北京大學
PEKING UNIVERSITY

ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features

Alec Helbling   Tuna Han Salih Meral   Benjamin Hoover   Pinar Yanardag   Duen Horng (Polo) Chau

Georgia Tech   Virginia Tech   IBM Research

ICML 2025 (Oral)

Presenter: Jinyi Luo
2025.07.28

Background

Diffusion Models are Black-Boxes

Input Image

Transformer,
Mid-Level Feature

Diffusion,
Mid-Timestep Latent

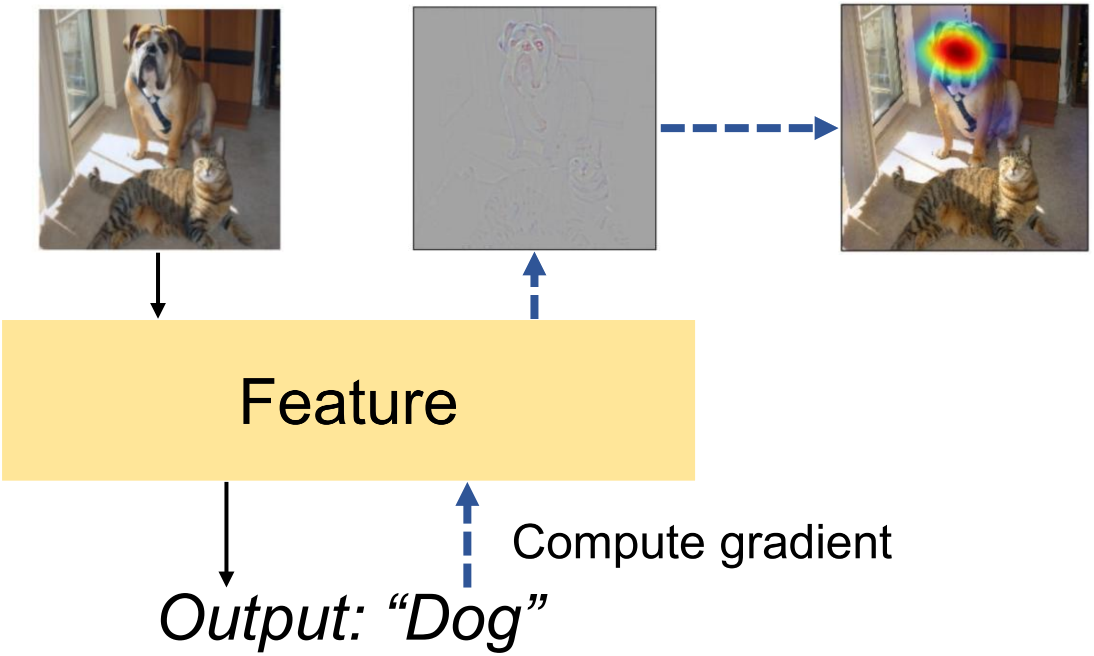
Representation for Restoration

Representation for Generation

Background

Existing Vision Interpretation Methods

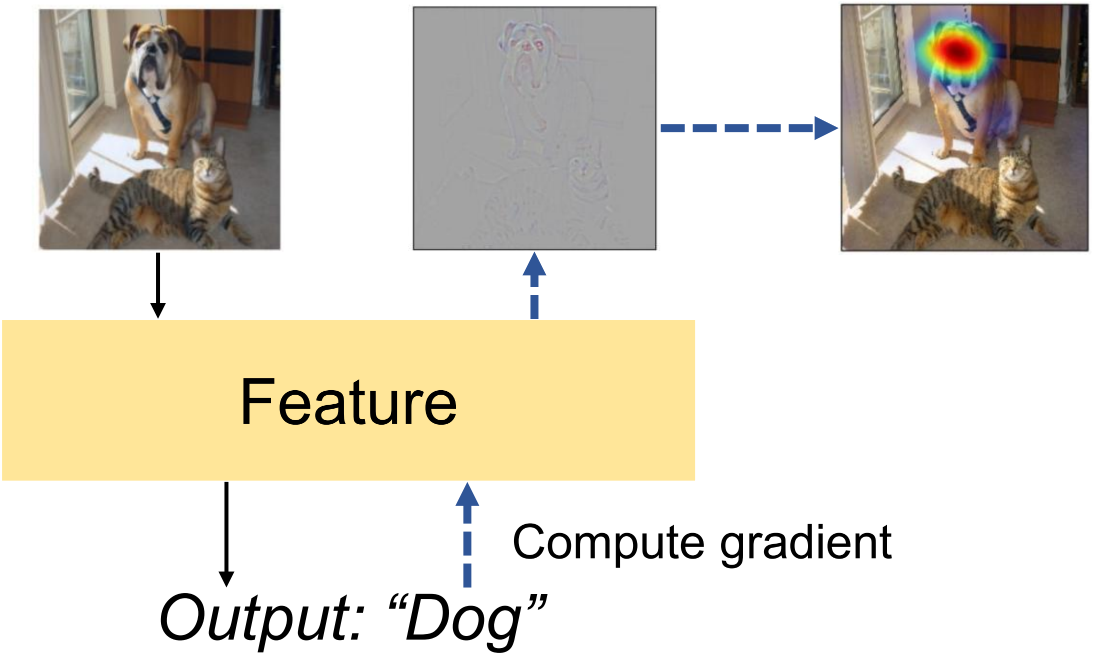
Gradient Interpretation



Hard to apply on generative models

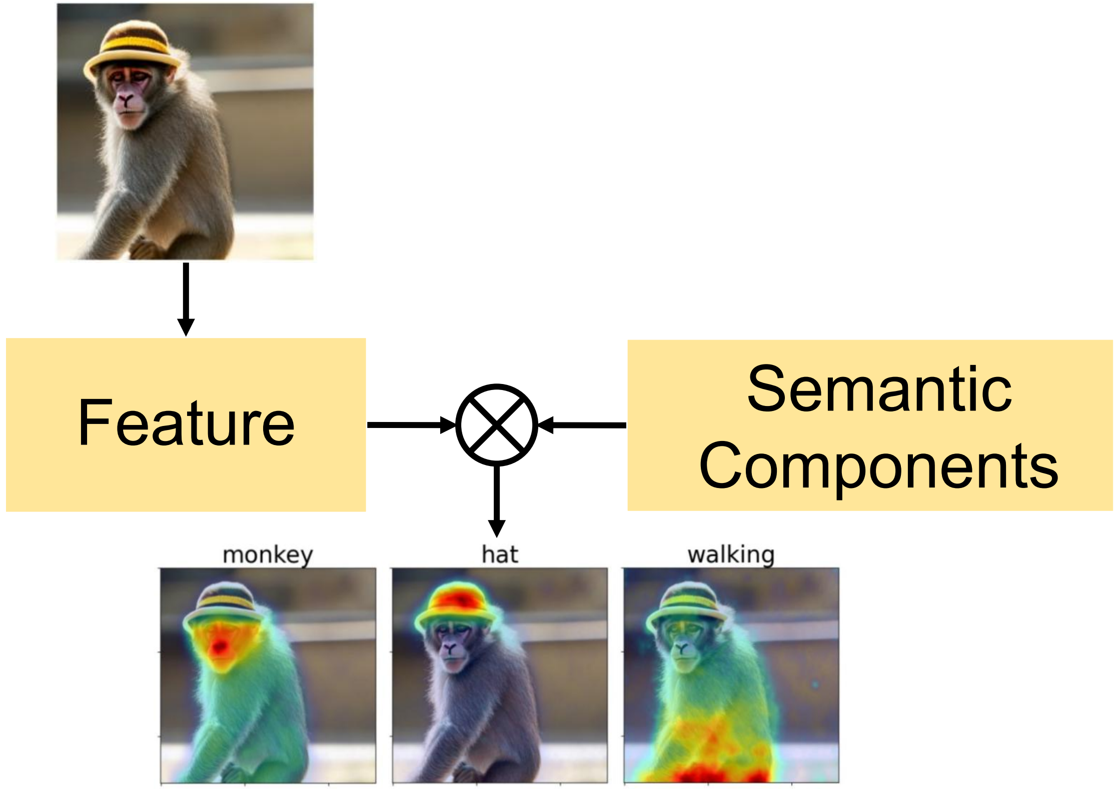
Existing Vision Interpretation Methods

Gradient Interpretation



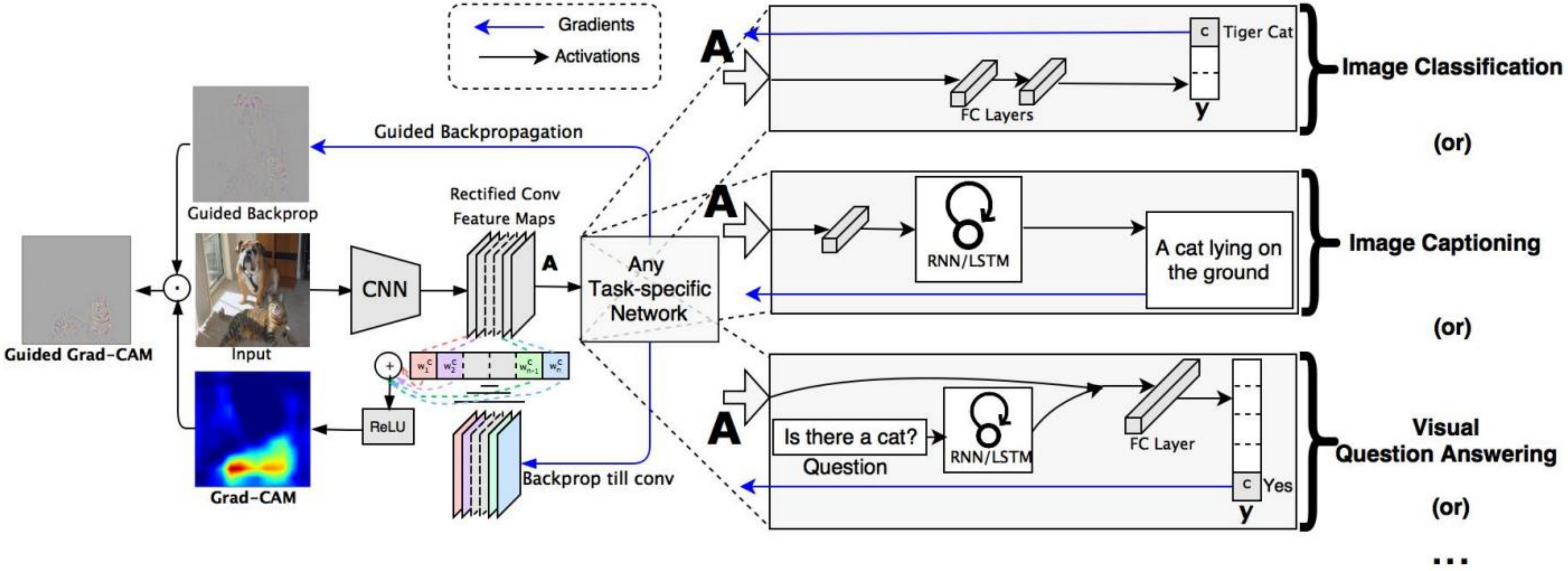
Hard to apply on generative models

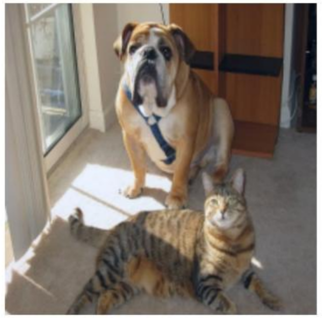
Representation Interpretation

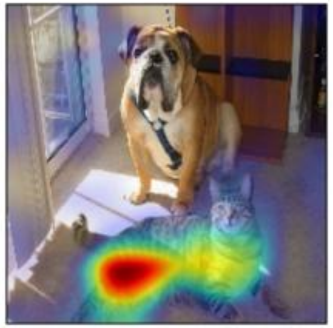


Requires model-specific design

Gradient-Based Interpretation




(a) Original Image


(c) Grad-CAM 'Cat'

