What the DAAM: Interpreting Stable Diffusion Using Cross Attention

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

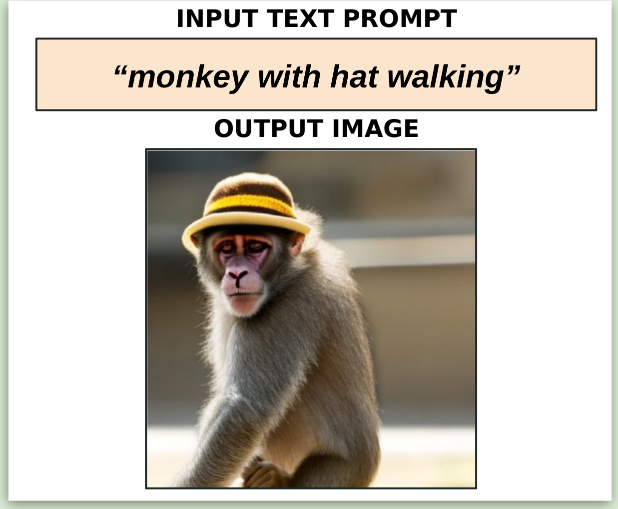
Presented by: Alex Lin, Jason Jabbour, Mark Mazumder

Outline

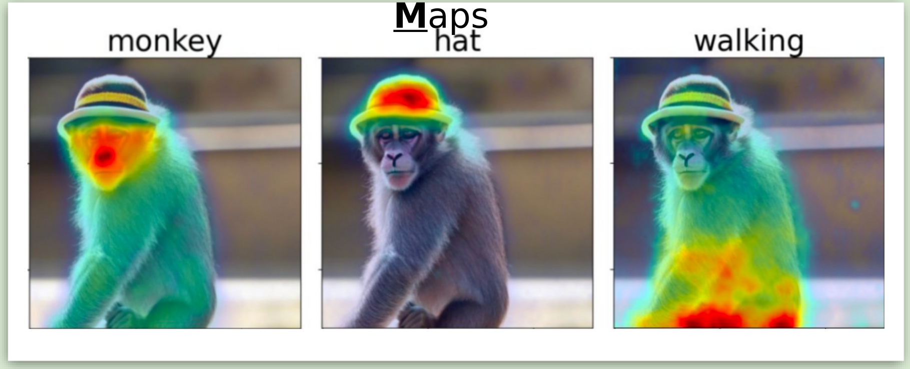
- Related work
- Background: (simplified) Stable Diffusion
- **DAAM** for text-to-image **attribution**
- Results
  - Attribution quality analysis
  - Syntax to pixels
  - Entanglement
- Limitations & discussion

DAAM estimates per-pixel attribution for each word in a prompt (post-hoc)

Stable Diffusion: Text to Image



DAAM: **D**iffusion **A**ttentive **A**ttribution **M**aps



Related Work

Cross-Attention

Vision-Language (VQA…)

Generative Models (GANs..)

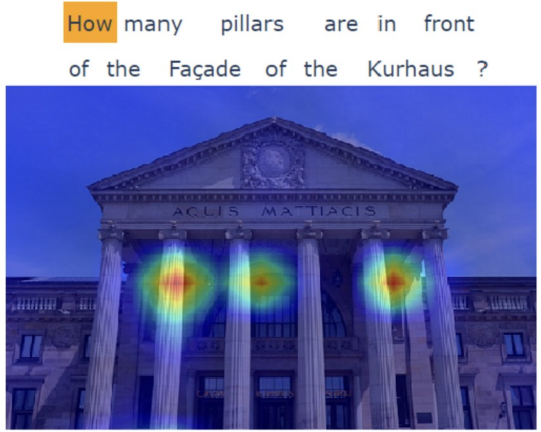
Attention is Not Explanation [Jain et al.]

Attention is Not *Not* Explanation [Wiegreffe et al.]

This work applies existing techniques (cross-attention) to an **open source, SOTA** diffusion model to <u>probe **limitations**</u>

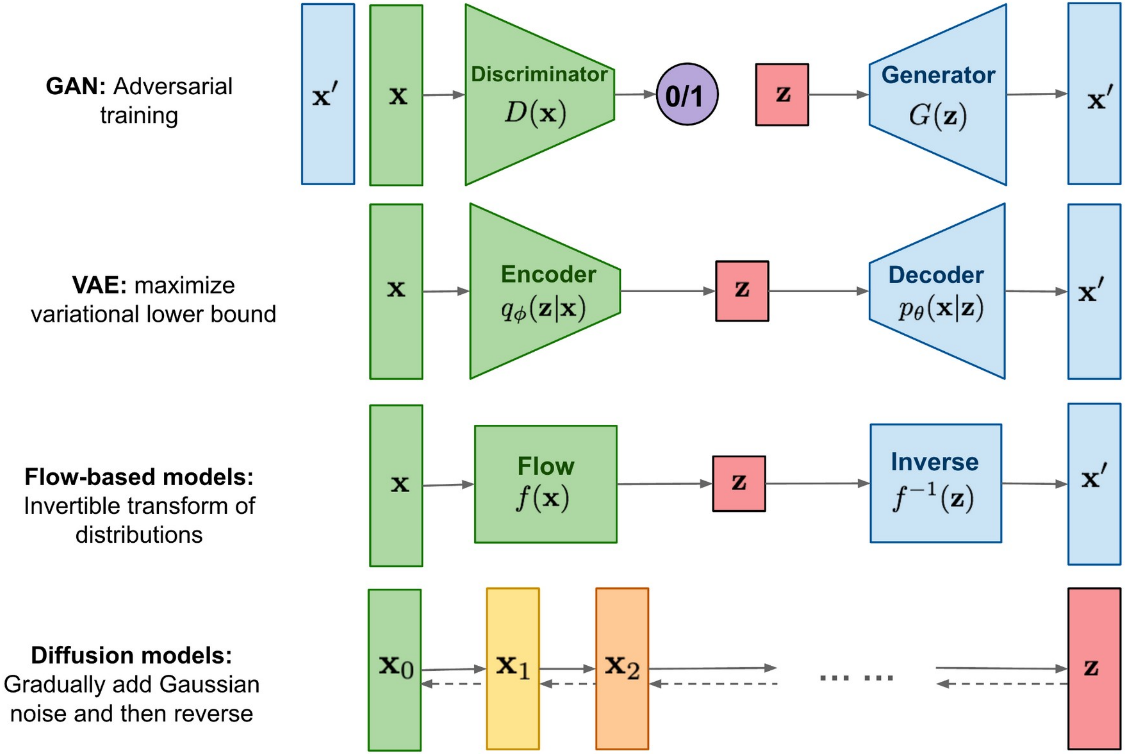
- **Textual perturbation** (Wallace et al., 2019), Attentional Visualization (Vig, 2019; Kovaleva et al., 2019, Shimoaka et al., 2016) and information bottlenecks (Jiang et al., 2020) to relate **important input tokens** to **outputs of large language models**
- Probing **vision transformers for verb understanding** (Hendricks and Nematzadeh, 2021)
- Enhancing diffusion models using **prompt engineering** (Hertz et al, 2020; Woolf, 2020)
- **Disentangling** e.g., style and spelling (Karras et al., 2019; Materzynska et al., 2022)

<u>[2203.17247] VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers</u> [Aflalo et al]



(b) **Predicted**: *There are 6 pillars in front of the Façade of the Kurhaus.*

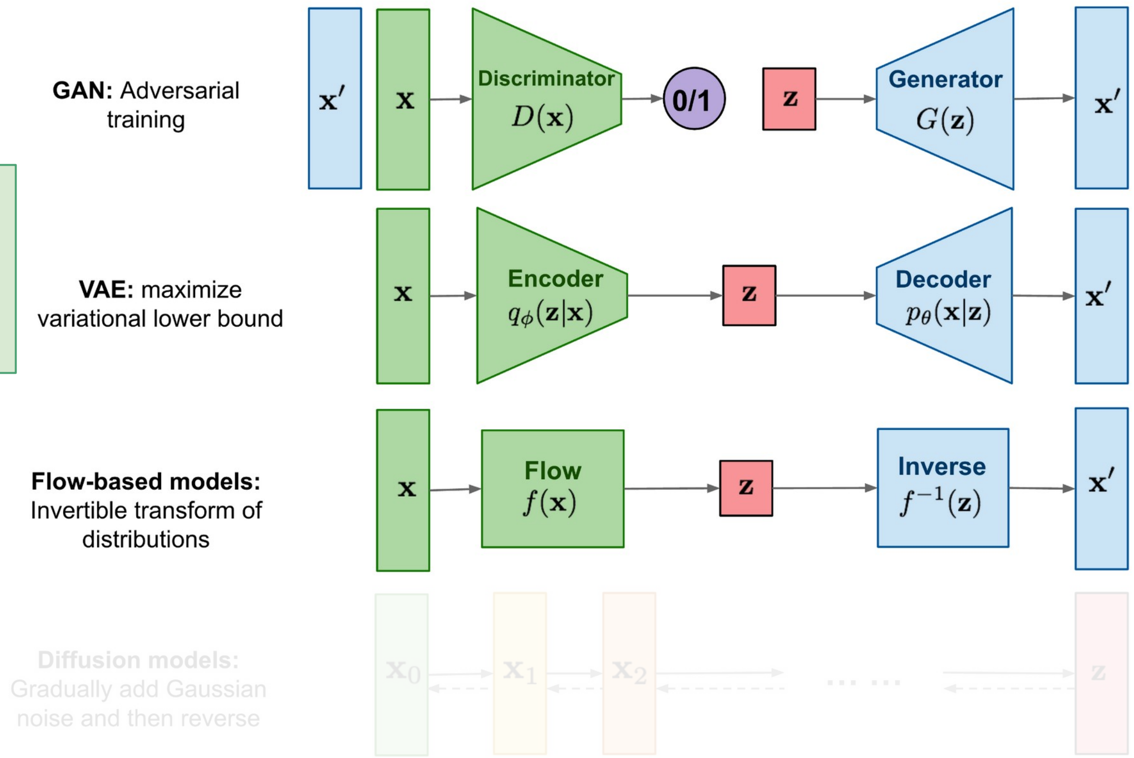
Diffusion Models vs: GANs, VAEs, Flow



Source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Diffusion Models vs: GANs, VAEs, Flow

