What the DAAM:
Interpreting Stable Diffusion Using Cross Attention

Raphael Tang Linqing Liu Akshat Pandey Zhiying Jiang Gefei Yang et al.
Comcast Applied AI
University College London
University of Waterloo

Abstract

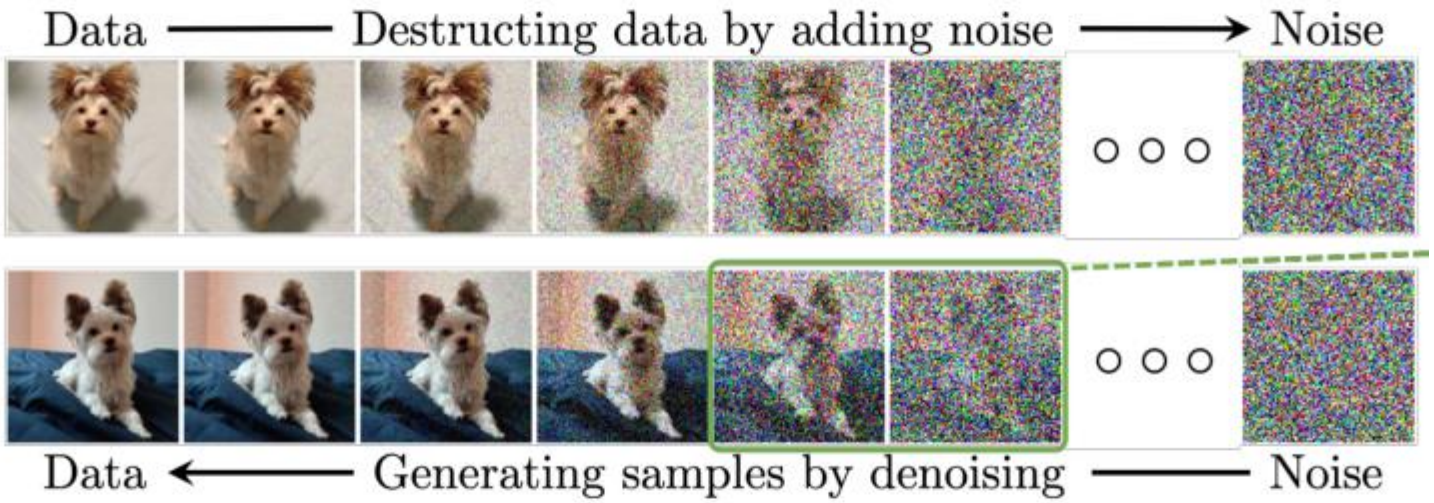
- Find out how individual words from the text prompt affects each output pix el, i.e., understand the conditional part of Stable Diffusion.

Figure 1: The original synthesized image and three DAAM maps for “monkey,” “hat,” and “walking,” from the prompt, “monkey with hat walking.”

Background: Diffusion Models

- Generative Model capable of state-of-the-art generation of pictures, videos, 3D models, etc.

- Consists of a *forward process*, where a datum is progressively noised, and a *reverse process*, where the datum is restored from noise.



Yang, Ling, et al. "Diffusion models: A comprehensive survey of methods and applications.“ *arXiv preprint arXiv:2209.00796* (2022).

Background: Diffusion Models

The *forward process*:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

progressively adds noise, until $q(\mathbf{x}_T) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$

i.e., data is transformed in to zero-mean isotropic Gaussian.