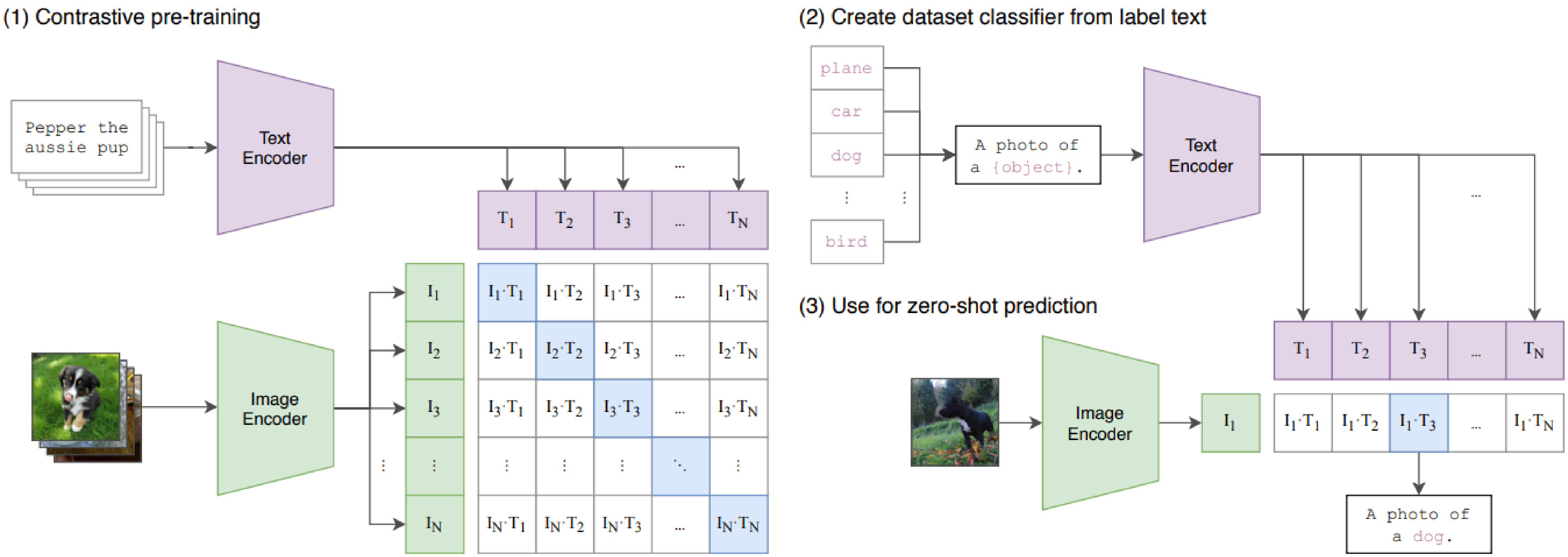

(i) Grad-CAM 'Dog'

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

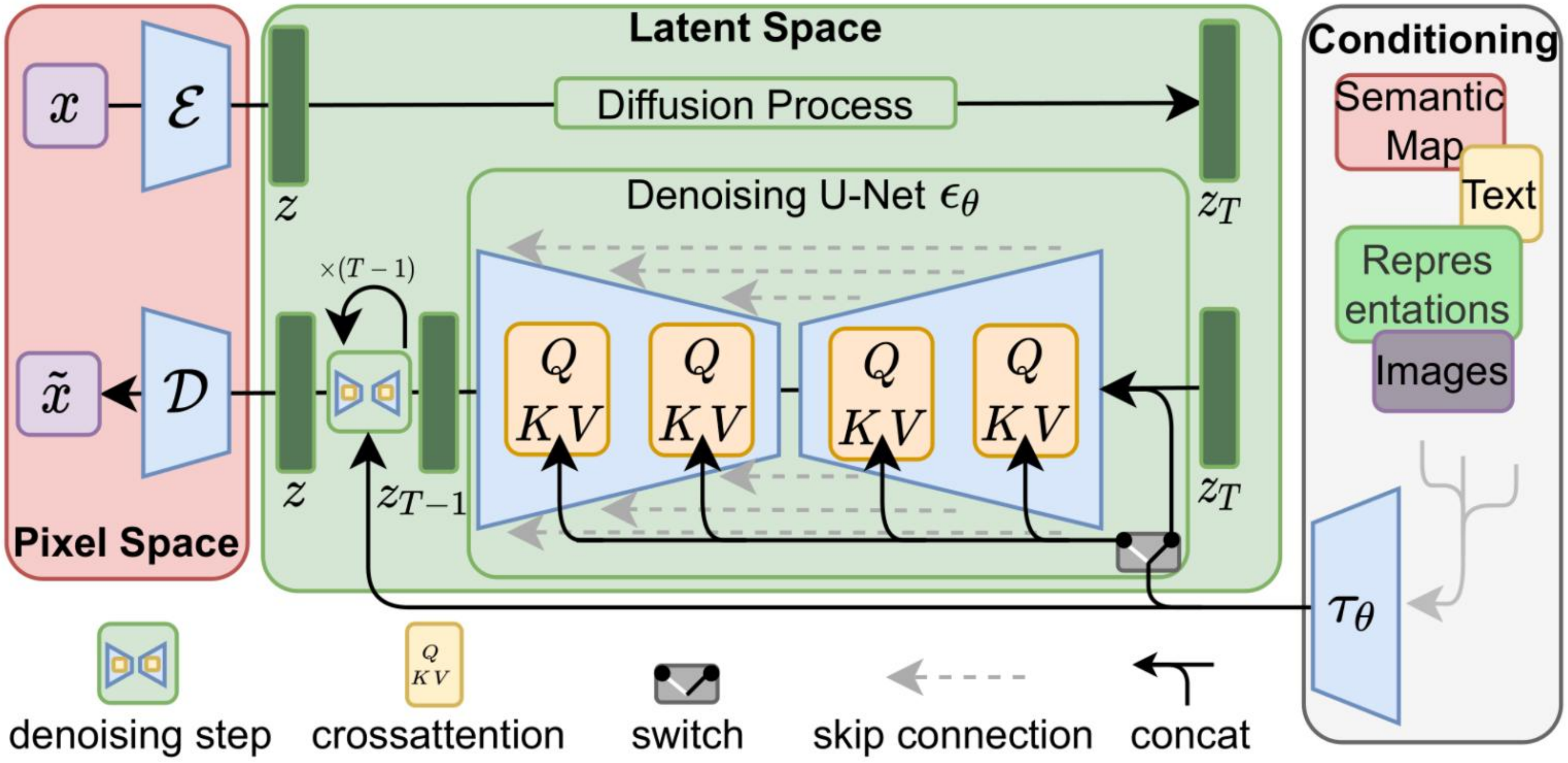
Ramprasaath et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017

Cross-Attention for Representation Interpretation



Radford et. al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021

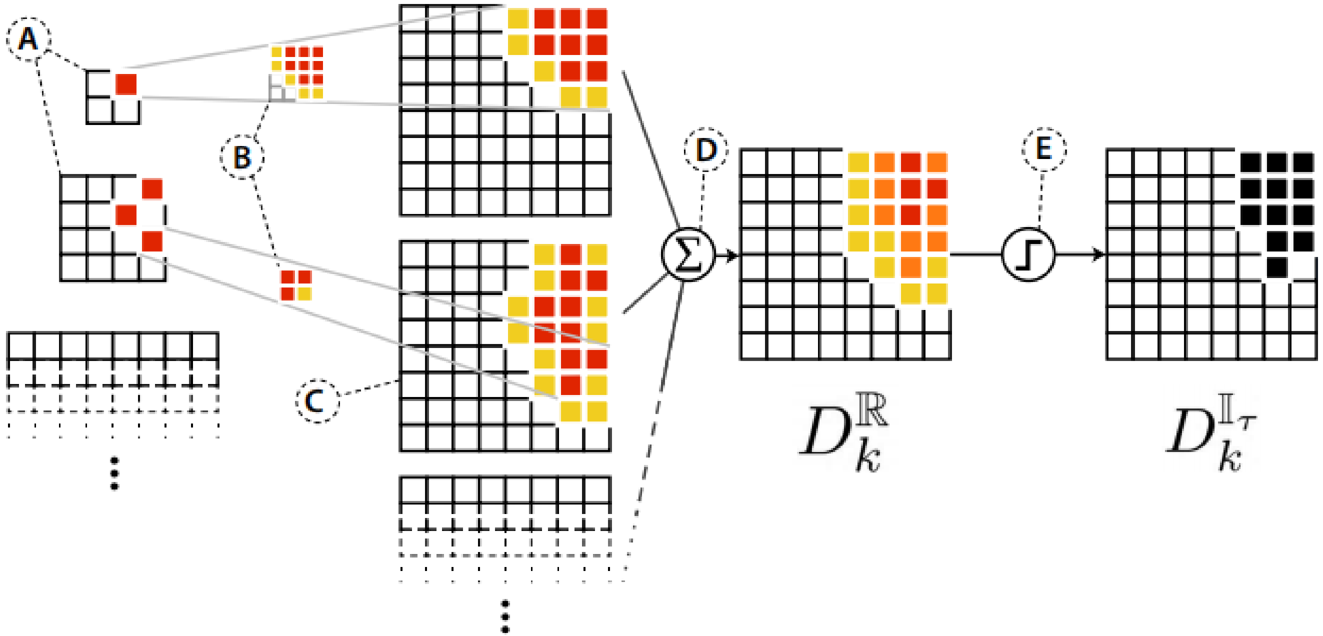
Cross-Attention for Representation Interpretation