“single-pass” encoder-decoder models

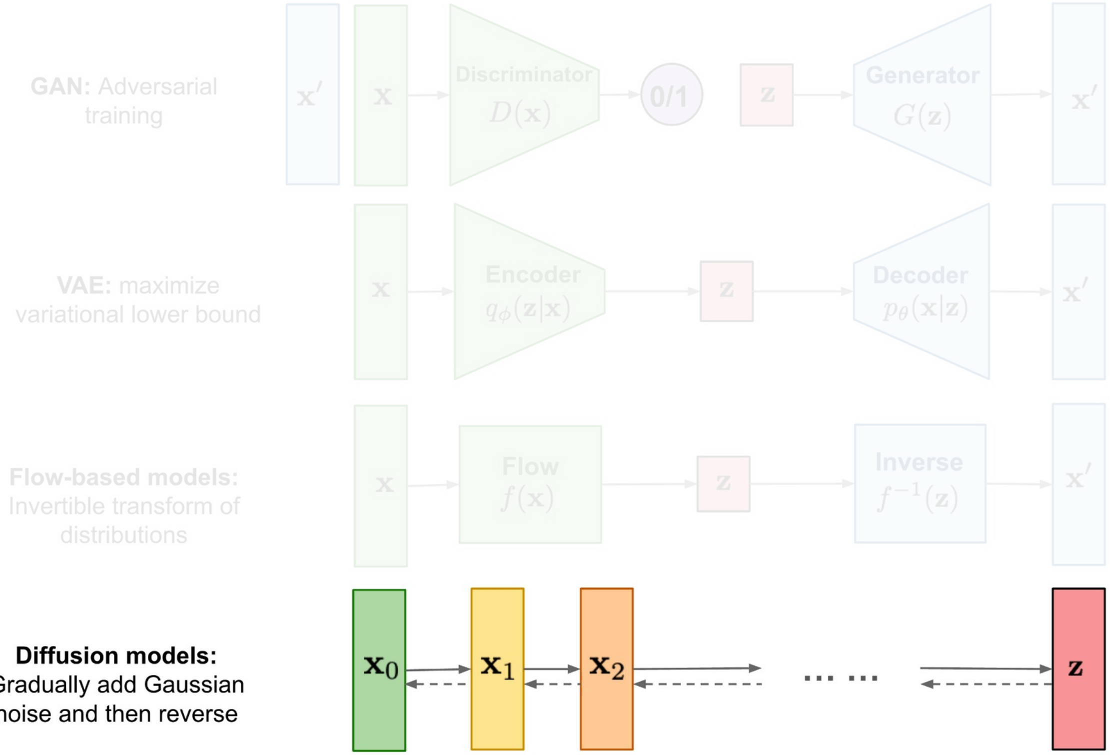


Source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Diffusion Models vs: GANs, VAEs, Flow

“single-pass” encoder-decoder models

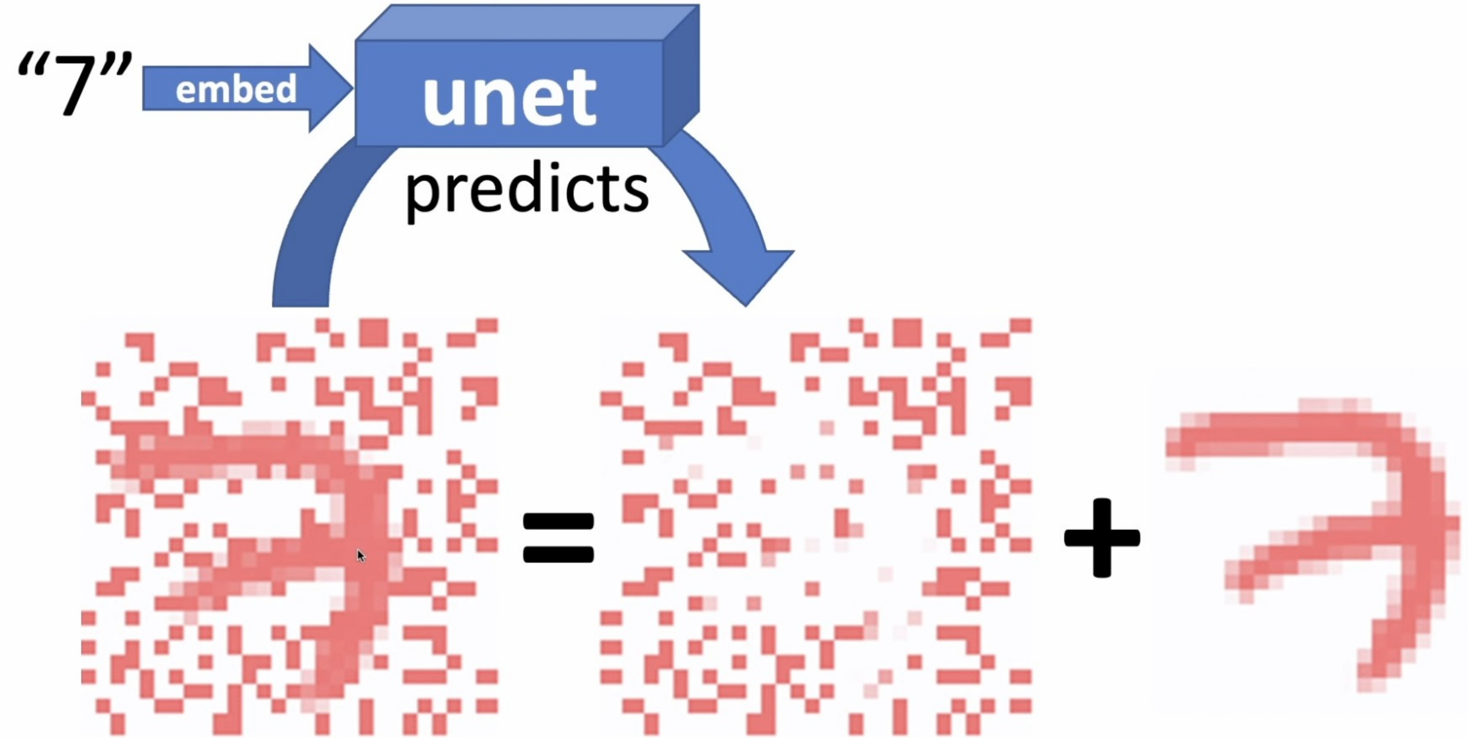
Multi-step time-series



Source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Image Generation: Predict and subtract noise

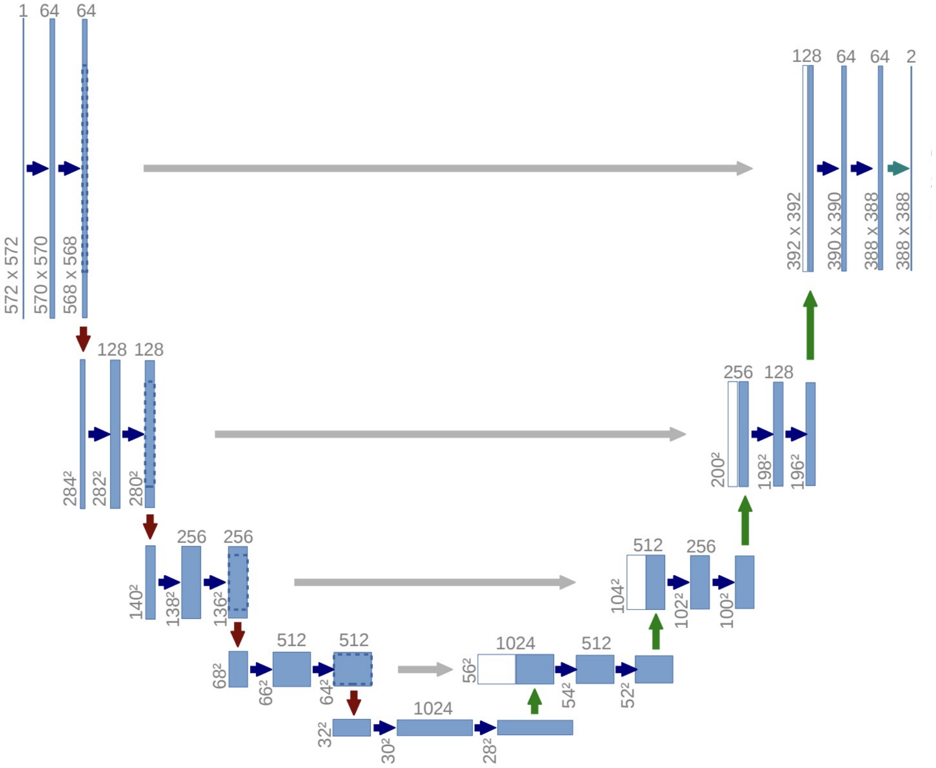
Conditioning:



Source:
[Lesson 10: Deep Learning Foundations to Stable Diffusion, 2022]

UNet Architecture

[1505.04597] U-Net: Convolutional Networks for Biomedical Image Segmentation



Forward Process: Noise Schedule

Fixed additive noise model: ϵ

Add noise for time steps *t* in {0, …, 30}

Image Source: [Diffusion models from scratch in PyTorch]

