

What the DAAM: Interpreting Stable Diffusion Using Cross Attention

Raphael Tang,^{*1} Linqing Liu,^{*2} Akshat Pandey,¹ Zhiying Jiang,³ Gefei Yang,¹
Karun Kumar,¹ Pontus Stenetorp,² Jimmy Lin,³ Ferhan Ture¹

¹Comcast Applied AI ²University College London ³University of Waterloo

¹{raphael_tang, akshat_pandey, gefei_yang, karun_kumar, ferhan_ture}@comcast.com

²{linqing.liu, p.stenetorp}@cs.ucl.ac.uk ³{zhiying.jiang, jimmylin}@uwaterloo.ca

Abstract

Large-scale diffusion neural networks represent a substantial 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 pixel-level attribution maps, we upscale and aggregate cross-attention word-pixel scores in the denoising subnetwork, naming our method DAAM. We evaluate its correctness by testing its semantic segmentation ability on nouns, as well as its generalized attribution quality on all parts of speech, rated by humans. We then apply DAAM to study the role of syntax in the pixel space, characterizing head-dependent heat map interaction patterns for ten common dependency relations. Finally, we study several semantic phenomena using DAAM, with a focus on feature entanglement, where we find that cohyponyms worsen generation quality and descriptive adjectives attend too broadly. To our knowledge, we are the first to interpret large diffusion models from a visiolinguistic perspective, which enables future lines of research. Our code is at <https://github.com/castorini/daam>.

1 Introduction

Diffusion neural networks trained on billions of image-caption pairs represent the state of the art in text-to-image generation (Yang et al., 2022), with some achieving realism comparable to photographs in human evaluation, such as Google’s Imagen (Saharia et al., 2022) and OpenAI’s DALL-E 2 (Ramesh et al., 2022). However, despite their quality and popularity, the dynamics of their image synthesis remain undercharacterized. Citing ethical concerns, these organizations have restricted the general public from using the models and their weights, preventing effective white-box (or even blackbox) analysis. To overcome this barrier,

^{*}Equal contribution.

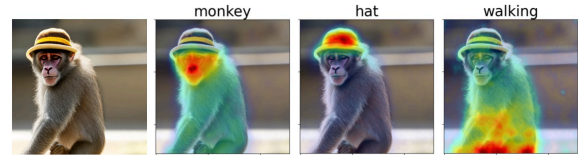


Figure 1: The original synthesized image and three DAAM maps for “monkey,” “hat,” and “walking,” from the prompt, “monkey with hat walking.”

Stability AI recently open-sourced Stable Diffusion (Rombach et al., 2022), a 1.1 billion-parameter latent diffusion model pretrained and fine-tuned on the LAION 5-billion image dataset (Schuhmann et al., 2022).

We probe Stable Diffusion to provide insight into the workings of large diffusion models. With a focus on text-to-image attribution, our central research question is, “How does an input word influence parts of a generated image?” To this, we first propose to produce two-dimensional attribution maps for each word by combining cross-attention maps in the model, as delineated in Section 2.2. A related work in prompt-guided editing from Hertz et al. (2022) conjectures that per-head cross attention relates words to areas in Imagen-generated images, but they fall short of constructing global per-word attribution maps. We name our method diffusion attentive attribution maps, or DAAM for short—see Figure 1 for an example.

To evaluate the veracity of DAAM, we apply it to a semantic segmentation task (Lin et al., 2014) on generated imagery, comparing DAAM maps with annotated segments. We attain a 58.9–64.8 mean intersection over union (mIoU) score, which is competitive with unsupervised segmentation models, described in Section 3.1. We further bolster these noun attribution results using a generalized study covering all parts of speech, such as adjectives and verbs. Through human annotation, we show that the mean opinion score (MOS) is above fair to good (3.4–4.2) on interpretable words.

Next, we characterize how relationships in the syntactic space of prompts relate to those in the pixel space of images. We assess head-dependent DAAM map interactions across ten common syntactic relationships, finding that, for some, the heat map of the dependent strongly subsumes that of the head, while the opposite is true for others. For still others, such as coreferent word pairs, the words’ maps greatly overlap, indicating identity. We assign visual intuition to our observations; for example, we conjecture that the maps of verbs contain those of their subjects, because verbs often contextualize both the subjects and their surroundings.