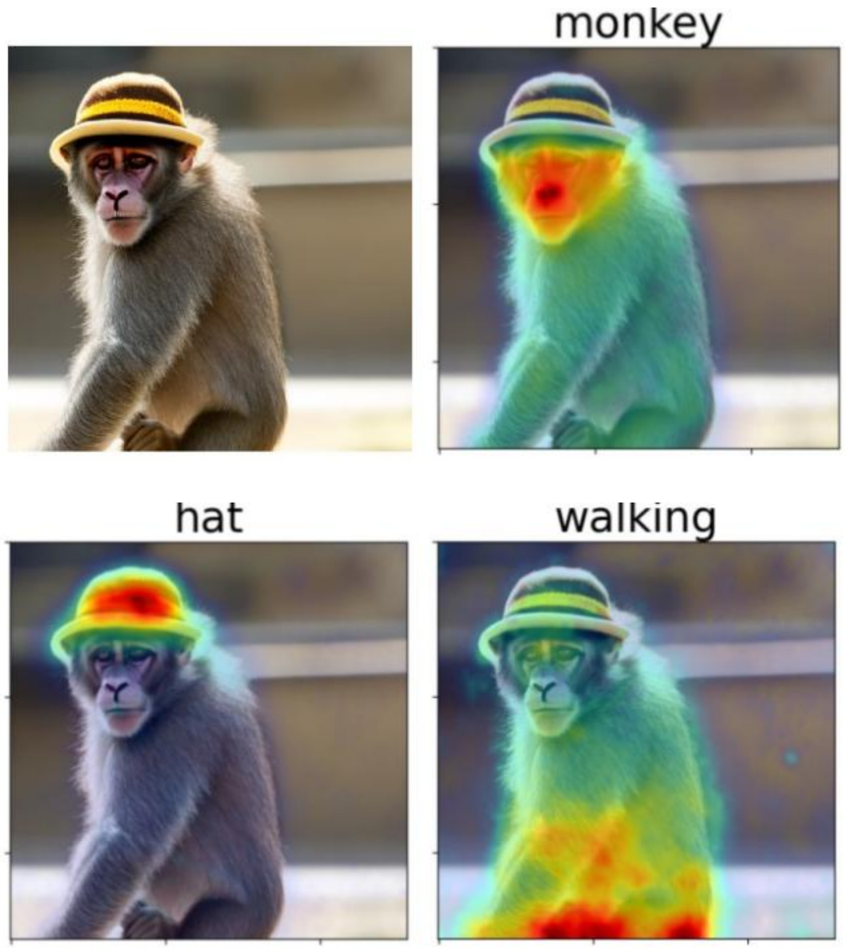


Rombach et. al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022

Cross-Attention for Representation Interpretation

- Summing attention from multiple layers
- More fine-grained saliency maps



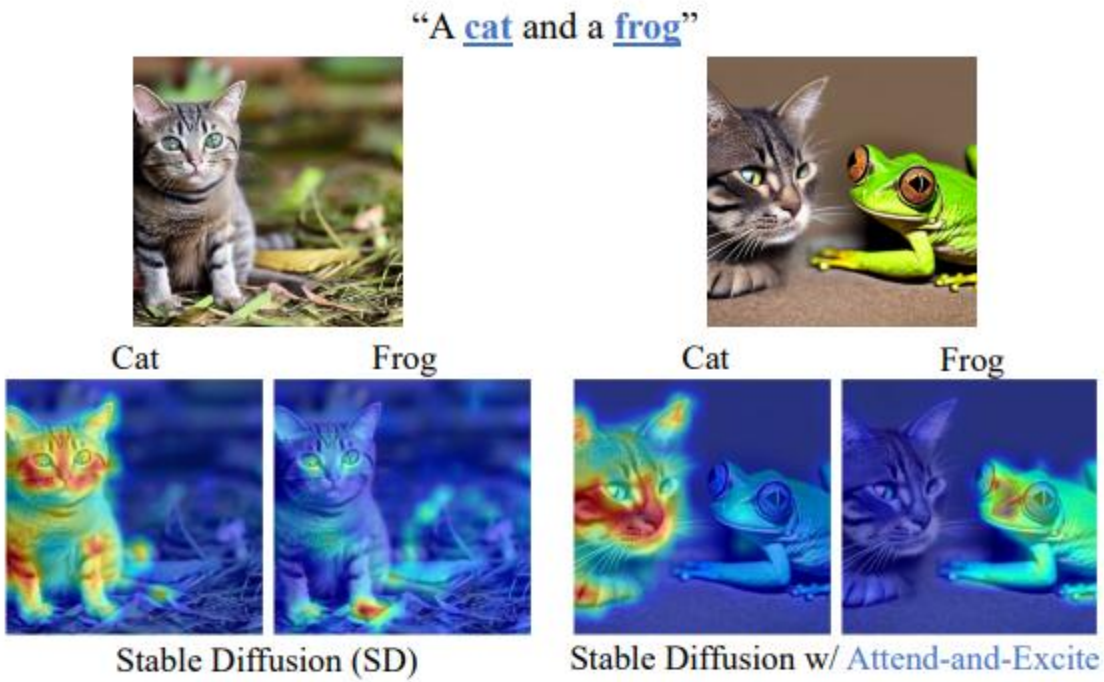


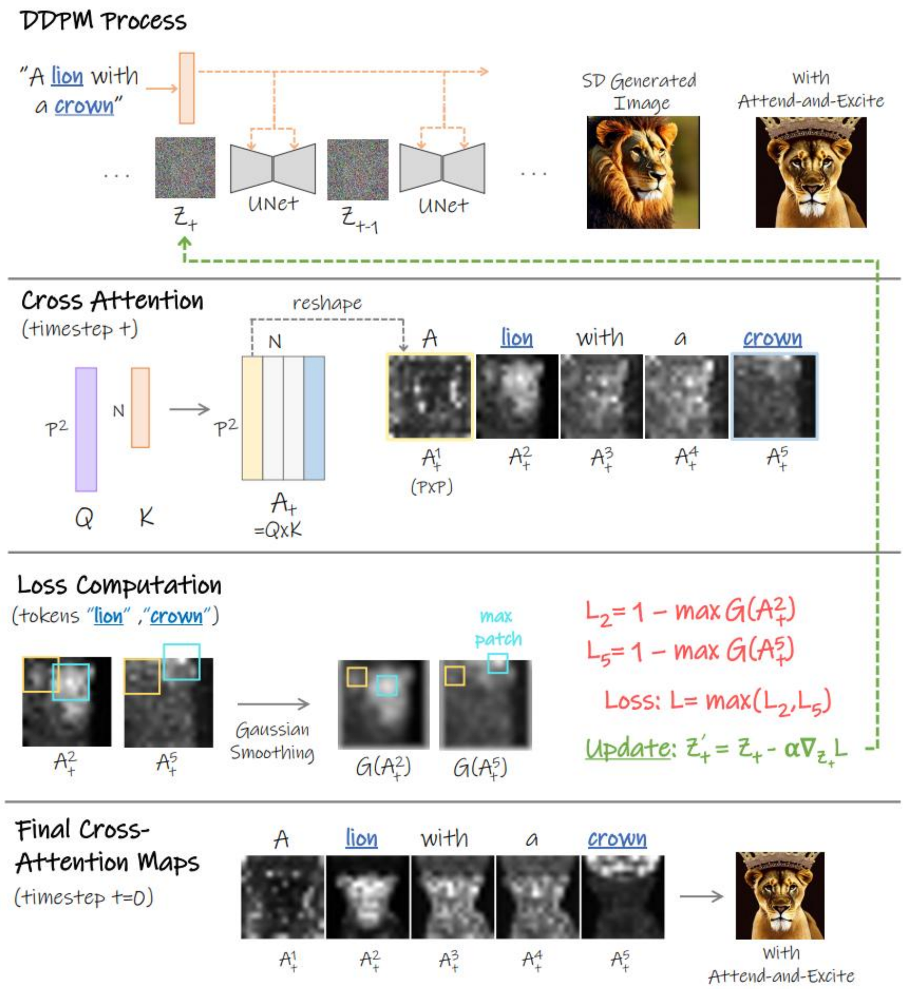
Raphael et. al., What the DAAM: Interpreting Stable Diffusion Using Cross Attention, arXiv 2022

Background

Cross-Attention for Representation Interpretation

- Loss for Instance Enhancement
- More fine-grained saliency maps

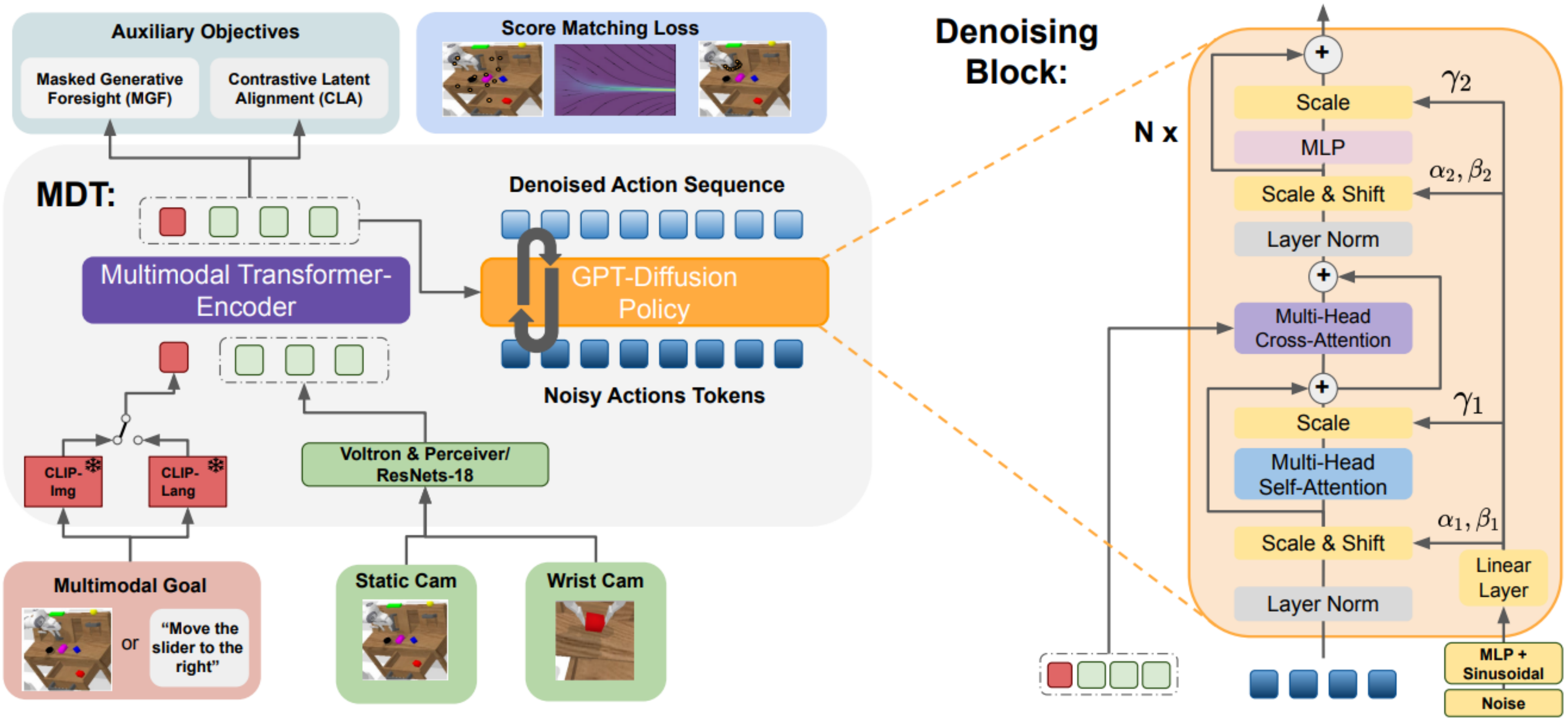




Chefer *et. al.*, Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models, TOG 2023