Reverse Process: Learned Noise Estimate

Fixed additive noise model: ϵ

Add noise for time steps *t* in {0, ..., 30}

Image Source: [Diffusion models from scratch in PyTorch]

Denoising model: ϵ_θ(image, timestep)

Diffusion Model Loss

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$

Error of **noise reconstruction**

Source: High-Resolution Image Synthesis with Latent Diffusion Models

Diffusion Model Loss

Added noise (from Normal distribution)

Denoising Model

Squared Euclidean Norm

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$

Image at time *t*

Error of **noise reconstruction**

Source: [High-Resolution Image Synthesis with Latent Diffusion Models]

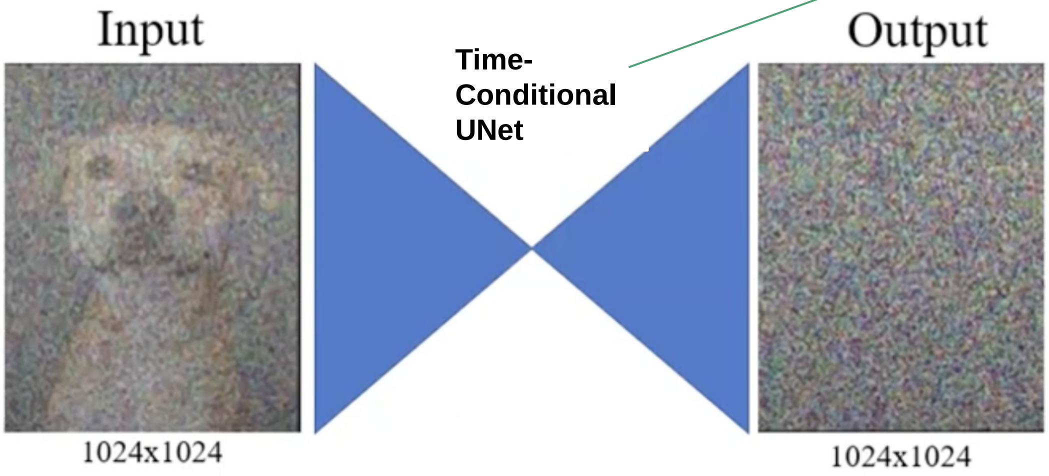
Diffusion Model Loss

Added noise (from Normal distribution)

Denoising Model

Squared Euclidean Norm

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$



Time-Conditional UNet

Image at time t

Reasoning about noise in pixel space is challenging

Source: High-Resolution Image Synthesis with Latent Diffusion Models

Diffusion Model Loss

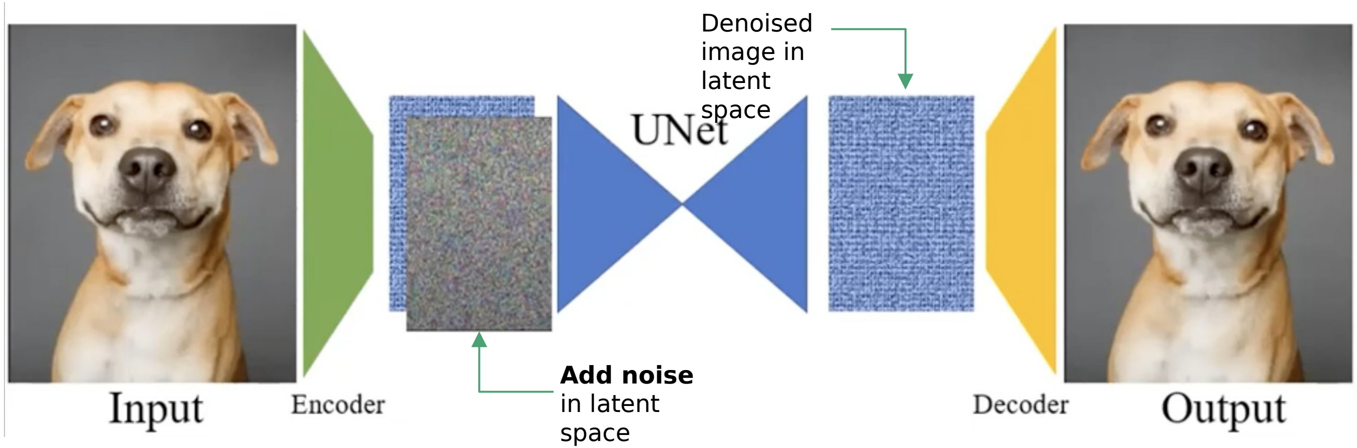
Added noise (from Normal distribution)

Denoising Model

Squared Euclidean Norm

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$

Image at time t



Denoised image in latent space

Add noise in latent space

Easier to estimate noise in latent space!

Source: High-Resolution Image Synthesis with Latent Diffusion Models

Latent Diffusion Model Loss

Added noise (from Normal distribution)

Denoising Model

Squared Euclidean Norm

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$

Image at time *t*

$$L_{LDM} := \mathbb{E}_{\underbrace{\mathcal{E}(x)}_{\text{Latent space encoder}},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]$$

Latent space encoder

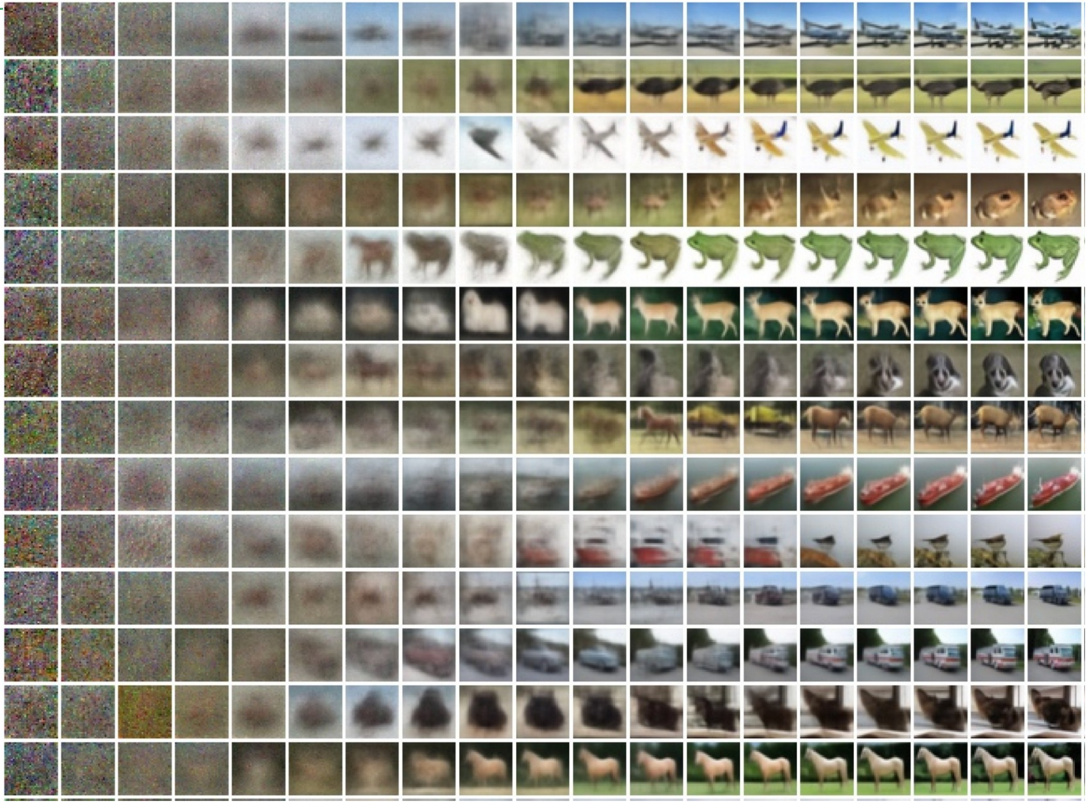
Latent representation of image at time *t*

Source: [High-Resolution Image Synthesis with Latent Diffusion Models]()