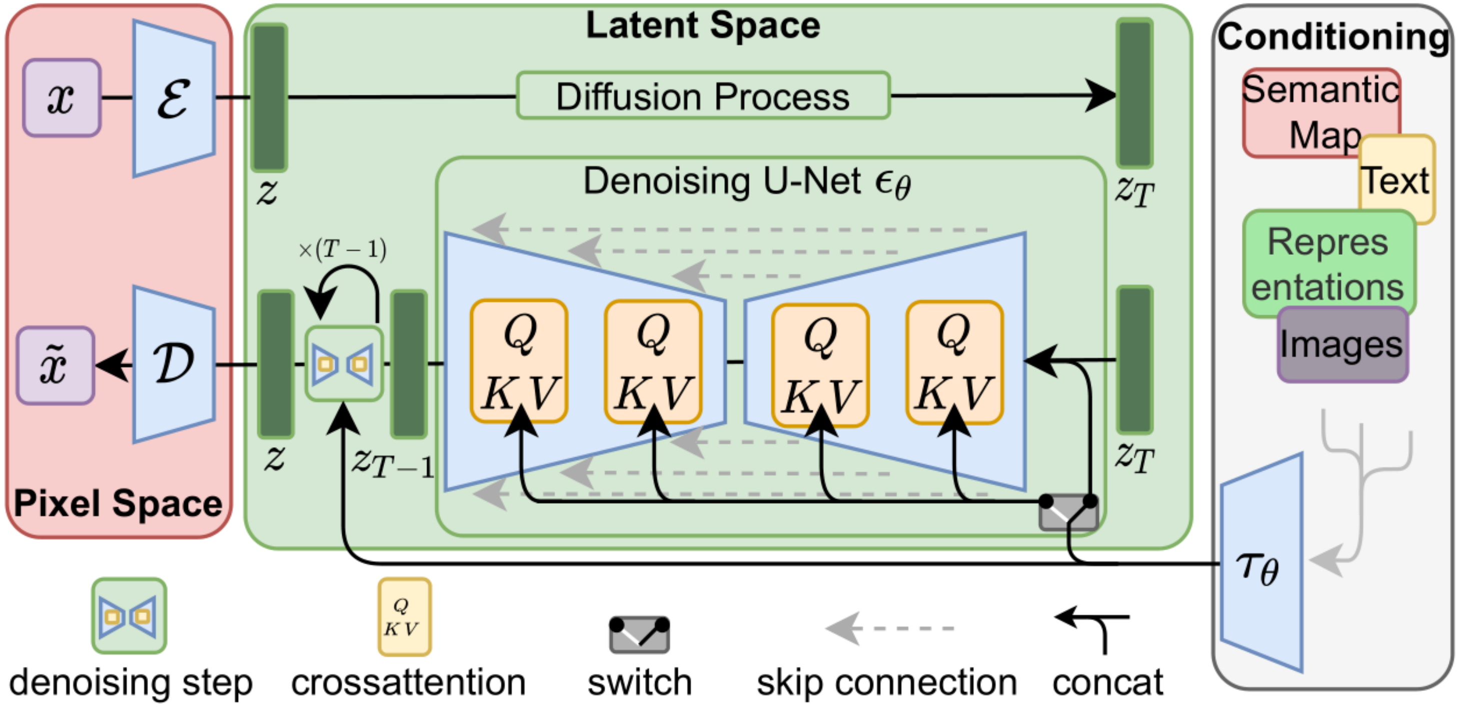
The *reverse process* transforms unit Gaussian noise back to original data by traversing the path backwards using a learnable kernel:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Background: Stable Diffusion

- Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
- Generates high-quality image given a text prompt (and others)