Multi-Modal Diffusion Transformers

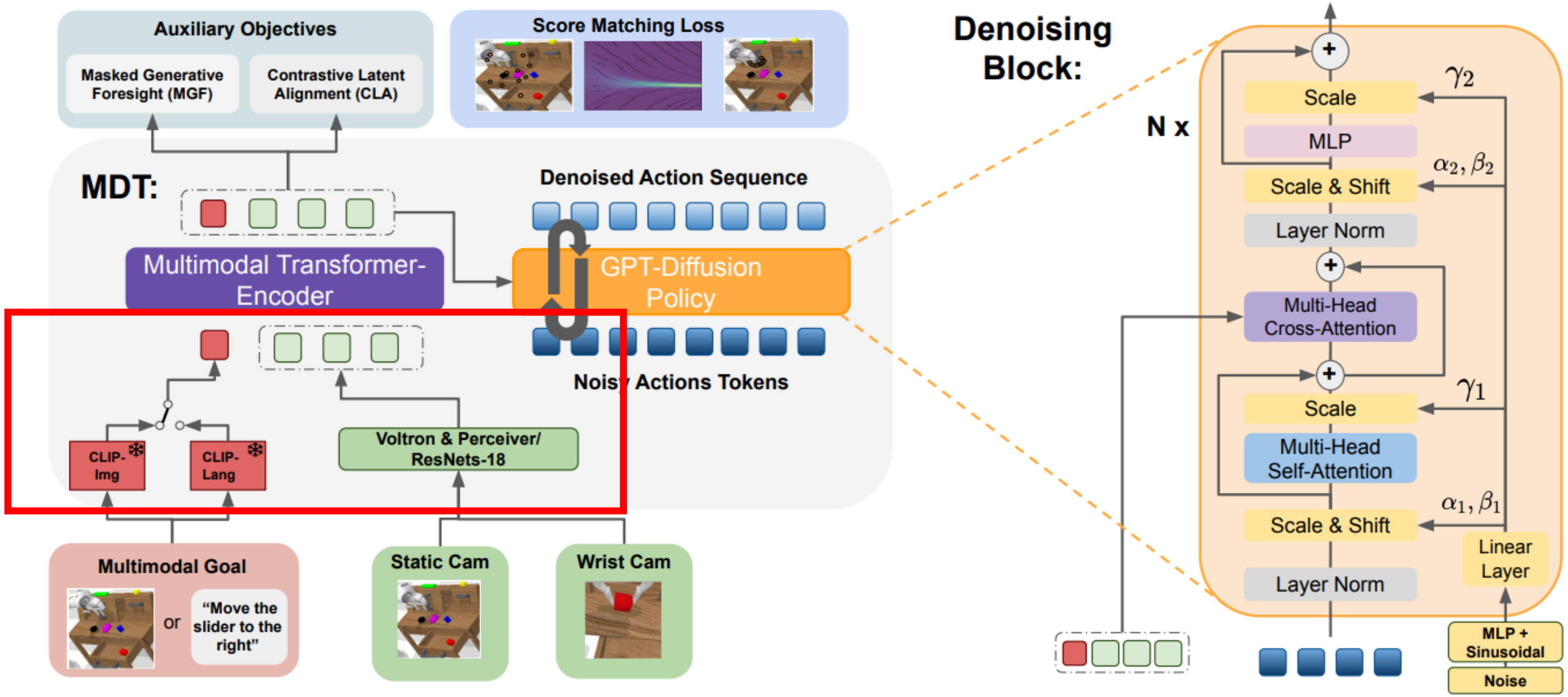
Uniform perspective of different modalities



Reuss et al., Multimodal Diffusion Transformer: Learning Versatile Behavior from Multimodal Goals, Robotics: Science and Systems 2024

Multi-Modal Diffusion Transformers

Uniform perspective of different modalities

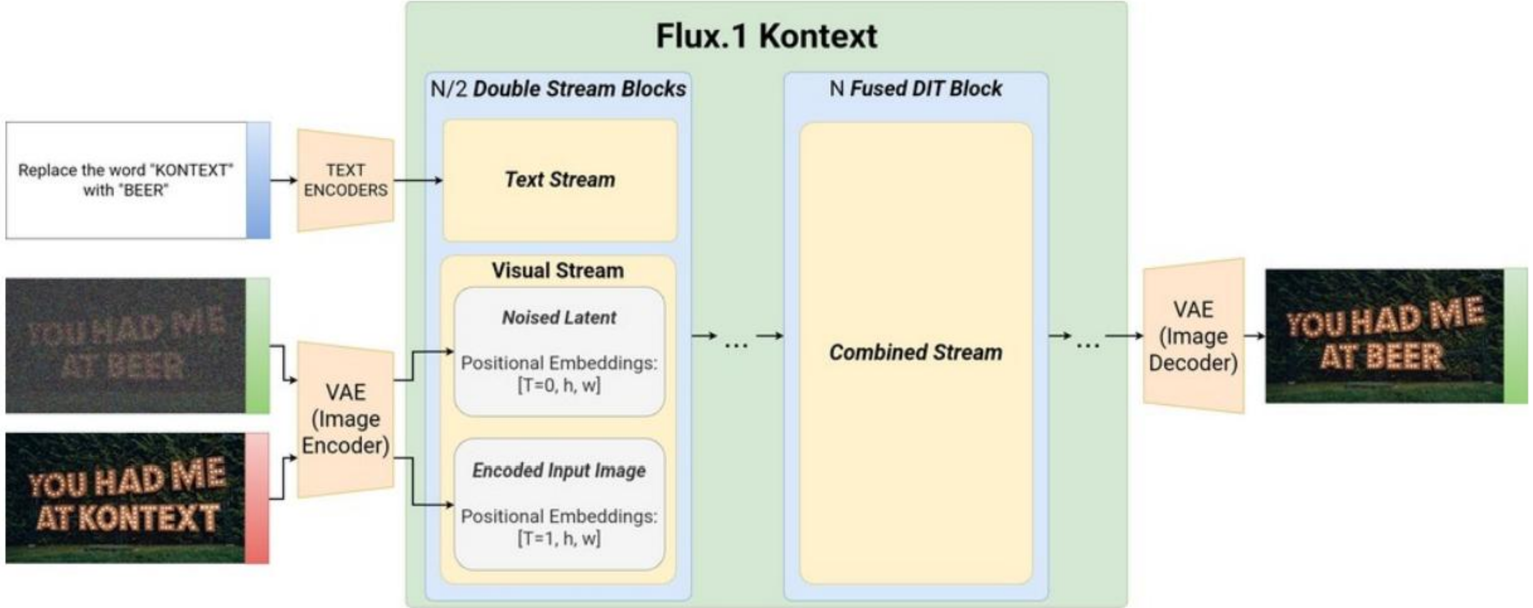


- Concat modalities
- Self-Attention
- Text prompt embeddings are also updated

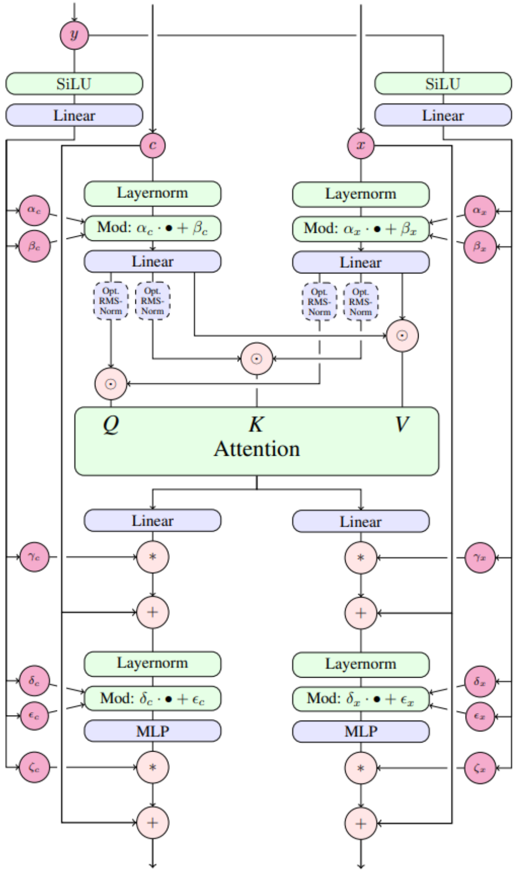
Reuss et al., Multimodal Diffusion Transformer: Learning Versatile Behavior from Multimodal Goals, Robotics: Science and Systems 2024

Background

DiTs with MMAttn: FLUX, SD 3+, …



Flux with both Double Stream Attention
and Fused Self-Attention



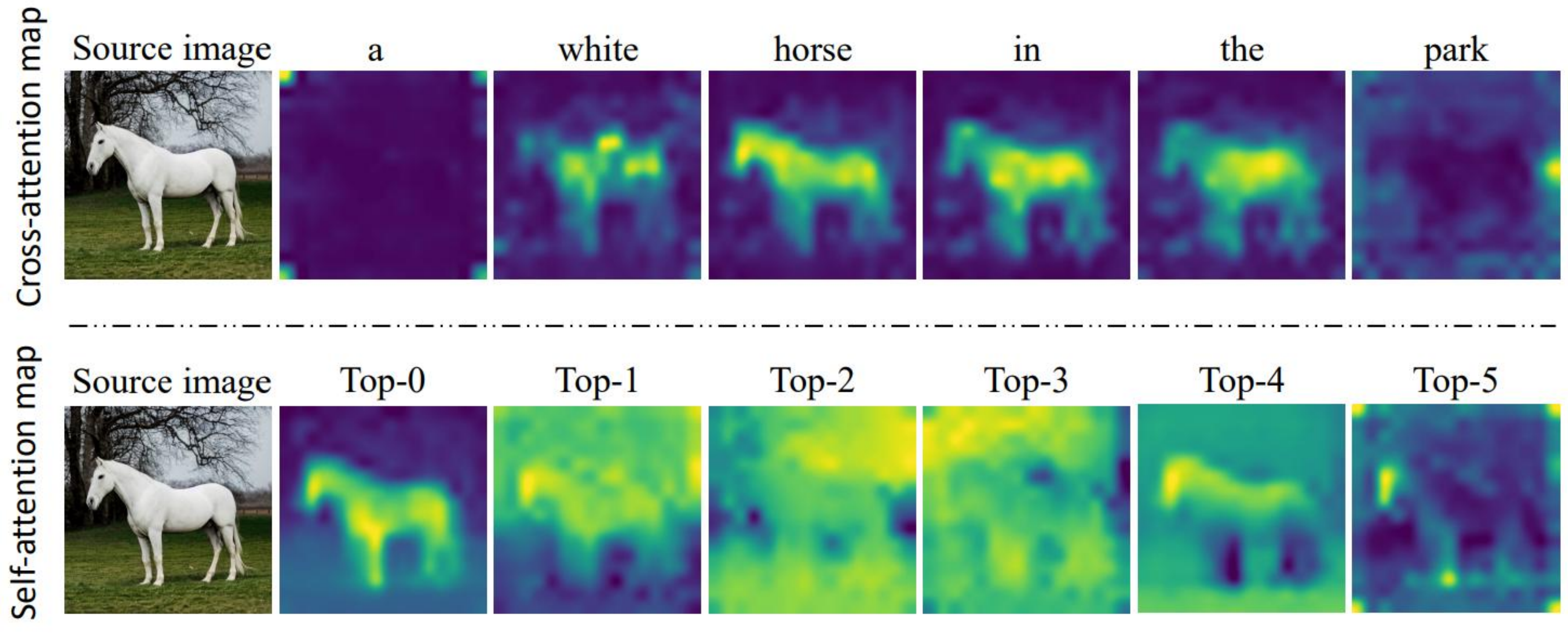
Stable Diffusion 3 with Noise-Condition
Fused Self-Attention

Patric et al., Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, ICML 2024
Batifol et al., FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space, arXiv 2025

Background

Attention Functionalities in MMDiT

- Semantic information from multi-modal cross-attention
- Geometric and shape details from self-attention



Class	Layer 3	Layer 6	Layer 9	Layer 10	Layer 12	Layer 14	Layer 16	Avg.
dog	1.00	1.00	1.00	1.00	0.89	0.76	1.00	0.95
horse	0.96	1.00	1.00	1.00	0.64	1.00	0.91	0.93
sheep	0.97	1.00	1.00	1.00	1.00	0.90	0.97	0.98
leopard	0.97	1.00	1.00	1.00	0.97	0.79	0.87	0.94
tiger	1.00	1.00	0.97	1.00	0.88	1.00	0.97	0.97
green	0.93	0.91	0.91	0.96	0.67	0.38	0.60	0.77
white	0.97	1.00	0.94	0.97	0.97	0.61	0.85	0.90
orange	0.97	1.00	0.94	0.92	0.89	0.94	0.83	0.93
yellow	0.96	0.77	1.00	0.98	1.00	0.36	0.68	0.82
red	0.97	0.97	0.93	0.85	0.70	0.23	0.65	0.76