Unconditional Sampling from Noise

https://arxiv.org/abs/2006.11239

Random Noise

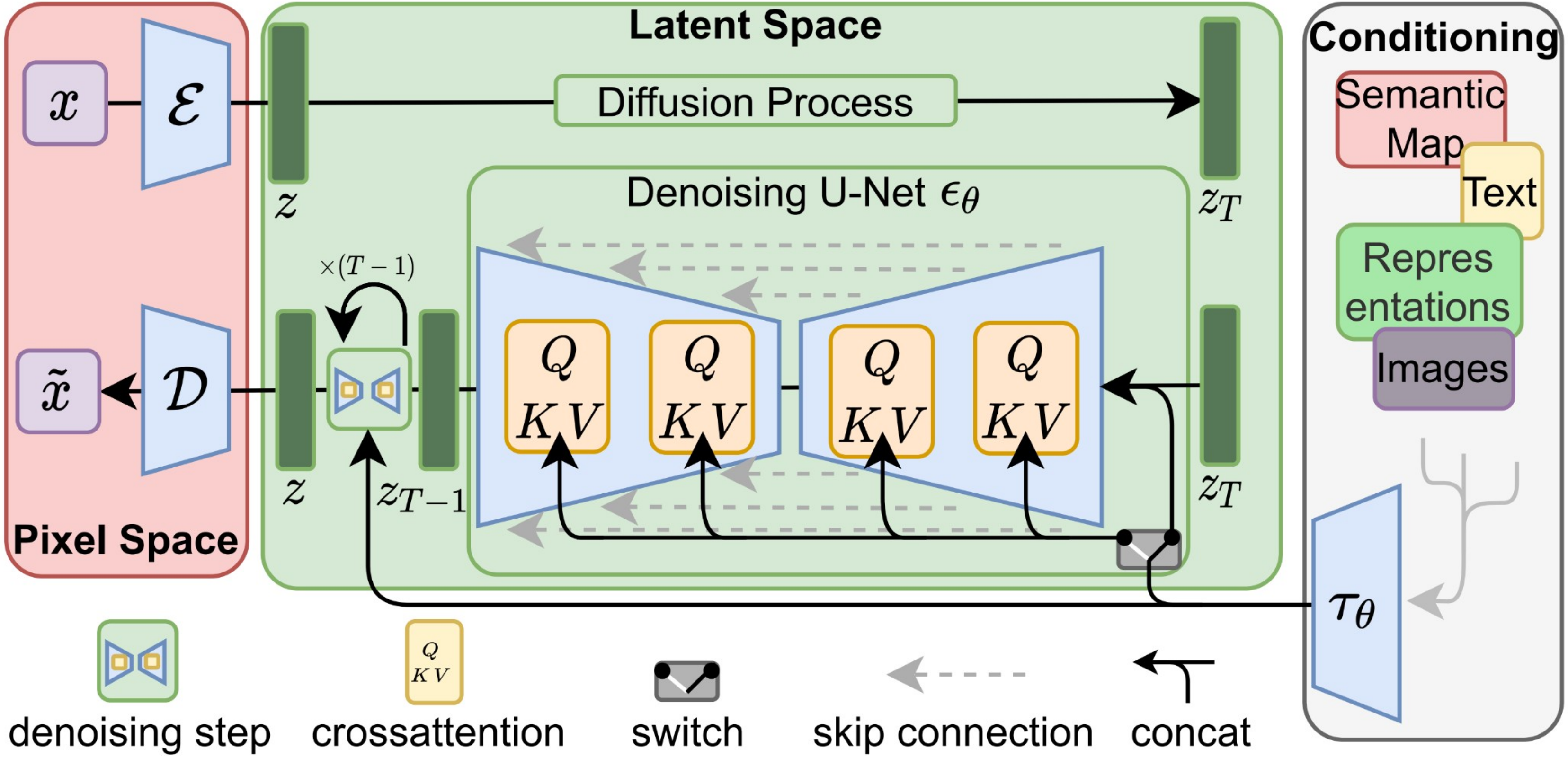


Conditional Generation?

How do we go from a text prompt to an image?

Stable Diffusion Architecture

[2112.10752] High-Resolution Image Synthesis with Latent Diffusion Models

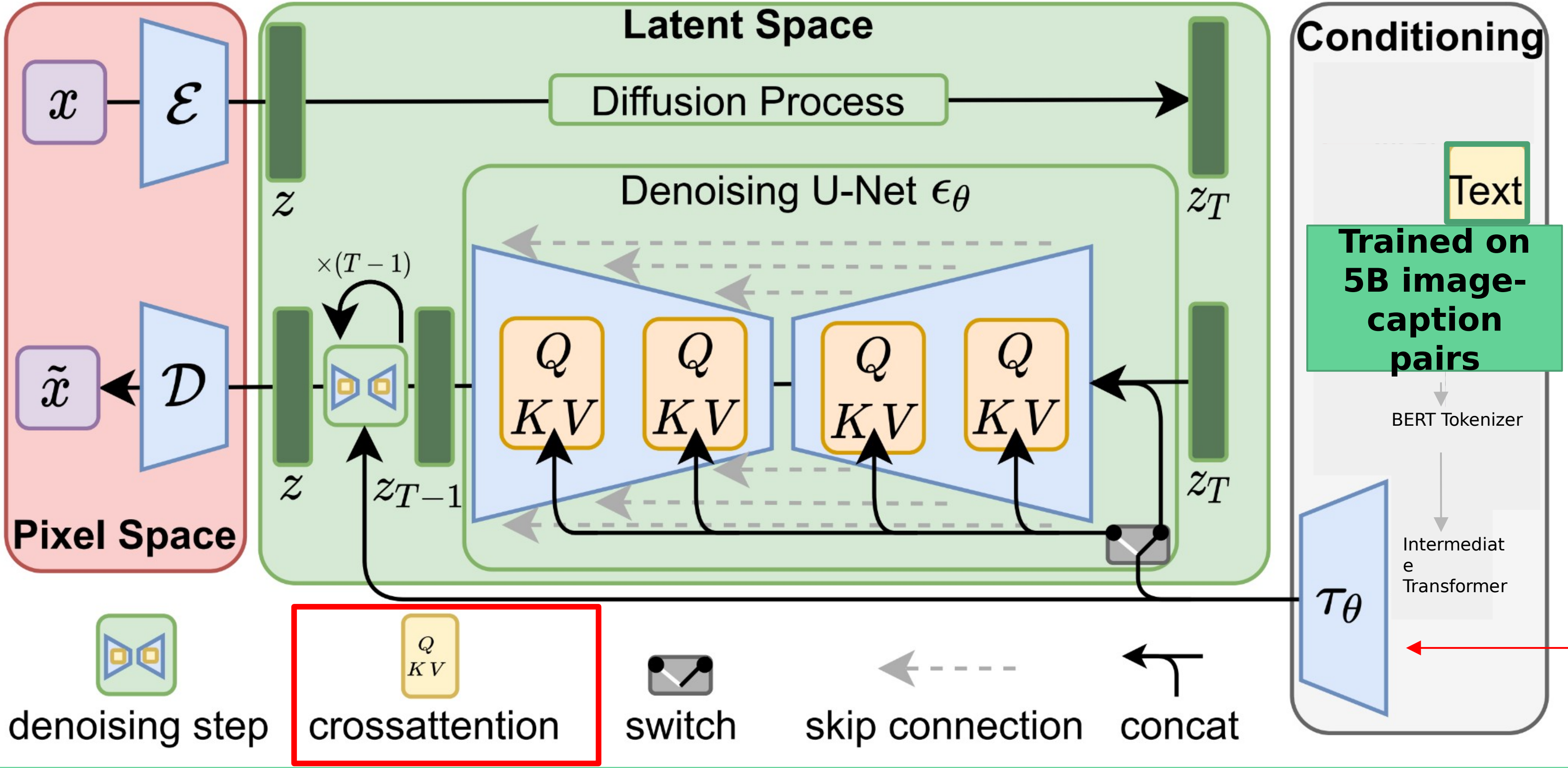


$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, \underbrace{\tau_\theta(y)})\|_2^2\Big]$$

Additionally parameterized on text encoding

Stable Diffusion Architecture

[2112.10752] High-Resolution Image Synthesis with Latent Diffusion Models



Cross-Attention for Text Conditioning

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$$

UNet block *i*

Text encoder

Text encoder

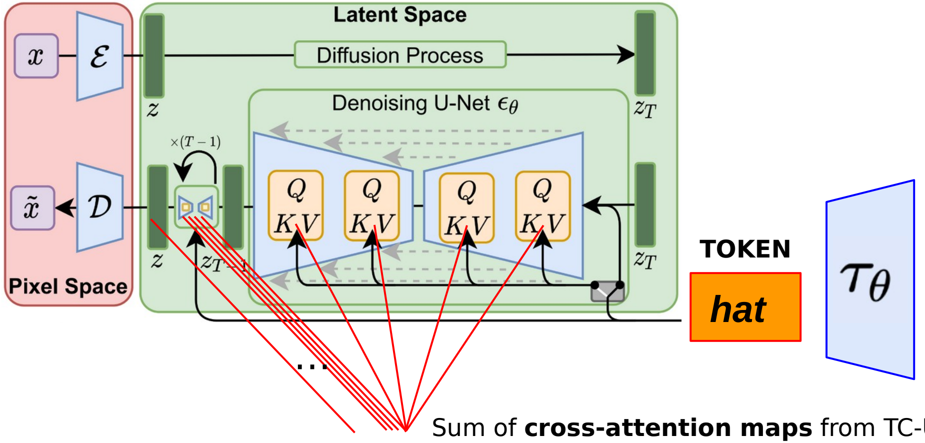
$$\underbrace{Q = W_Q^{(i)} \cdot \varphi_i(z_t)}_{\textbf{Queries: }\text{image tokens}}, \ \underbrace{K = W_K^{(i)} \cdot \tau_\theta(y), \ V = W_V^{(i)} \cdot \tau_\theta(y)}_{\textbf{Keys, Values: }\text{prompt tokens}}$$

Queries: image tokens

Keys, Values: prompt tokens

DAAM estimates **per-word attribution** via this cross-attention **for each subset** of prompt tokens

DAAM Per-Word Heatmaps

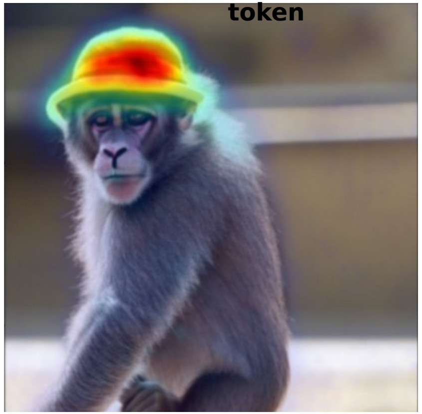


INPUT TEXT PROMPT

"monkey with hat walking"

Sum of **cross-attention maps** from TC-UNet for 30 timesteps wrt ***"hat"*** **token**