Background: Stable Diffusion



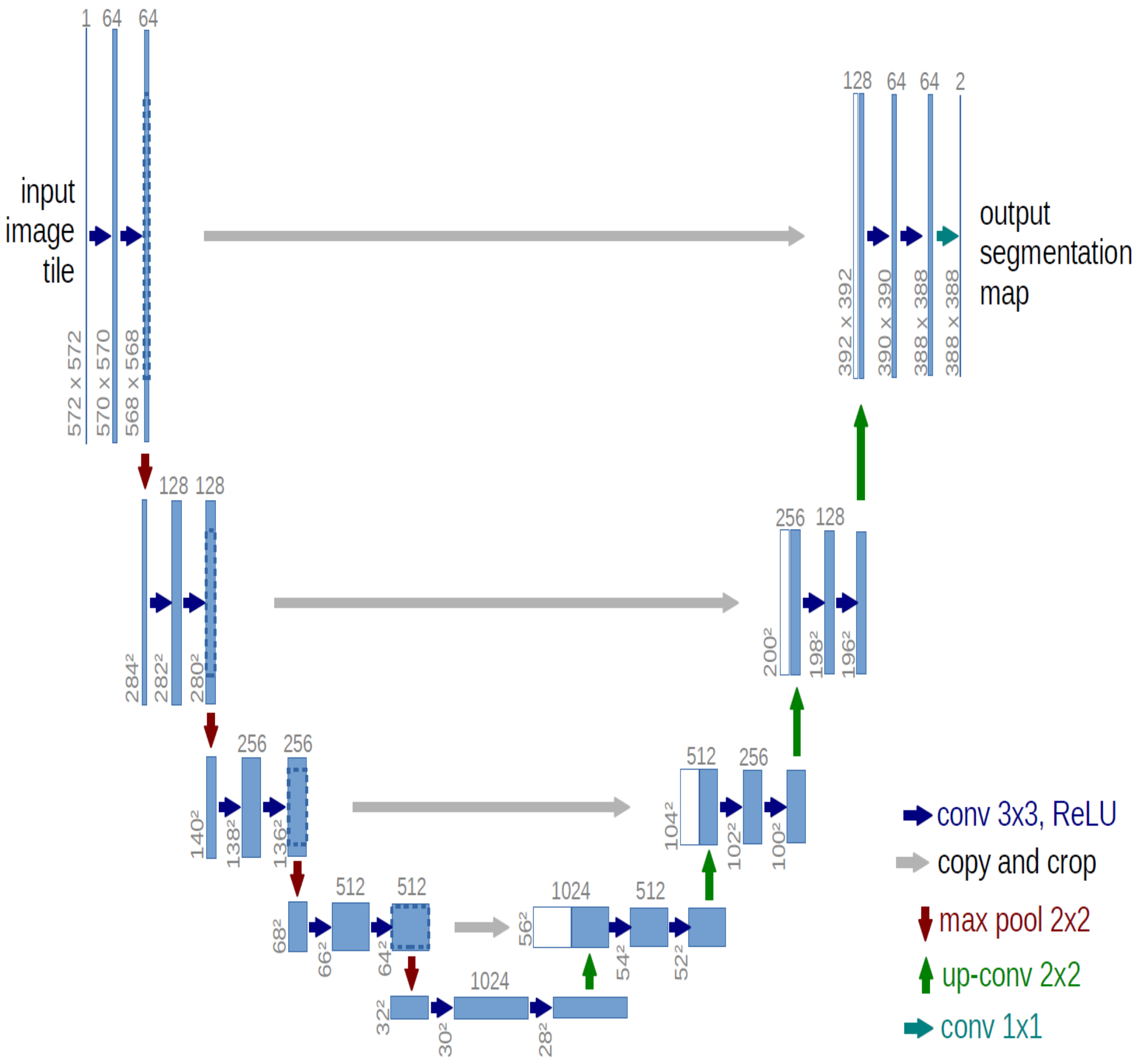
Observations:

- Image is generated in latent space, and restored via Convolutional VAE.
- Denoising kernel is conditioned on the prompt text.