Class	Layer 3	Layer 6	Layer 9	Layer 10	Layer 12	Layer 14	Layer 16	Avg.
dog	0.53	0.60	0.78	0.60	0.53	0.47	0.38	0.55
horse	0.50	0.70	0.82	0.65	0.68	0.53	0.28	0.59
sheep	0.53	0.45	0.25	0.45	0.62	0.53	0.25	0.44
leopard	0.47	0.65	0.57	0.60	0.47	0.65	0.60	0.57
tiger	0.23	0.12	0.55	0.20	0.45	0.42	0.53	0.36
green	0.00	0.00	0.05	0.00	0.05	0.00	0.12	0.03
white	0.00	0.05	0.30	0.55	0.03	0.15	0.25	0.19
orange	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
yellow	0.00	0.42	0.07	0.05	0.00	0.30	0.07	0.13
red	0.00	0.15	0.28	0.20	0.00	0.20	0.10	0.13

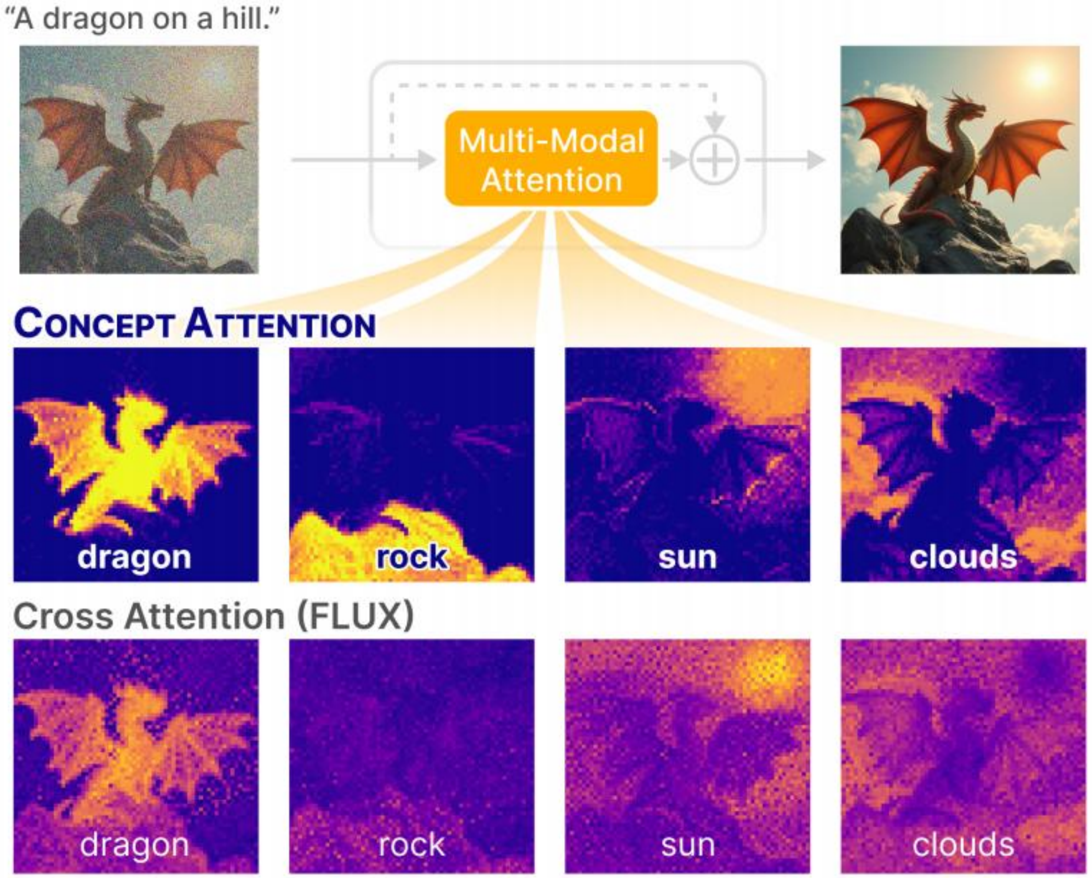
Liu et. al., Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing, CVPR 2025

PEKING UNIVERSITY

Attention in MMDiT contains rich semantic information as well.

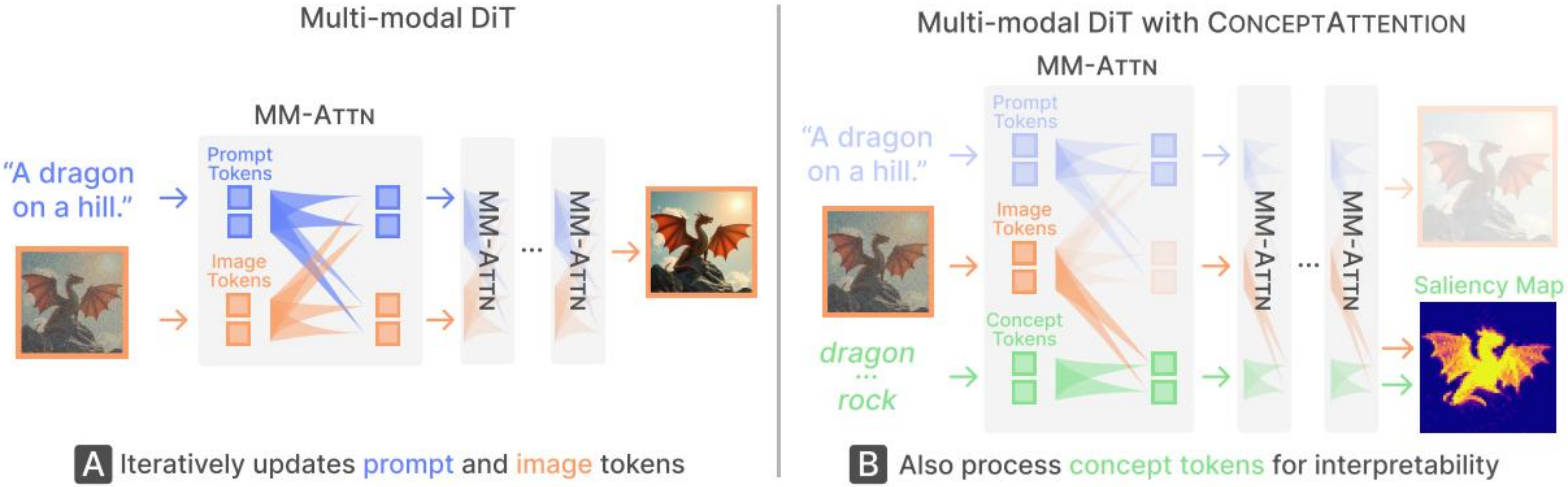
How to interpret it and acquire accurate saliency map?

Raw Cross-Attention is not accurate in MMDiTS

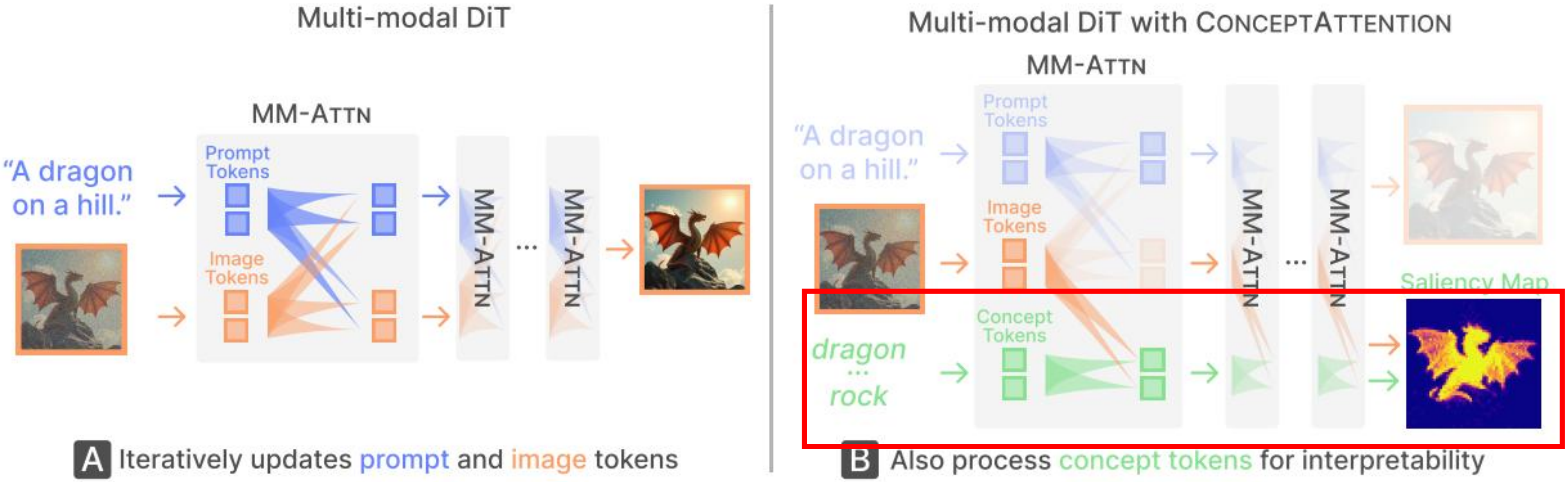


- The prompt domain is updated alongside image domain
- Fine semantic in deeper level cannot be accessed

Concept Residual Stream



Concept Residual Stream



Concept tokens are updated through each layer, but does not interact generation

PEKING UNIVERSITY 北京大學

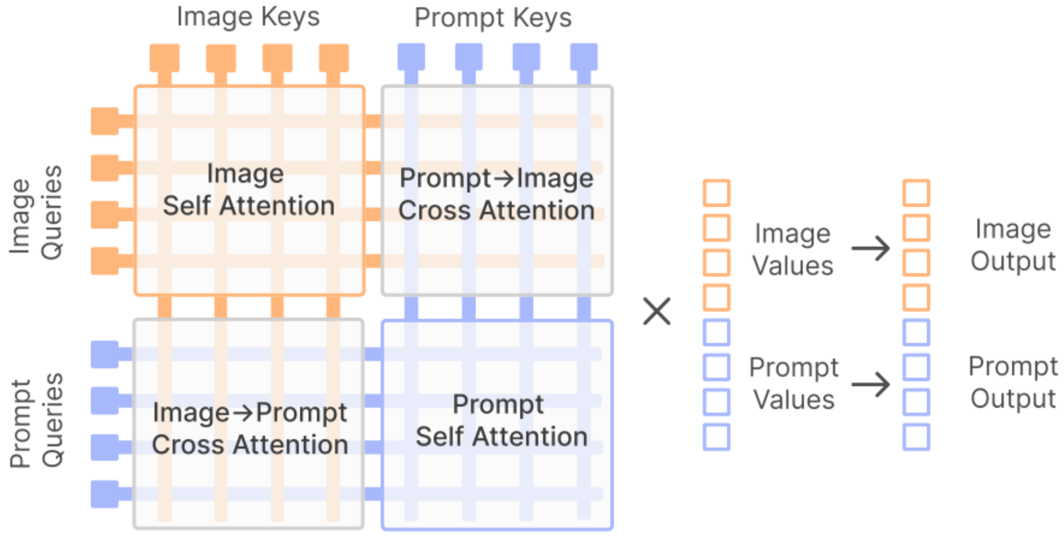
Single-Directional Concept Cross-Attention