Overlay attention maps on generated image



DAAM Attention Maps

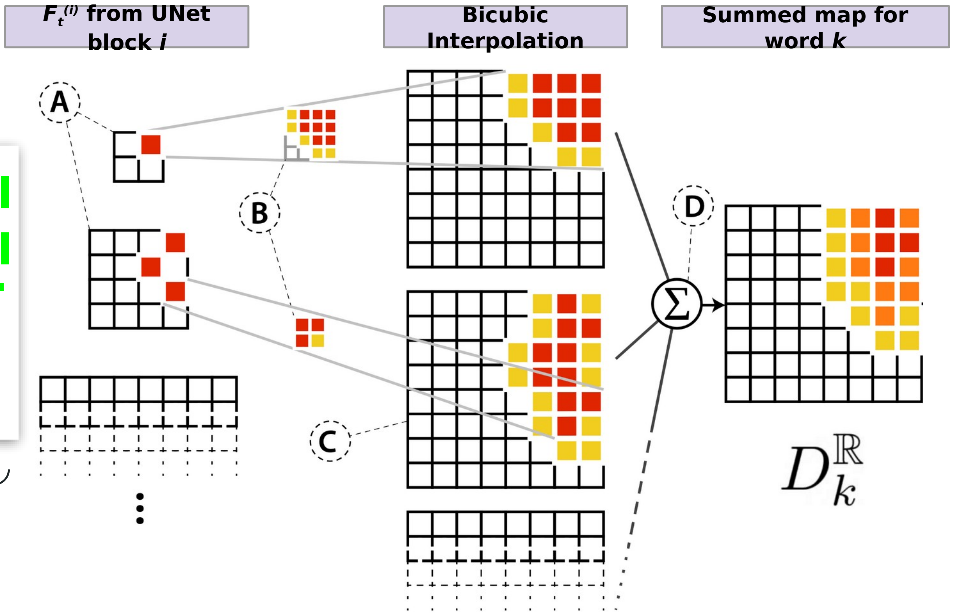
$$F_t^{(i)\downarrow} \in \mathbb{R}^{\underbrace{\lceil \frac{w}{c^i} \rceil \times \lceil \frac{h}{c^i} \rceil}_{\text{Width, height of downsampled UNet block}} \times l_H \times \boxed{l_W}}$$

Cross-Attention Map for block *i* at time *t*

Width, height of downsampled UNet block

Number of attention heads

Number of words in prompt



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

Summed map for each word *k*, $k \in \{1,.., l_W\}$

Downsampling blocks

Upsampling blocks

Sum across: UNet blocks *i*, timesteps *j*, attention heads *l*

Outline

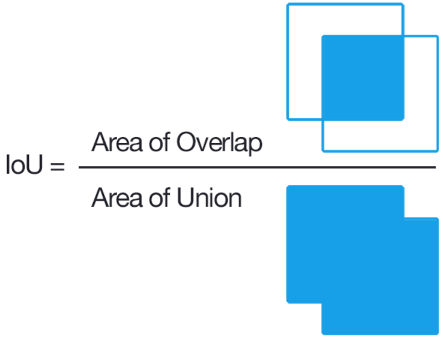
- Background: Stable Diffusion
- DAAM for text-to-image attribution
- Results
  - Attribution quality analysis
  - Syntax to pixels
  - Entanglement
- Related work
- Limitations & discussion

Attribution Analysis Part 1: Object Attribution

We can evaluate DAAM as an image-segmentation tool

DAAM segments Stable Diffusion images

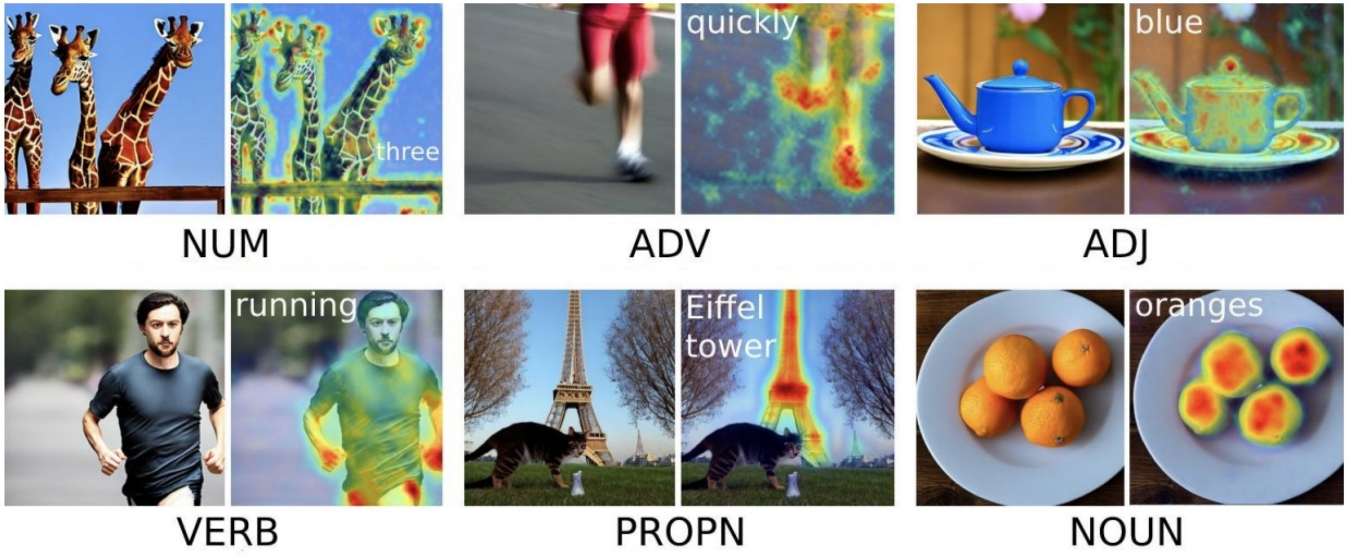
#	Method	COCO-Gen mIoU^{80}	COCO-Gen mIoU$^{\infty}$	Unreal-Gen mIoU^{80}	Unreal-Gen mIoU$^{\infty}$
	Supervised Methods				
1	Mask R-CNN (ResNet-101)	82.9	32.1	76.4	31.2
2	QueryInst (ResNet-101-FPN)	80.8	31.3	78.3	35.0
3	Mask2Former (Swin-S)	**84.0**	32.5	**80.0**	36.7
4	CLIPSeg	78.6	**71.6**	74.6	**70.9**
	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



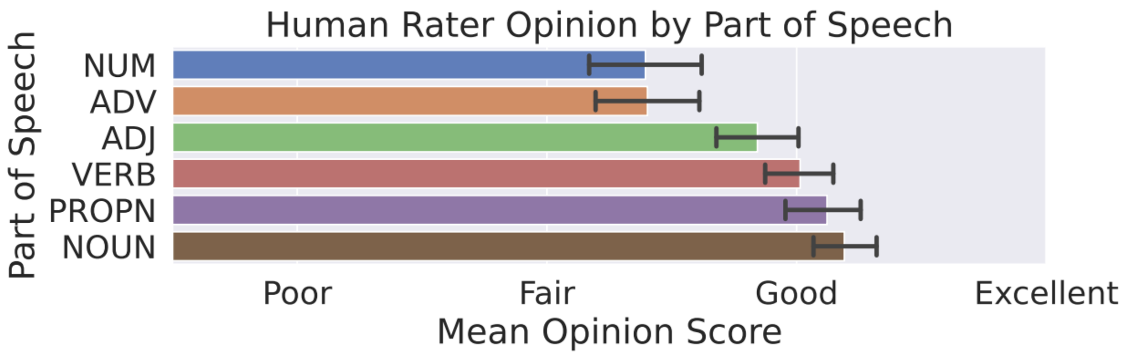
- DAAM does not require explicit segmentation labels
- DAAM is “open vocabulary” – can segment for any text input (not limited to known classes)

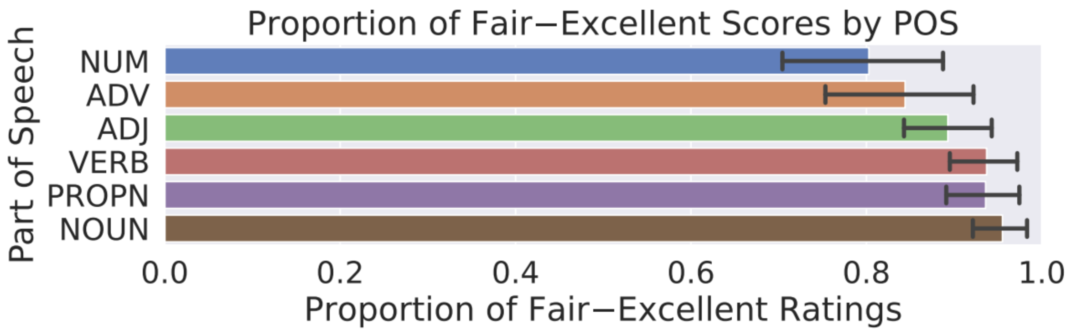
Attribution Analysis Part 2: Generalized Attribution

DAAM can segment beyond nouns

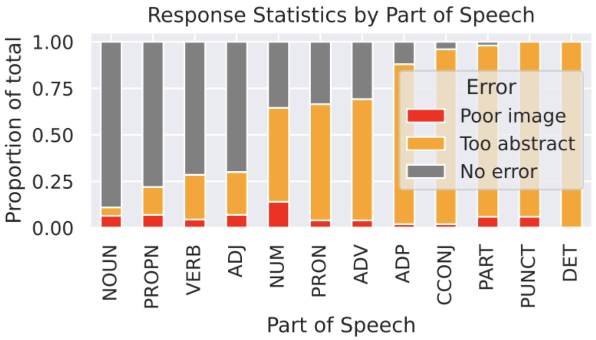


Evaluating DAAM with a user study





- 50 annotators, none see more than 18% of images
- Every image has three raters
- Abstract words / poor images were thrown out of evaluation



Visuosyntatic Analysis

How does **syntax** relate to the generated **pixels**?