Approach: Denoiser

- Denoiser $\epsilon_\theta(l, t; \boldsymbol{X})$
- Where $\boldsymbol{X}$ is a list of word (CLIP) emb eddings: $\boldsymbol{X} = [\boldsymbol{x}_1; \ldots; \boldsymbol{x}_{l_w}]$

- Denoiser is a convolutional U-Net:
  - Downsampling block i $(i = 1, \ldots, K)$ outp ut: $\boldsymbol{h}^{\downarrow}_{i,t} \in \mathbf{R}^{\left\lceil\frac{w}{c^i}\right\rceil \times \left\lceil\frac{h}{c^i}\right\rceil}$
  - Using multi-headed cross-attention layer : $\boldsymbol{h}^{\downarrow}_{i,t} = F_t^{(i)}(\widehat{\boldsymbol{h}}^{\downarrow}_{i,t}, \boldsymbol{X}) \cdot (W_v^{(i)} \boldsymbol{X})$, i.e., att. sc ores btw $\widehat{\boldsymbol{h}}^{\downarrow}_{i,t}$ (Q) are calculated for each word embedding $\boldsymbol{X}$ (K, V).



Approach: Denoiser

- Due to the convolutional nature of the U-Net and the VAE, we can upscale each U-Net block to original image size and aggregate them:

$$D_k^{\mathbb{R}}[x, y] := \sum_{i,j,\ell} \tilde{F}^{(i)\downarrow}_{t_j,k,\ell}[x, y] + \tilde{F}^{(i)\uparrow}_{t_j,k,\ell}[x, y], \qquad (6)$$

k: word, l: head

- Thresholded for segmentation tasks:

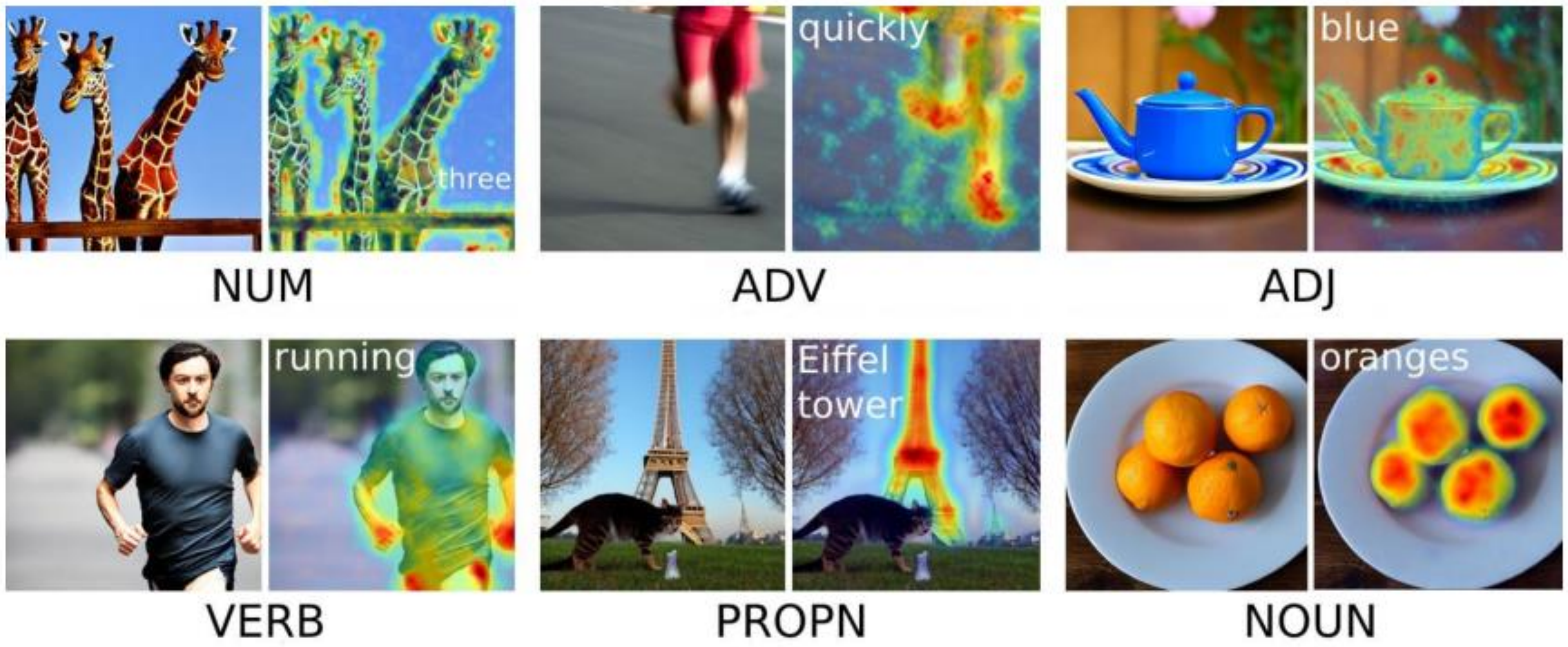
$$D_k^{\mathbb{I}_\tau}[x, y] := \mathbb{I}\left(D_k^{\mathbb{R}}[x, y] \geq \tau \max_{i,j} D_k^{\mathbb{R}}[i, j]\right), \qquad (7)$$