- Multi-Modal Self-Attention:

$$k_{xp} = [K_x x_1, \ldots, K_p p_1 \ldots]$$

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

(a) Multi-modal Attention



PEKING UNIVERSITY 北京大學

Single-Directional Concept Cross-Attention

- Multi-Modal Self-Attention:

$$k_{xp} = [K_x x_1, \ldots, K_p p_1 \ldots]$$

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

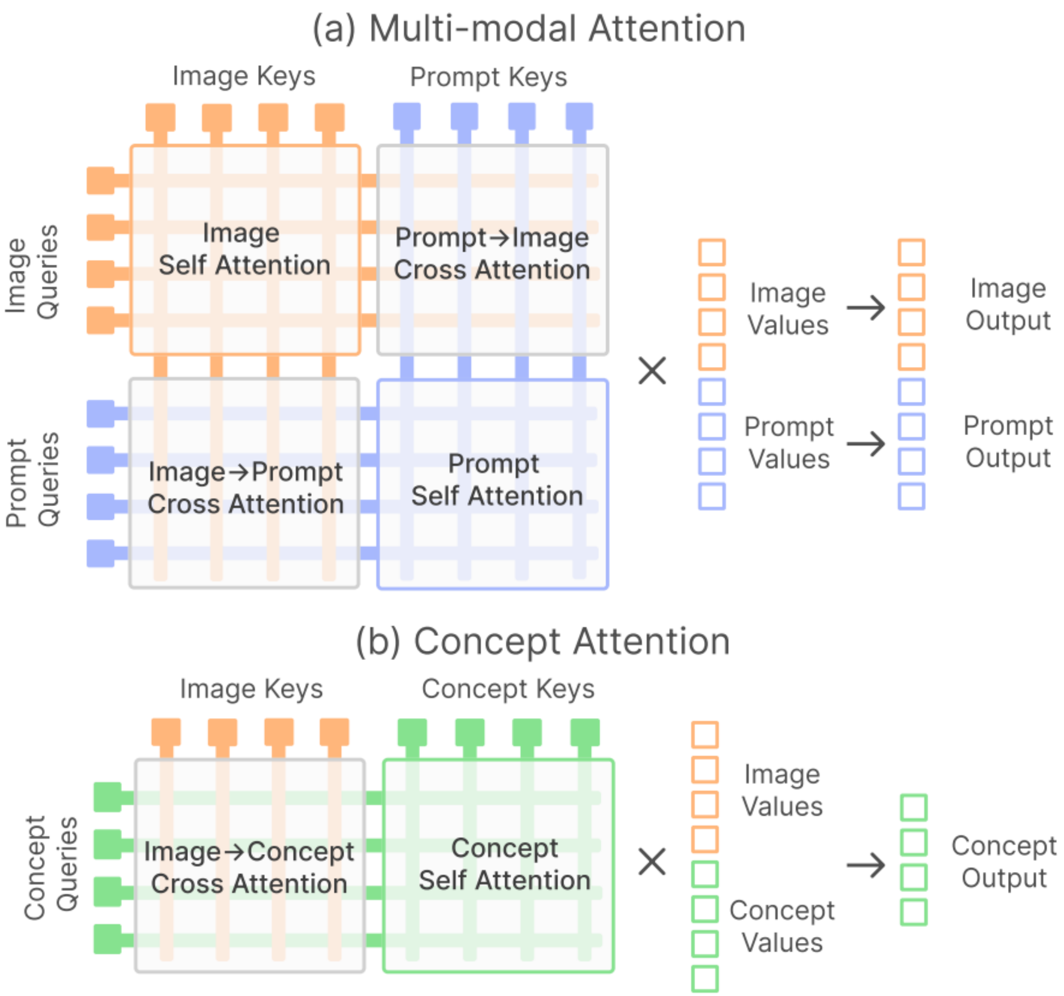
- Concept Cross-Attention:

$$q_c = [Q_p c_1, \ldots]$$

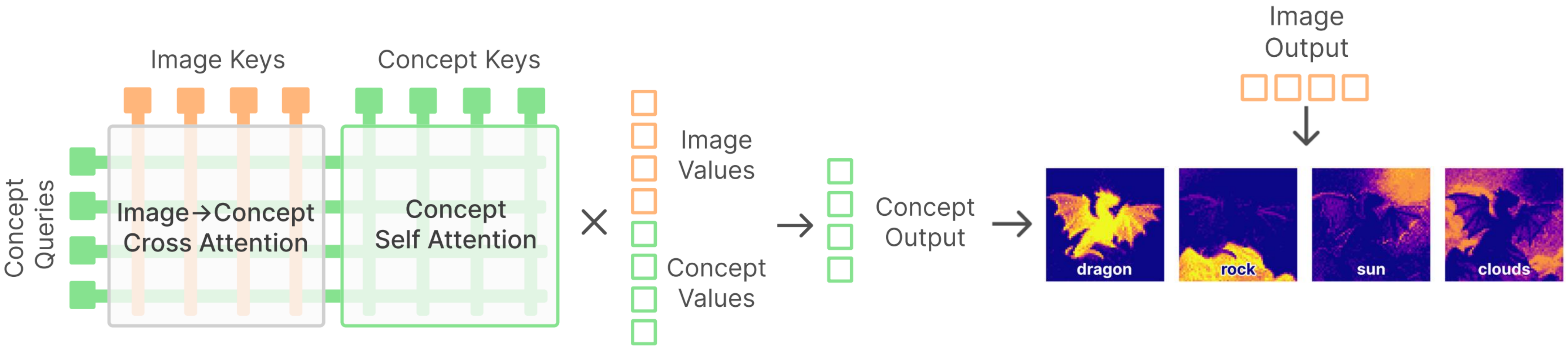
$$k_{xc} = [K_x x_1 \ldots, K_x x_n, K_p c_1 \ldots, K_p c_r]$$