Visuosyntatic Analysis

Head-Dependent Pairs

The cat sat on the mat

The red car

She ate the pizza

Visuosyntatic Analysis

Head-Dependent Pairs

The cat sat on the mat

Head because subject of sentence

Dependent because mat is the object of the sentence

The red car

Dependent because “red” is an adjective that describes the noun “car”

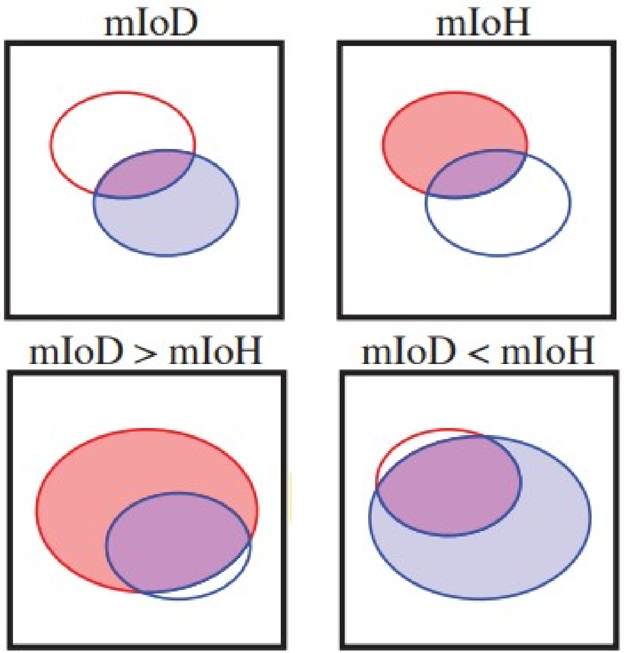
Head because main noun that represents the object being described

She ate the pizza

Head because main verb that describes the action being performed

Dependent because it is the noun that receives the action being performed

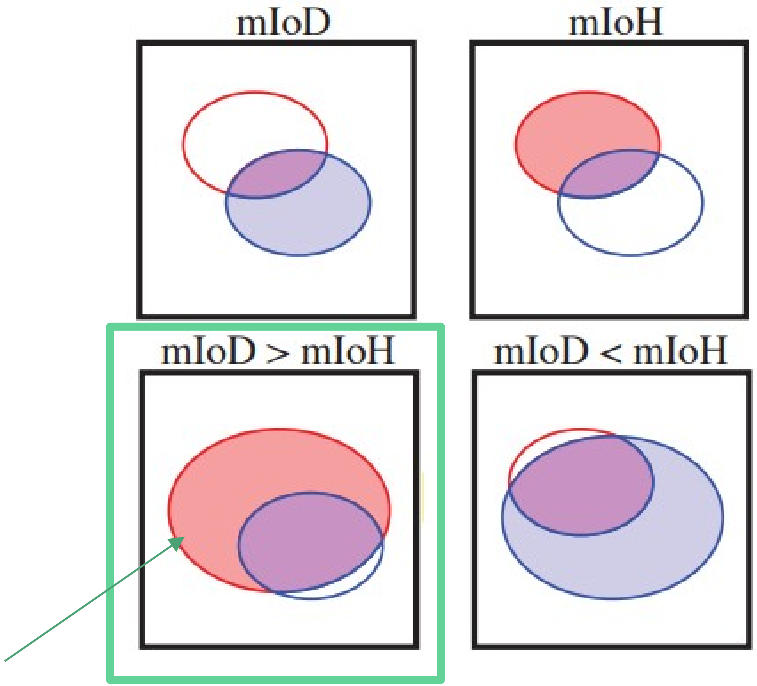
Visuosyntatic Analysis



Measures of Overlap

- Mean Intersection over Union (mIoU)
- Mean Intersection over the Dependent (mIoD)
- Mean Intersection over the Head (mIoH)

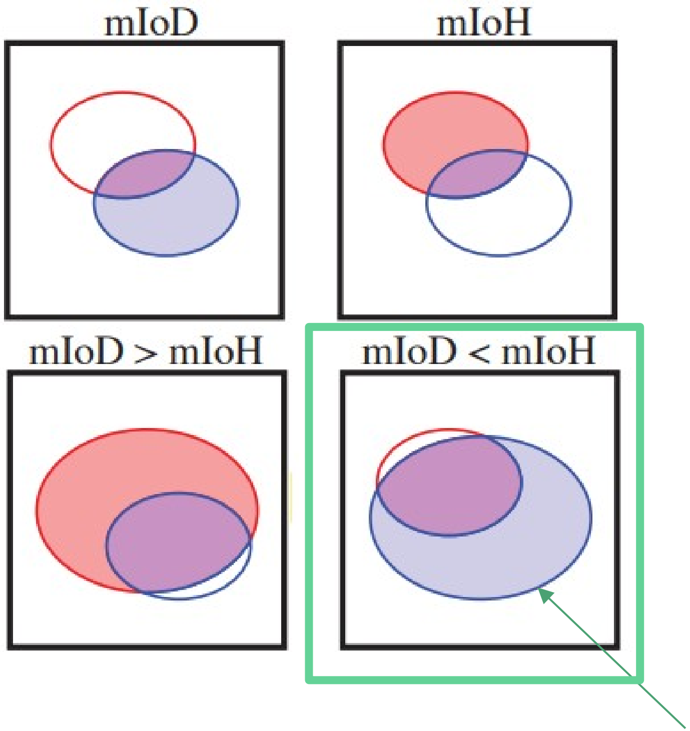
Visuosyntatic Analysis



Measures of Overlap

- Mean intersection over union (mIoU)
- Mean Intersection over the Dependent (mIoD)
- Mean Intersection over the Head (mIoH)

Visuosyntatic Analysis



Measures of Overlap

- Mean intersection over union (mIoU)
- Mean Intersection over the Dependent (mIoD)
- Mean Intersection over the Head (mIoH)

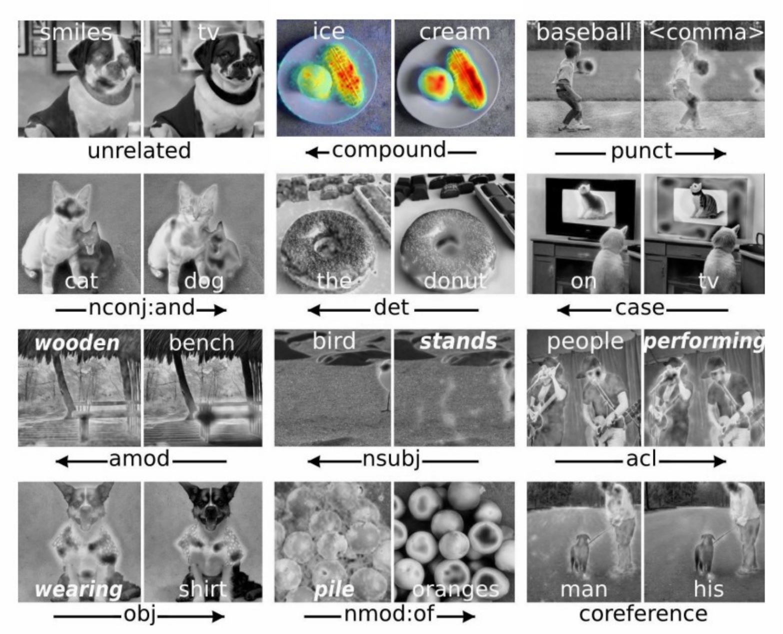
Visuosyntatic Analysis

#	Relation	mIoD	mIoH	Δ	mIoU
1	Unrelated pairs	65.1	66.1	1.0	47.5
2	All head–dependent pairs	62.3	62.0	0.3	43.4
3	compound	71.3	71.5	0.2	51.1
4	punct	68.2	70.0	1.8	49.5
5	nconj:and	58.0	56.1	1.9	38.2
6	det	54.8	52.2	2.6	35.0
7	case	51.7	58.1	6.4	36.9
8	acl	**67.4**	79.3	12.	55.4
9	nsubj	76.4	**63.9**	12.	52.2
10	amod	**62.4**	77.6	15.	51.1
11	nmod:of	73.5	**57.9**	16.	47.5
12	obj	75.6	**46.3**	29.	55.4
14	Coreferent word pairs	84.8	77.4	7.4	**66.6**

Noun Compounds

ice + cream

Visuosyntatic Analysis



Noun Compounds

ice + cream

Takeaway:

No dominance

Complement one another

Visuosyntatic Analysis

#	Relation	mIoD	mIoH	Δ	mIoU
1	Unrelated pairs	65.1	66.1	1.0	47.5
2	All head–dependent pairs	62.3	62.0	0.3	43.4
3	compound	71.3	71.5	0.2	51.1
4	punct	68.2	70.0	1.8	49.5
5	nconj:and	58.0	56.1	1.9	38.2
6	det	54.8	52.2	2.6	35.0
7	case	51.7	58.1	6.4	36.9
8	acl	**67.4**	79.3	12.	55.4
9	nsubj	76.4	**63.9**	12.	52.2
10	amod	**62.4**	77.6	15.	51.1
11	nmod:of	73.5	**57.9**	16.	47.5
12	obj	75.6	**46.3**	29.	55.4
14	Coreferent word pairs	84.8	77.4	7.4	**66.6**

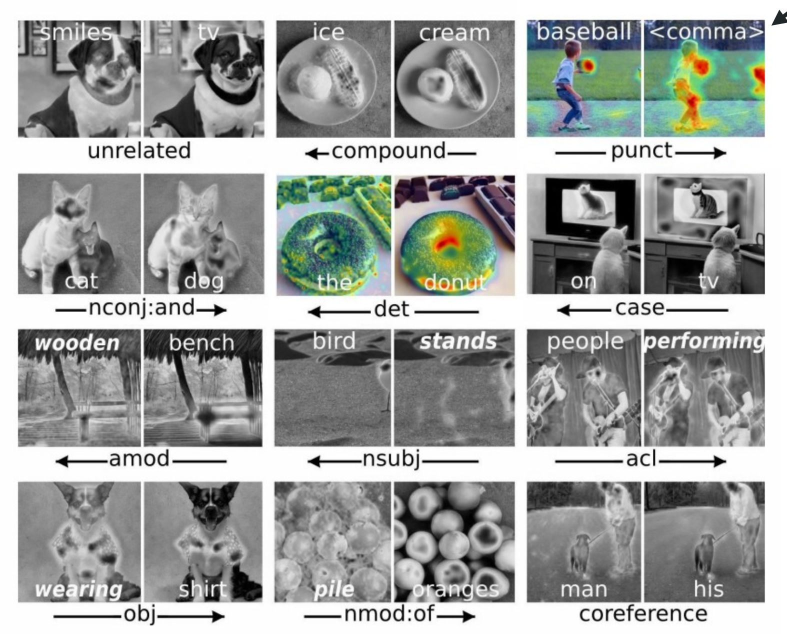
Punctuation & Articles

the + donut

baseball + <comma>

Visuosyntatic Analysis

(normalized)