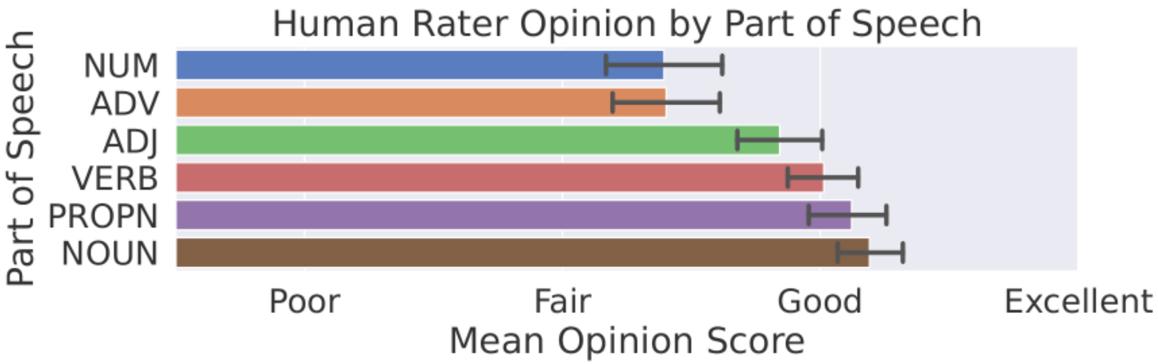
Result: Object Attribution

- Synthesize images based on COCO image captions dataset, hand-seg ment each **noun**, and compare results with image segmentation models:

#	Method	COCO-Gen mIoU^{80}	COCO-Gen mIoU$^{\infty}$	Unreal-Gen mIoU^{80}	Unreal-Gen mIoU$^{\infty}$
	Supervised Methods				
1	Mask R-CNN (ResNet-101)	82.9	32.1	76.4	31.2
2	QueryInst (ResNet-101-FPN)	80.8	31.3	78.3	35.0
3	Mask2Former (Swin-S)	**84.0**	32.5	**80.0**	36.7
4	CLIPSeg	78.6	**71.6**	74.6	**70.9**
	Unsupervised Methods				
5	Whole image mask	20.4	21.1	19.5	19.3
6	PiCIE + H	31.3	25.2	34.9	27.8
7	STEGO (DINO ViT-B)	35.8	53.6	42.9	54.5
8	Our DAAM-0.3	64.7	59.1	59.1	**58.9**
9	Our DAAM-0.4	**64.8**	**60.7**	**60.8**	58.3
10	Our DAAM-0.5	59.0	55.4	57.9	52.5

