosen. However, when no relevant concept is presented, CONCEPTATTENTION will fall back on the most relevant one, in this case “car“ for the bike patches.

Figure from the paper

Generalization to video generation

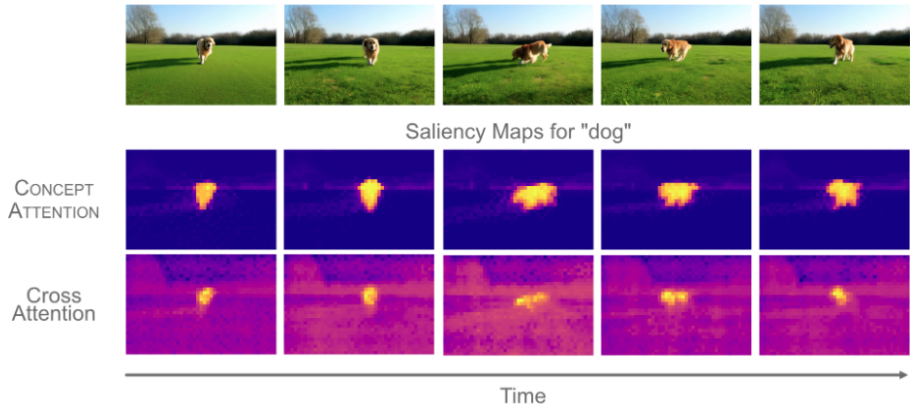
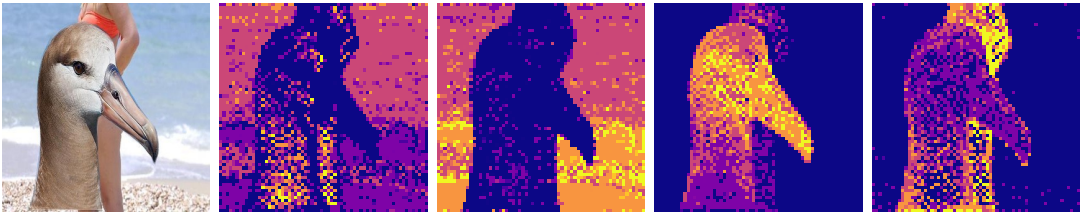


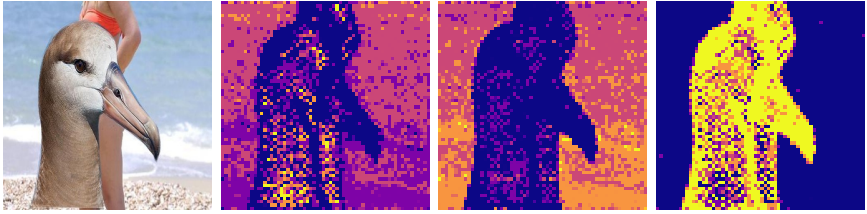
Figure 8. **CONCEPTATTENTION generalizes seamlessly to video generation MMDiT models like CogVideoX.**

Figure from the paper

My Results 1/2



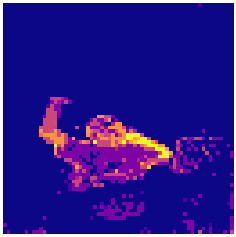
(a) Original (b) Sea (c) Sky (d) Bird (e) Person



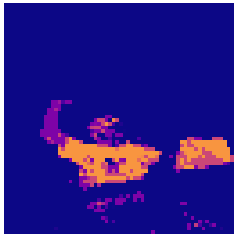
(a) Original (b) Sea (c) Sky (d) Person

My Results 2/2

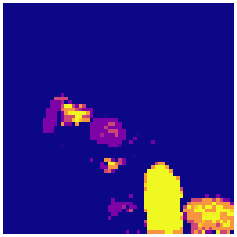
Figure: Original



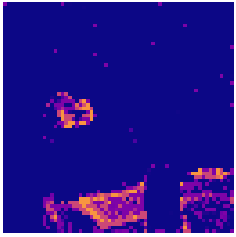
(a) Person



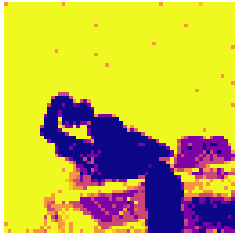
(b) Jacket



(c) Drink



(d) Paper



(e) Background

References

GeeksforGeeks (2021).
What is saliency map?
Last updated 29 October 2021.

Helbling, A., Meral, T. H. S., Hoover, B., Yanardag, P., and Chau, D. H. (2025).
Conceptattention: Diffusion transformers learn highly interpretable features.
arXiv preprint arXiv:2502.04320.

Peebles, W. and Xie, S. (2023).
Scalable diffusion models with transformers.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022).
High-resolution image synthesis with latent diffusion models.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020).
Score-based generative modeling through stochastic differential equations.
arXiv preprint arXiv:2011.13456.

Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., and Ture, F. (2022).
What the daam: Interpreting stable diffusion using cross attention.
arXiv preprint arXiv:2210.04885.

The End