Punctuation & Articles

the + donut

baseball + <comma>

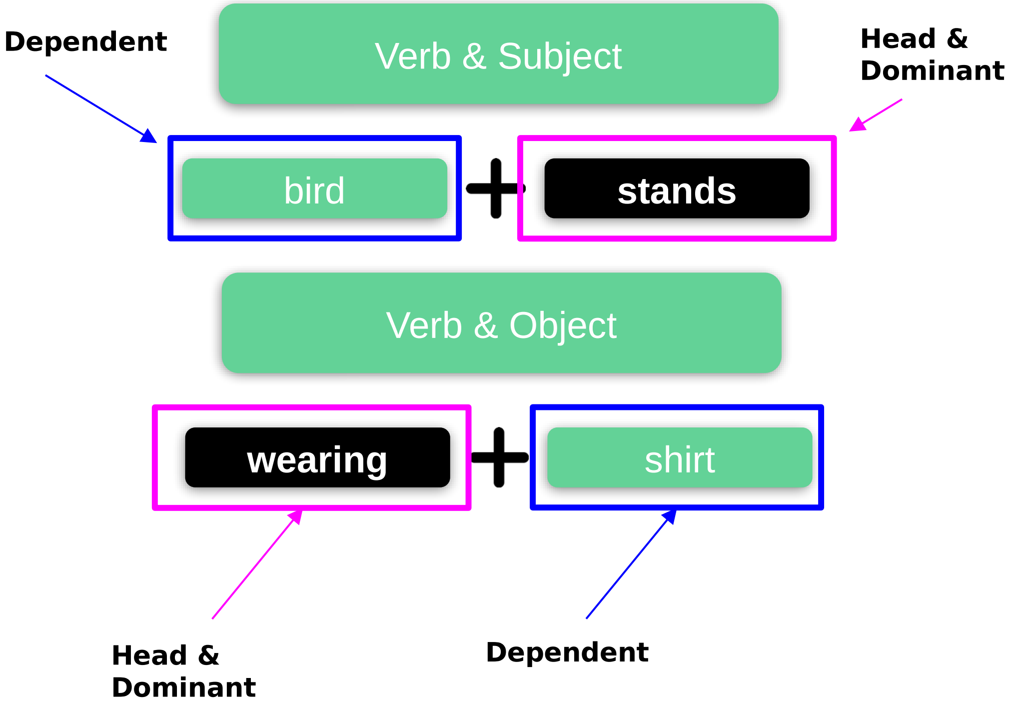
Takeaway:

No dominance

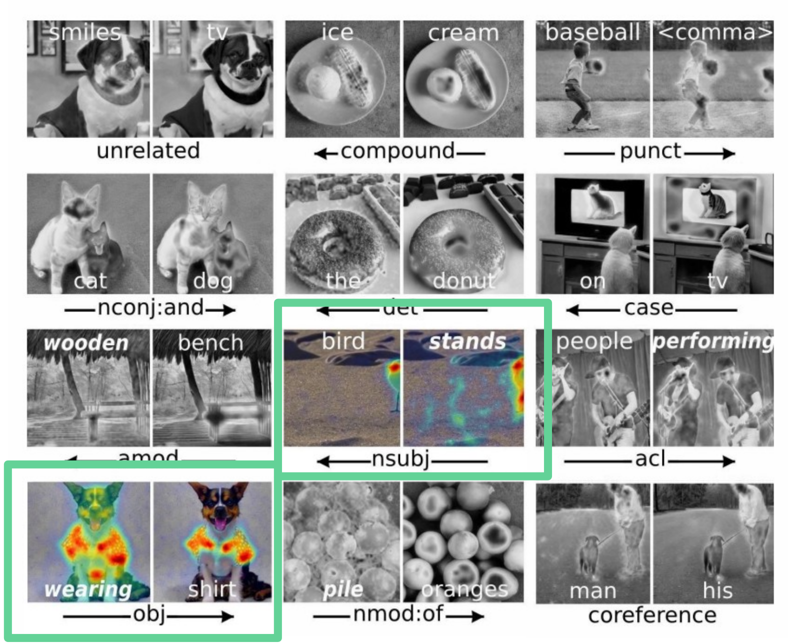
Little semantic meaning

Visuosyntatic Analysis

#	Relation	mIoD	mIoH	Δ	mIoU
1	Unrelated pairs	65.1	66.1	1.0	47.5
2	All head–dependent pairs	62.3	62.0	0.3	43.4
3	compound	71.3	71.5	0.2	51.1
4	punct	68.2	70.0	1.8	49.5
5	nconj:and	58.0	56.1	1.9	38.2
6	det	54.8	52.2	2.6	35.0
7	case	51.7	58.1	6.4	36.9
8	acl	**67.4**	79.3	12.	55.4
9	nsubj	76.4	**63.9**	12.	52.2
10	amod	**62.4**	77.6	15.	51.1
11	nmod:of	73.5	**57.9**	16.	47.5
12	obj	75.6	**46.3**	29.	55.4
14	Coreferent word pairs	84.8	77.4	7.4	**66.6**



Visuosyntatic Analysis



Dependent

Verb & Subject

Head & Dominant

bird + **stands**

Verb & Object

wearing + shirt

Head & Dominant

Dependent

Takeaway:

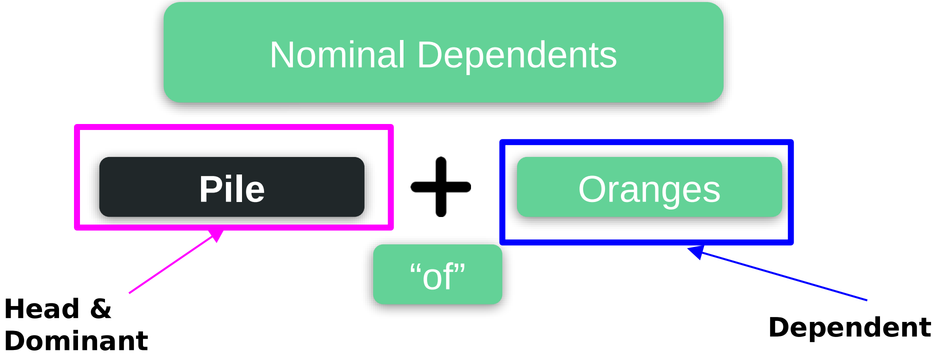
Semi-intuitive

Head verb dominates subject/object

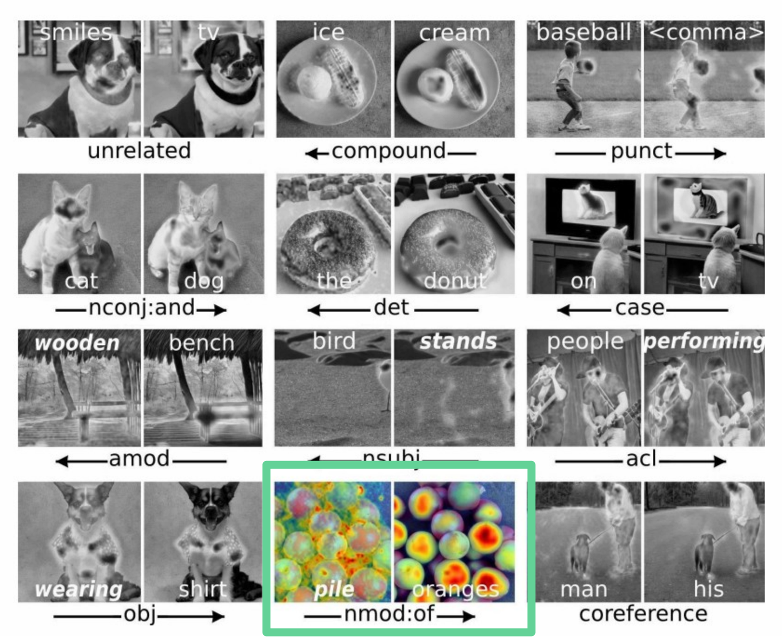
Verb contextualises the subject/object in its surroundings

Visuosyntatic Analysis

#	Relation	mIoD	mIoH	Δ	mIoU
1	Unrelated pairs	65.1	66.1	1.0	47.5
2	All head–dependent pairs	62.3	62.0	0.3	43.4
3	`compound`	71.3	71.5	0.2	51.1
4	`punct`	68.2	70.0	1.8	49.5
5	`nconj:and`	58.0	56.1	1.9	38.2
6	`det`	54.8	52.2	2.6	35.0
7	`case`	51.7	58.1	6.4	36.9
8	`acl`	**67.4**	79.3	12.	55.4
9	`nsubj`	76.4	**63.9**	12.	52.2
10	`amod`	**62.4**	77.6	15.	51.1
11	`nmod:of`	73.5	**57.9**	16.	47.5
12	`obj`	75.6	**46.3**	29.	55.4
14	Coreferent word pairs	84.8	77.4	7.4	**66.6**