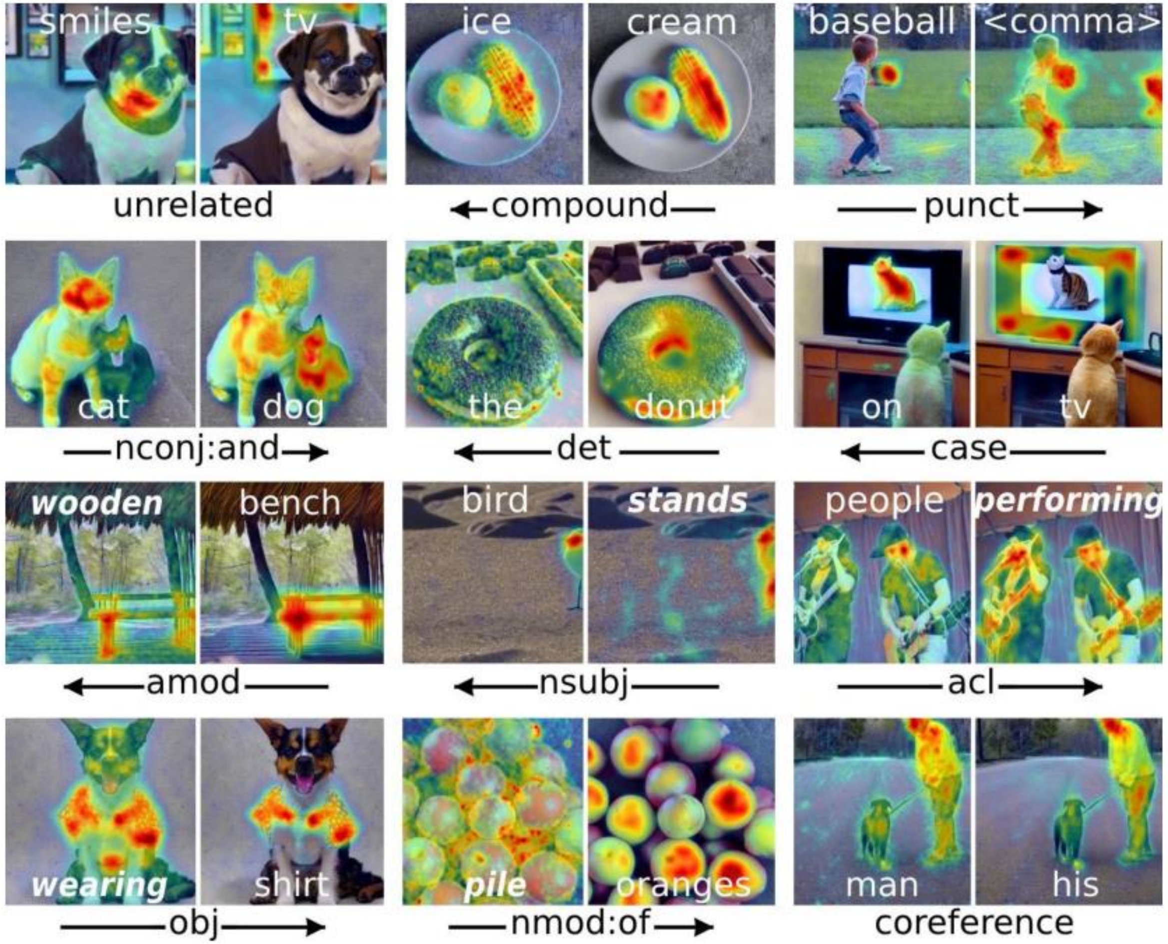
Result: Generalized Attribution





Visuosyntactic Analysis

- How syntax relates to generated pixels by measuring mIoU ($\frac{|A \cap B|}{|A \cup B|}$), mIoD ($\frac{|A \cap B|}{|A|}$), and mIoH ($\frac{|A \cap B|}{|B|}$).