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

$$o_c = \mathrm{softmax}(q_c k_{xc}^T) v_{xc}$$



Producing Saliency Map



$$o_c = \mathrm{softmax}(q_c k_{xc}^T) v_{xc}$$

$$\phi(o_x, o_c) = \mathrm{softmax}(o_x o_c^T).$$

Method

北京大學 PEKING UNIVERSITY

Pseudo-code

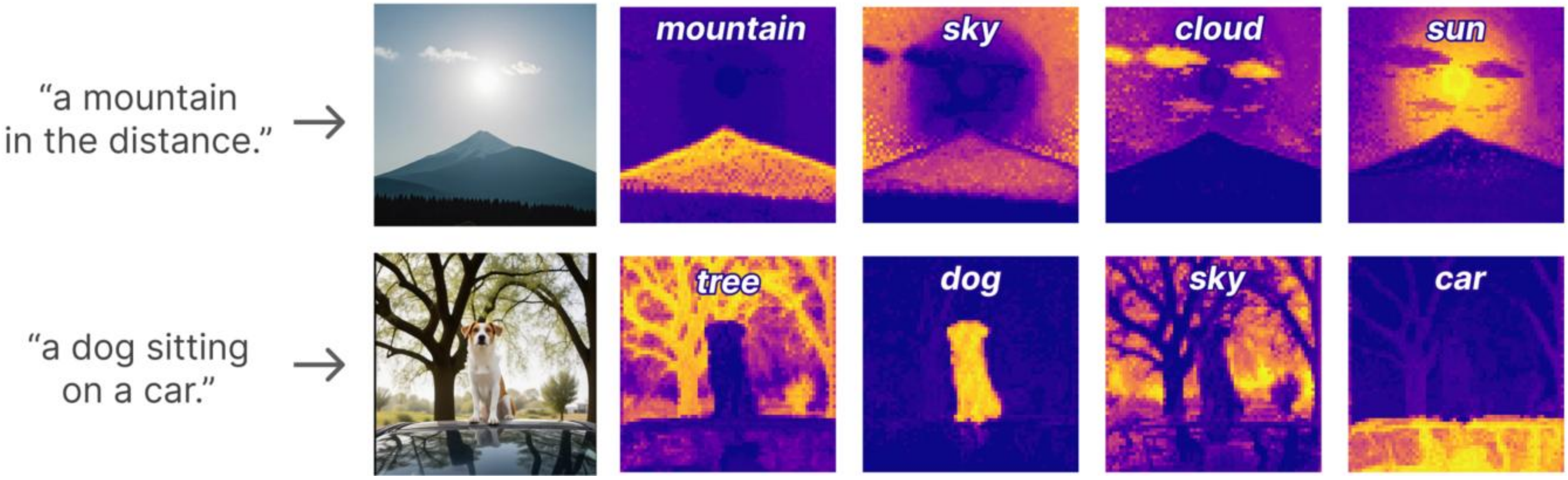
(a) Multi-Modal Attention

```
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

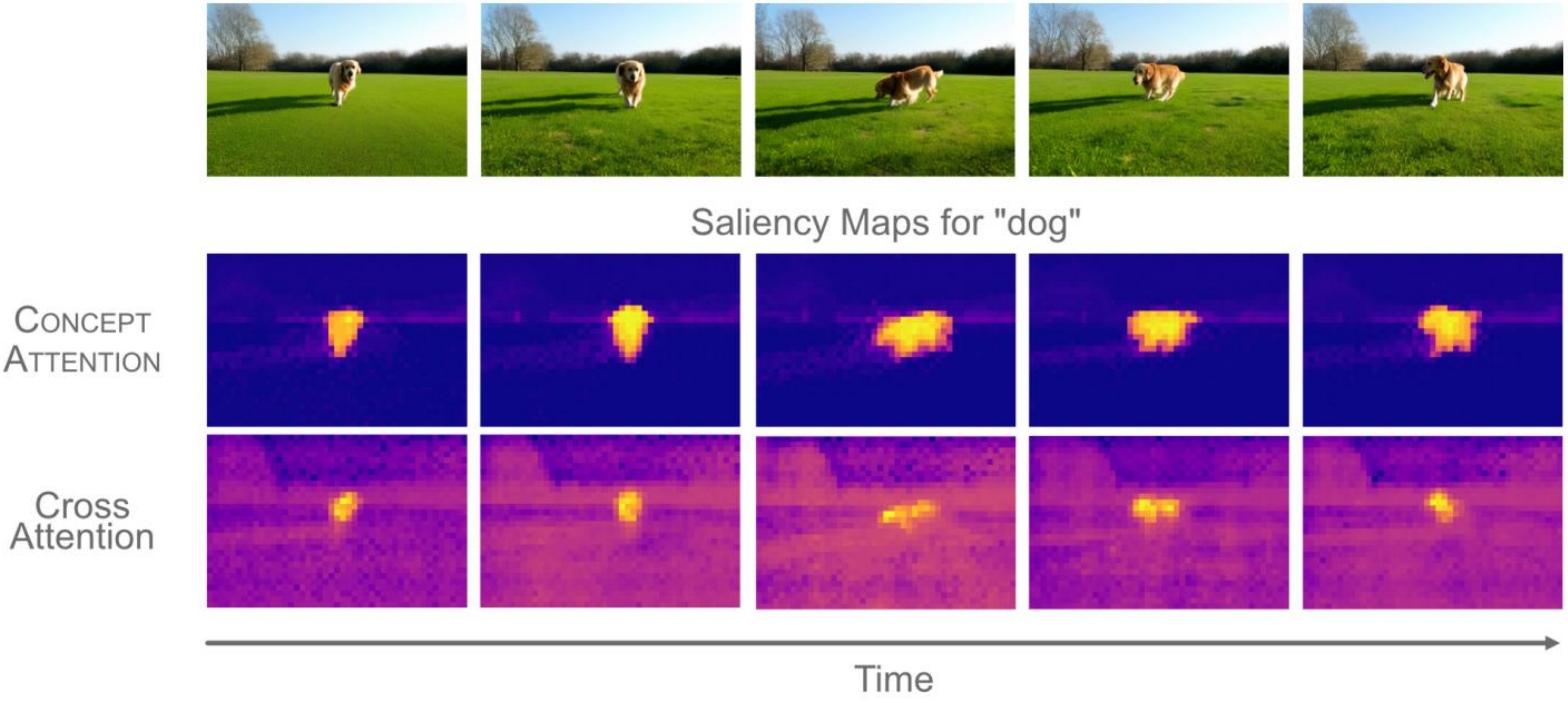
```
+ def multi_modal_attn_with_concept_attn(img, txt, concepts):
      # Compute the keys, queries, and values
      img_k, img_q, img_v = img_projection(img)
      txt_k, txt_q, txt_v = txt_projection(txt)
+     concept_k, concept_q, concept_v = txt_projection(concepts)
      # Concat the image and 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
```

Produced Saliency Maps



PEKING UNIVERSITY 北京大學

Comparison with Cross-Attention on Videos



Experiments

Quantitative Benchmarks:

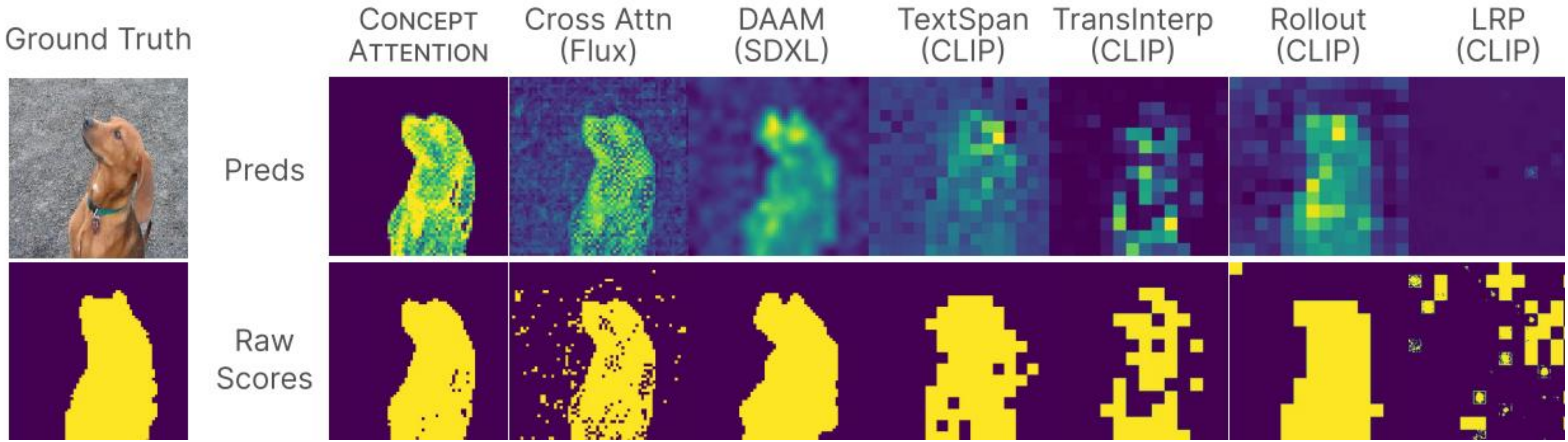
- Zero-shot semantic segmentation
- ImageNet-Segmentation & PascalVOC 2012

Baselines:

- Interpretation results on Transformer encoder features
- Attention and Interpretation on Unet-SDs
- Raw cross-attention in DiTs

Experiments

Open Vocabulary Semantic Segmentation Result



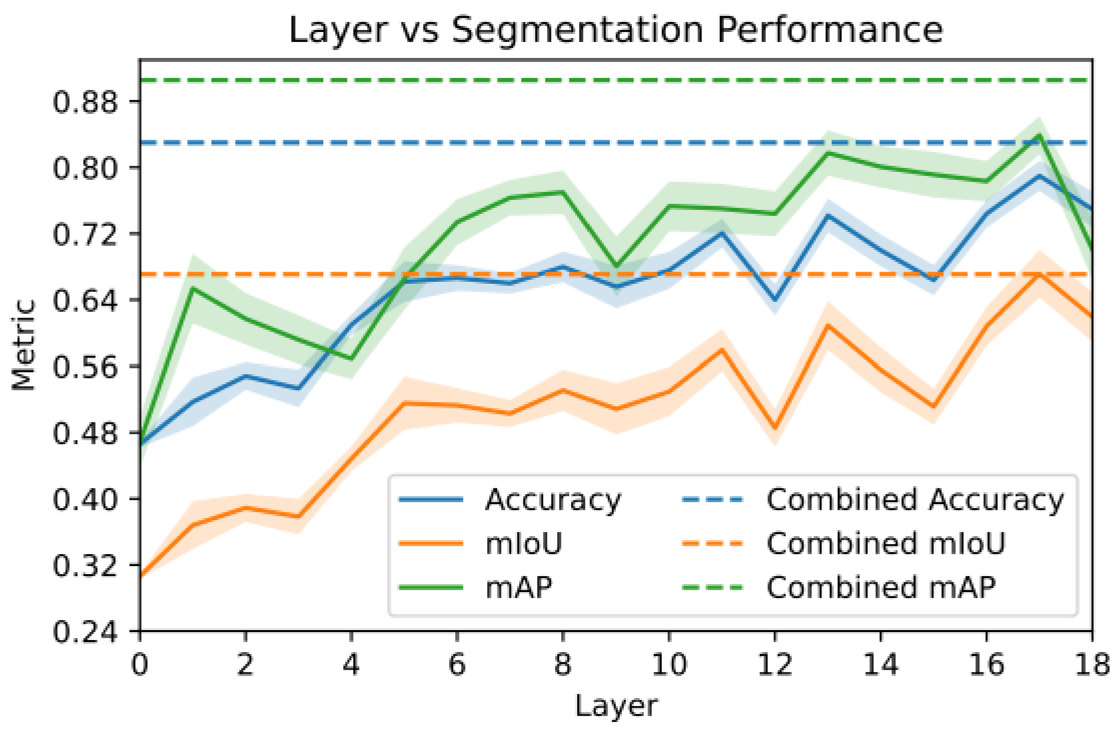
北京大學 PEKING UNIVERSITY

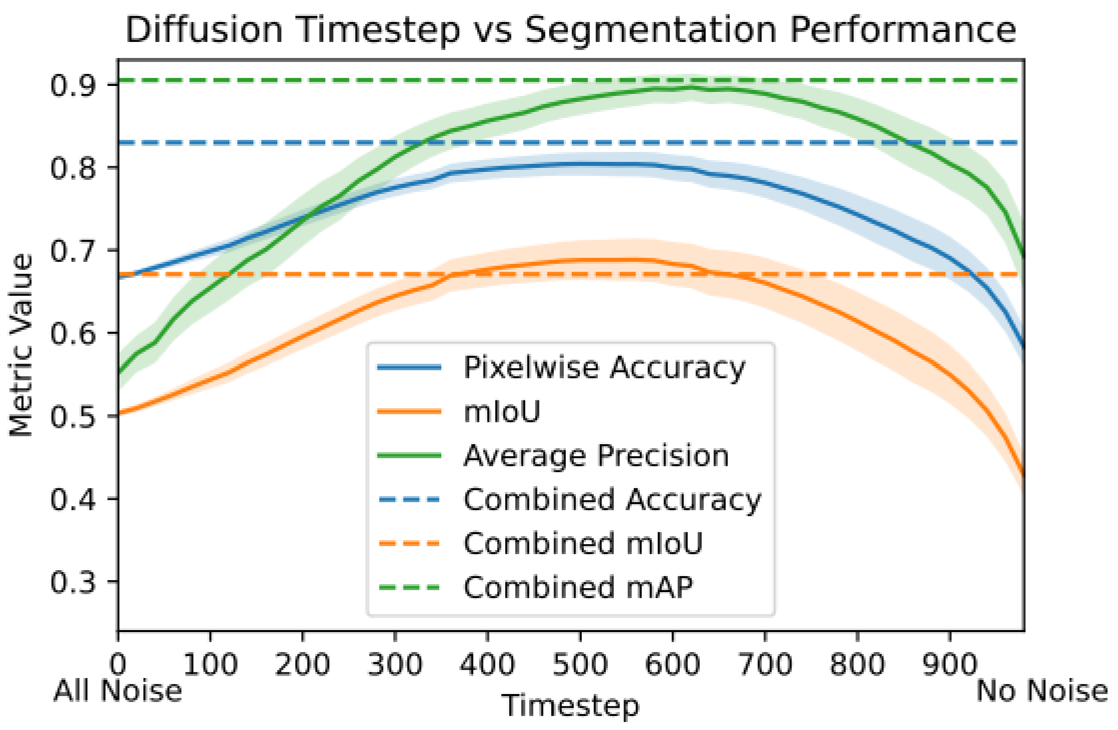
Zero-Shot Semantic Segmentation Performance