Visuosyntatic Analysis



Nominal Dependents

Pile + Oranges

“of”

Head & Dominant

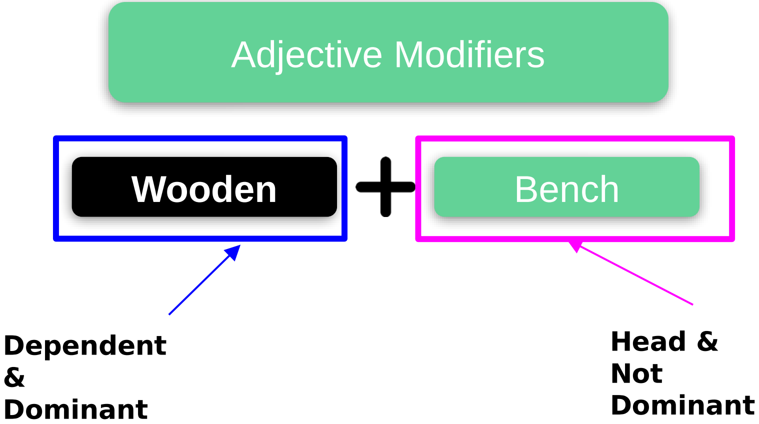
Dependent

Takeaway:

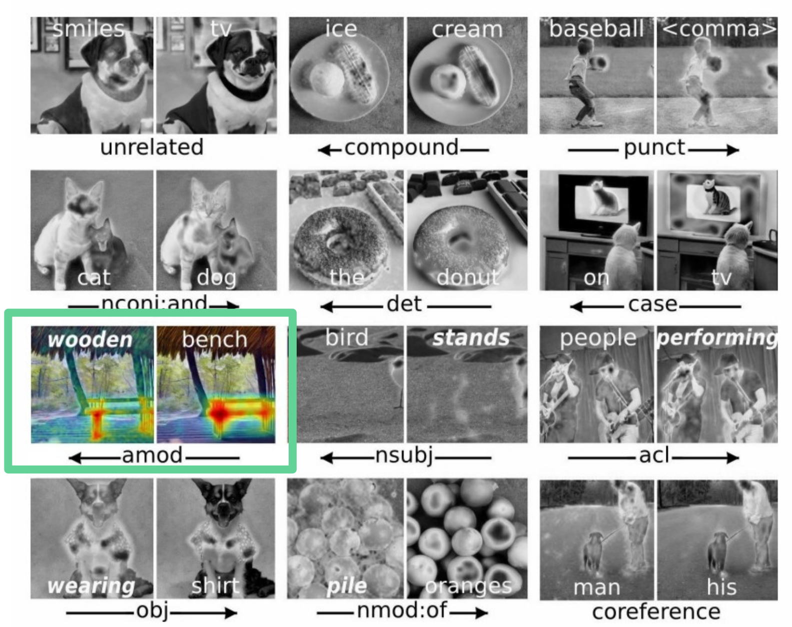
Dominance is intuitive

Visuosyntatic Analysis

#	Relation	mIoD	mIoH	Δ	mIoU
1	Unrelated pairs	65.1	66.1	1.0	47.5
2	All head–dependent pairs	62.3	62.0	0.3	43.4
3	compound	71.3	71.5	0.2	51.1
4	punct	68.2	70.0	1.8	49.5
5	nconj:and	58.0	56.1	1.9	38.2
6	det	54.8	52.2	2.6	35.0
7	case	51.7	58.1	6.4	36.9
8	acl	**67.4**	79.3	12.	55.4
9	nsubj	76.4	**63.9**	12.	52.2
10	amod	**62.4**	77.6	15.	51.1
11	nmod:of	73.5	**57.9**	16.	47.5
12	obj	75.6	**46.3**	29.	55.4
14	Coreferent word pairs	84.8	77.4	7.4	**66.6**



Visuosyntatic Analysis



Adjective Modifiers

Wooden + Bench

Dependent & Dominant

Head & Not Dominant

Takeaway:

Dominance is counter intuitive

Visuosyntatic Analysis

How does **syntax** relate to the generated **pixels**?

Takeaway:

Attribution map of the **dependent subsumes** that of the **head**, and **opposite** for others

Dominance is **intuitive** in some cases but **counter intuitive** in others

Visuosemantic Analysis

Do semantically **similar words** have **worse** generation quality?

Visuosemantic Analysis: **Cohyponym Entanglement**

Prompt Structure:

“a(n) <noun> and a(n) <noun>”

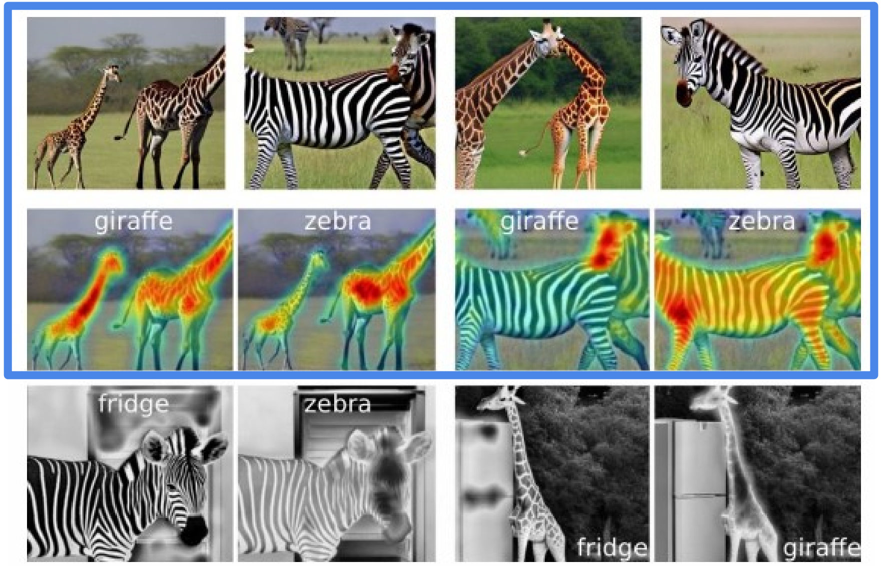
Cohyponym Example:

“a giraffe and a zebra”

Non-Cohyponym Example:

“a zebra and a fridge”

Visuosemantic Analysis: **Cohyponym Entanglement**



Cohyponym

"a giraffe and a zebra"

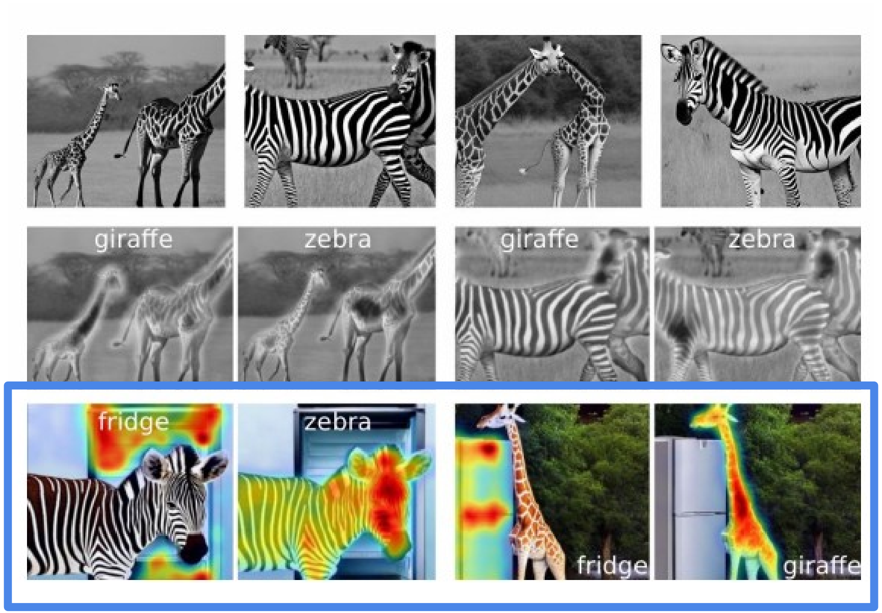
Takeaway

Stable-Diffusion image generation **worsens**

Generates **one** of the nouns but **not both**

Attribution **maps** for the two nouns **overlap … Feature Entanglement**

Visuosemantic Analysis: **Cohyponym Entanglement**



Non-Cohyponym

“a zebra and a fridge”

Takeaway

Generates **both** of the nouns

Attribution **maps** for the two nouns are **distinct**

Visuosemantic Analysis: **Adjectival Entanglement**

Prompt Structure:

“<adj> <noun> <verb phrase>”

Example:

“a [rusty] shovel sitting in a clean shed”

“a [bumpy] ball rolling down a hill”

Visuosemantic Analysis: **Adjectival Entanglement**

Expected Behavior

If **no entanglement**, background should **not gain attributes** pertaining to that adjective

Visuosemantic Analysis: **Adjectival Entanglement**

Takeaway

Attribution maps for adjectives attend **too broadly** across images **beyond nouns** they modify … **Feature Entanglement**

Summary

- DAAM provides pixel-level attribution maps for Stable Diffusion, a state-of-the-art text-to-image generator
- These maps appear to be informative, as evaluated through segmentation tasks and user study
- DAAM can be a useful tool for further understanding and analyzing Stable Diffusion – e.g. through visuosyntactic analysis and visuosemantic analysis

Class Discussion

- Does DAAM give a clear understanding about how a large-scale latent diffusion model synthesizes text to image and which parts of an image is influenced the most?
- Does DAAM explain all the dynamics of how images are synthesized? If not, how should DAAM be modified to better explain image generation?
    - Other explanatory tools besides **attention** and **segmentation proposals**?
- DAAM pointed out failure cases of stable-diffusion. Are there further interpretability methods needed to understand why **feature entanglement** is occurring and how it could be improved?