Visuosemantic Analysis: Cohyponym Entanglement

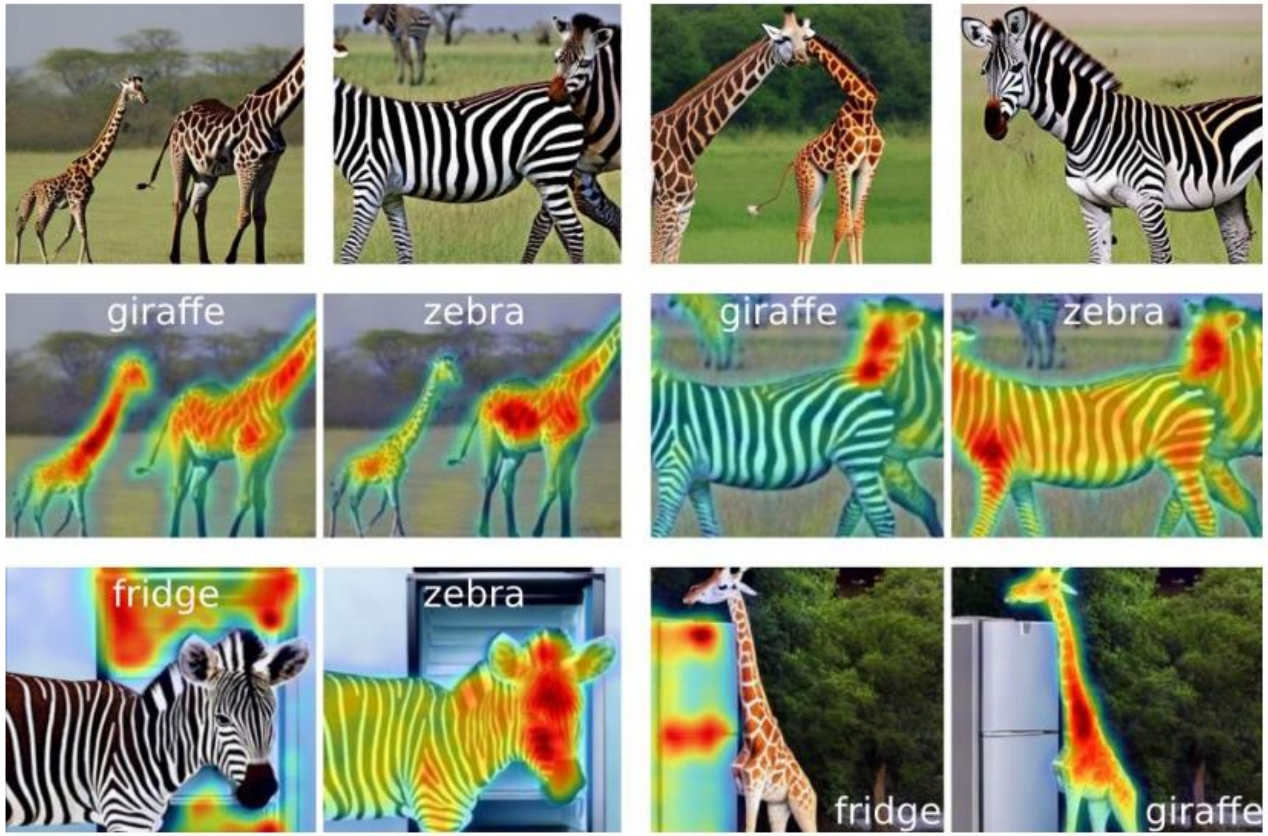


Figure 7: Rows starting from the top: generated images for cohyponyms "a giraffe and a zebra," heat maps for the first two images, and heat maps for non-cohyponymic zebra–fridge and giraffe–fridge prompts.

Visuosemantic Analysis: Adjectival Entanglement

Figure 8: First row: a DAAM map for “rusty” and three generated images for “a <adj> shovel sitting in a clean shed;” second row: a map for “bumpy” and images for “a <adj> ball rolling down a hill.”

Conclusions

- Study visuolinguistic phenomena in diffusion models by interpreting word-pixel cross-attention maps, and the attribution method is proven correct us ing experiments.
- Find feature entanglement.