Method	Architecture	ImageNet-Segmentation Acc ↑	mIoU↑	mAP↑	PascalVOC (Single Class) 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**

Experiments

Ablation on Different Layers and Timesteps





PEKING UNIVERSITY

Algorithm Ablations

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**

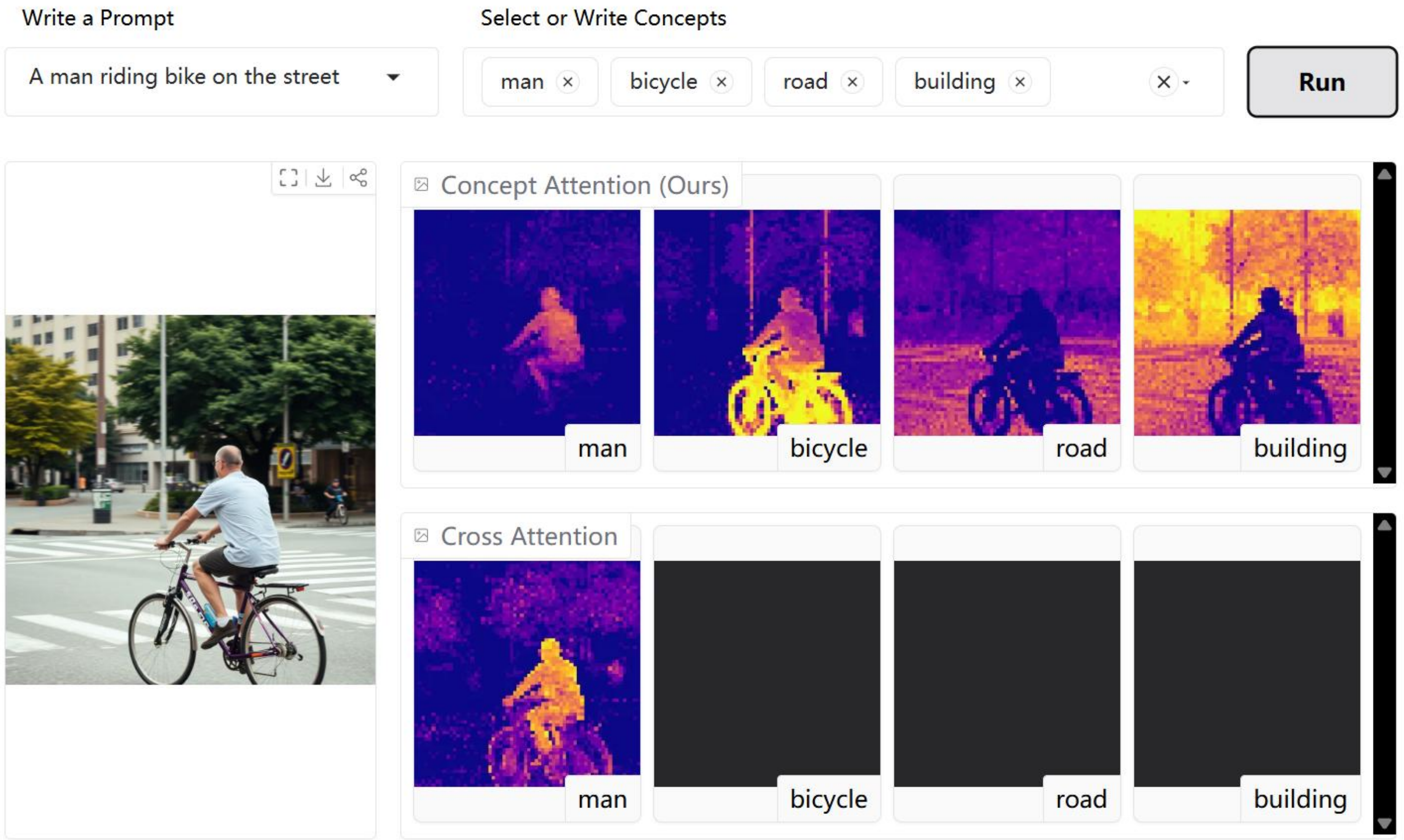
Space to Obtain Saliency Map

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**

Concept Utilization

Experiments

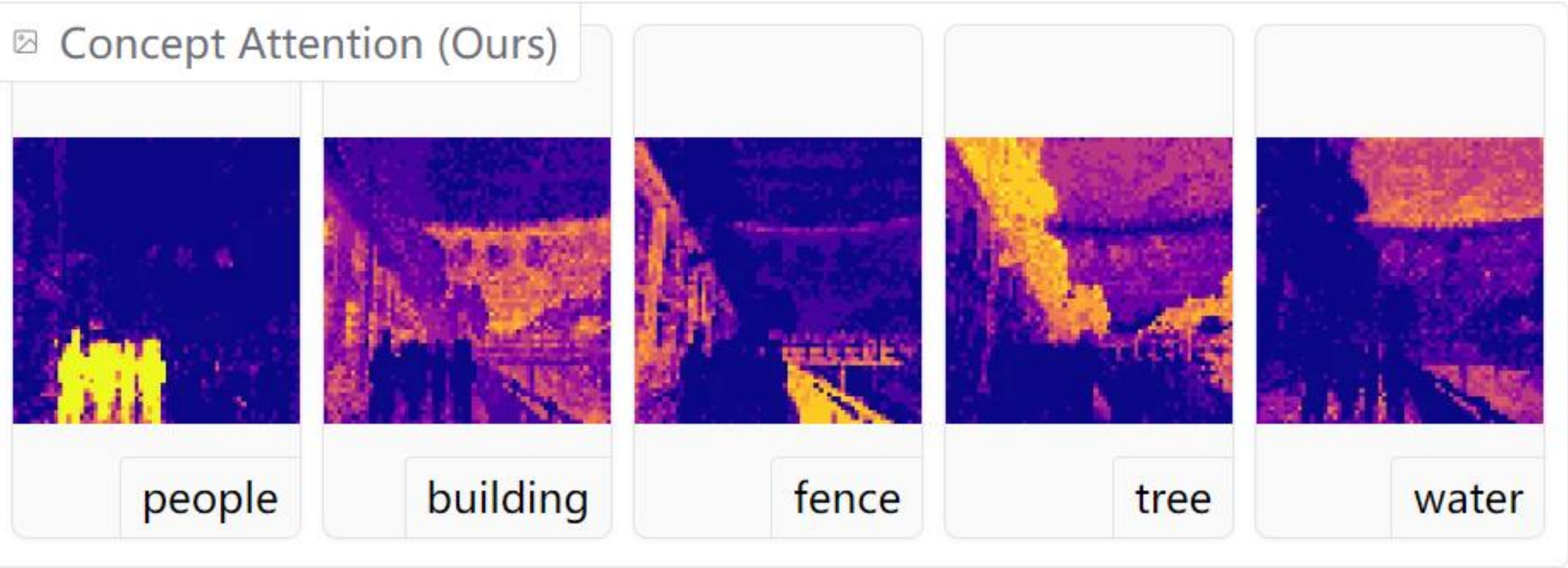
Demo



https://huggingface.co/spaces/helblazer811/ConceptAttention/

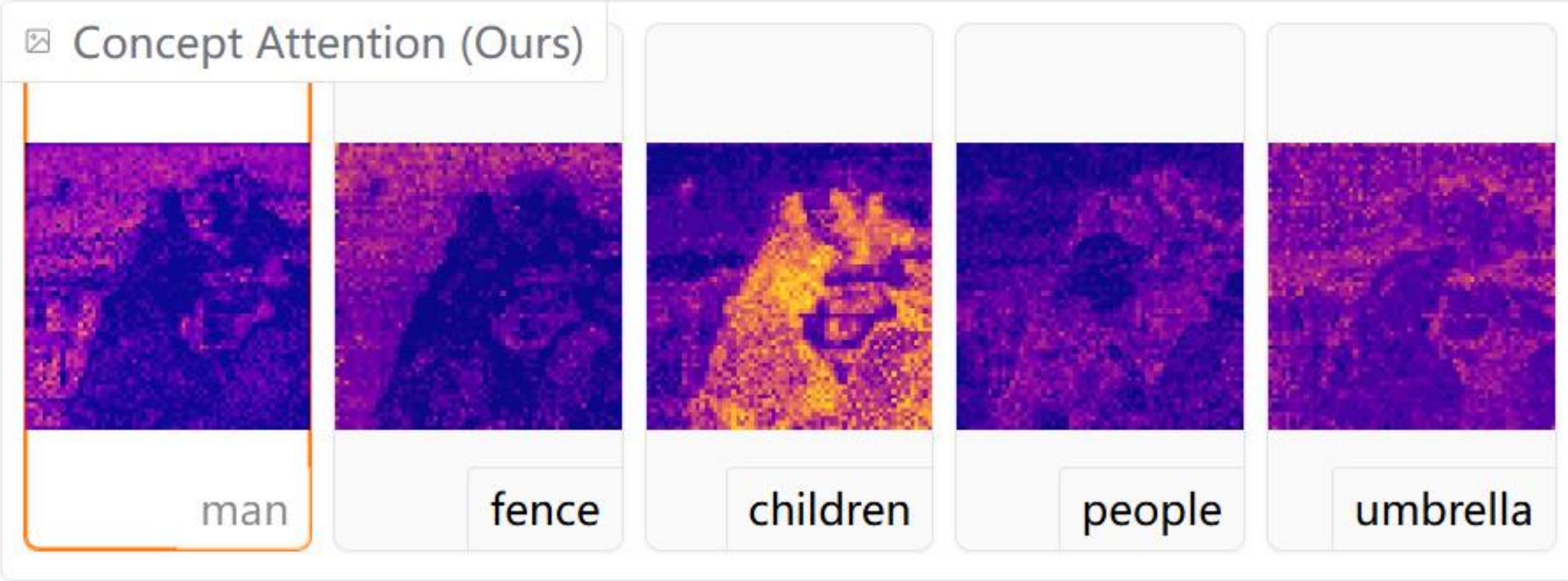
Experiments

Demo



Limitations

Zero-Shot Low Resolution Results:

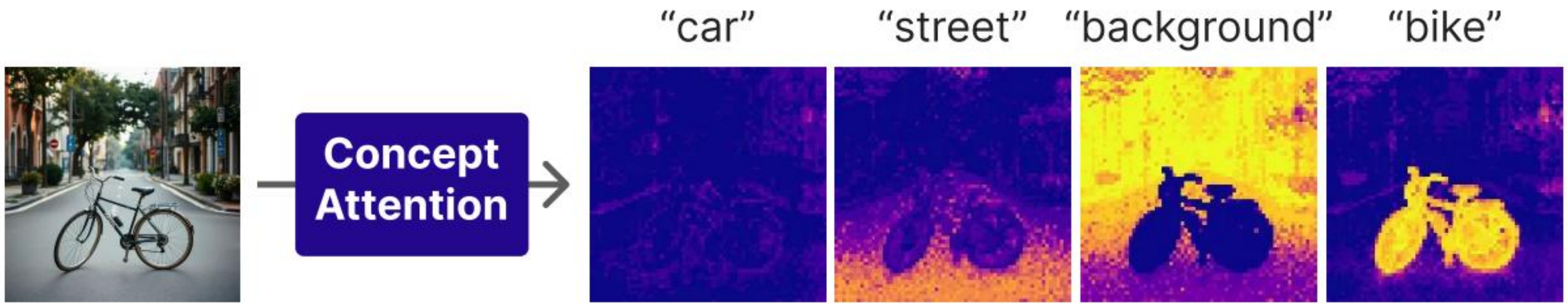


Limitations

Correct concept “bike” chosen over similar concept “car” when both are given

Concepts

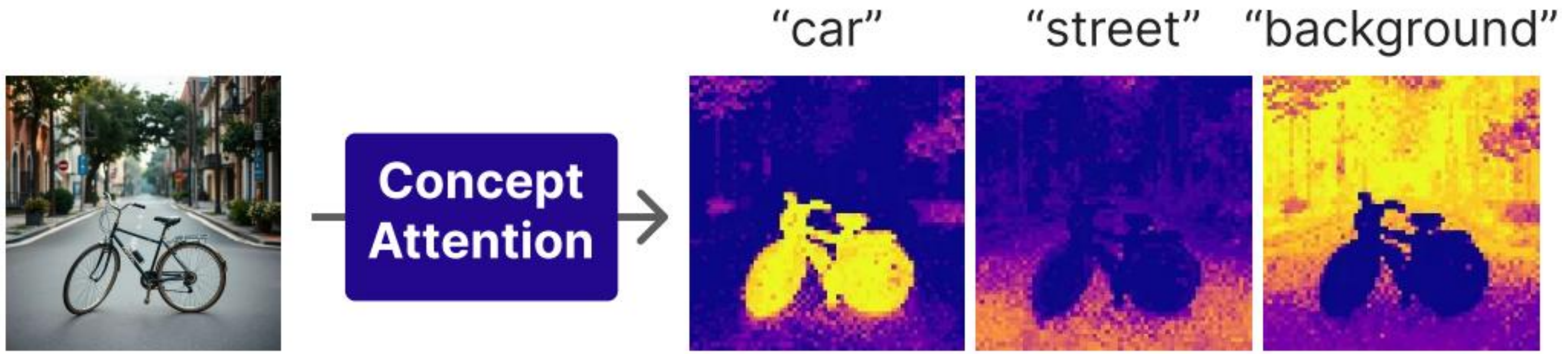
“car”
“bike”
“background”
“street”
\+



Closest concept “car” chosen when correct concept “bike” is not present

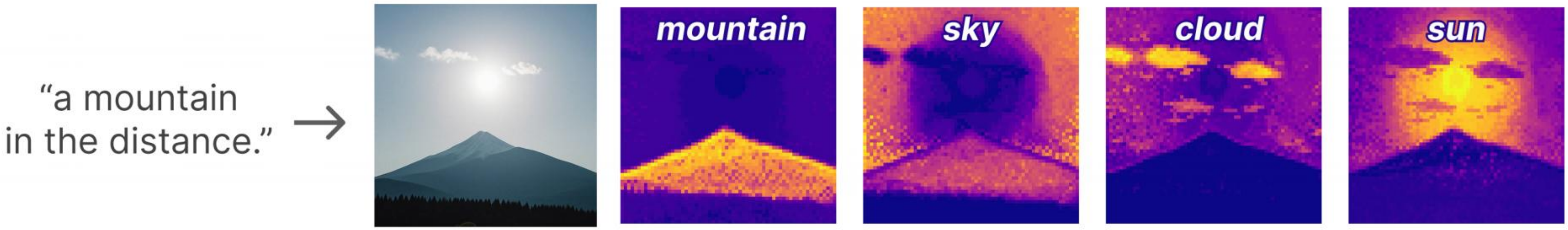
Concepts

“car”
“background” +
“street”



Limitations

Cannot Deal with Overlapping Concepts:



Conclusion

- **Rich semantic representation in MM-DiT Attention**
- **A training-free approach to extract saliency maps**
- **Excels on open-vocabulary semantic segmentation**
- **Fails on LQ data domain**

Thanks for listening!

Presenter: Jinyi Luo
2025.07.28