Finally, we form hypotheses to further examine our syntactic findings, studying semantic phenomena through the lens of DAAM, particularly those affecting the generation quality. In Section 5.1, we demonstrate that, in constructed prompts with two distinct nouns, cohyponyms have worse quality, e.g., “a giraffe and a zebra” generates either a giraffe *or* a zebra, but not both. We observe that cohyponym status and generation incorrectness each increases the amount of overlap between the heat maps. We also show in Section 5.2 that descriptive adjectives attend too broadly across the image, far beyond the nouns they modify. If we hold the scene layout fixed (Hertz et al., 2022) and vary only the adjective, the entire image changes, not just the noun. These two phenomena suggest feature entanglement, where objects are entangled with both the scene and other objects.

In summary, our contributions are as follows: (1) we propose and evaluate an attribution method, novel within the context of interpreting diffusion models, measuring which parts of the generated image the words influence most; (2) we provide new insight into how syntactic relationships map to generated pixels, finding evidence for directional imbalance in head-dependent DAAM map overlap, alongside visual intuition (and counterintuition) in the behaviors of nominals, modifiers, and function words; and (3) we shine light on failure cases in diffusion models, showing that descriptive adjectival modifiers and cohyponyms result in entangled features and DAAM maps.

2 Our Approach

2.1 Preliminaries

Latent diffusion models (Rombach et al., 2022) are a class of denoising generative models that are trained to synthesize high-fidelity images from ran-

dom noise through a gradual denoising process, optionally conditioned on text. They generally comprise three components: a deep language model like CLIP (Radford et al., 2021) for producing word embeddings; a variational autoencoder (VAE; Kingma and Welling, 2013) which encodes and decodes latent vectors for images; and a time-conditional U-Net (Ronneberger et al., 2015) for gradually denoising latent vectors. To generate an image, we initialize the latent vectors to random noise, feed in a text prompt, then iteratively denoise the latent vectors with the U-Net and decode the final vector into an image with the VAE.

Formally, given an image, the VAE encodes it as a latent vector $\ell_{t_0} \in \mathbb{R}^d$. Define a forward “noise injecting” Markov chain $p(\ell_{t_i} | \ell_{t_{i-1}}) := \mathcal{N}(\ell_{t_i}; \sqrt{1 - \alpha_{t_i}} \ell_{t_0}, \alpha_{t_i} \mathbf{I})$ where $\{\alpha_{t_i}\}_{i=1}^T$ is defined following a schedule so that $p(\ell_{t_T})$ is approximately zero-mean isotropic. The corresponding denoising reverse chain is then parameterized as

$$p(\ell_{t_{i-1}} | \ell_{t_i}) := \mathcal{N}(\ell_{t_{i-1}}; \frac{1}{\sqrt{1 - \alpha_{t_i}}}(\ell_{t_i} + \alpha_{t_i} \epsilon_{\theta}(\ell_{t_i}, t_i)), \alpha_{t_i} \mathbf{I}), \quad (1)$$

for some denoising neural network $\epsilon_{\theta}(\ell, t)$ with parameters θ . Intuitively, the forward process iteratively adds noise to some signal at a fixed rate, while the reverse process, equipped with a neural network, removes noise until recovering the signal. To train the network, given caption-image pairs, we optimize

$$\min_{\theta} \sum_{i=1}^T \zeta_i \mathbb{E}_{p(\ell_{t_i} | \ell_{t_0})} \|\epsilon_{\theta}(\ell_{t_i}, t_i) - \nabla_{\ell_{t_i}} \log p(\ell_{t_i} | \ell_{t_0})\|_2^2, \quad (2)$$

where $\{\zeta_i\}_{i=1}^T$ are constants computed as $\zeta_i := 1 - \prod_{j=1}^i (1 - \alpha_j)$. The objective is a reweighted form of the evidence lower bound for score matching (Song et al., 2021). To generate a latent vector, we initialize $\hat{\ell}_{t_T}$ as Gaussian noise and iterate

$$\hat{\ell}_{t_{i-1}} = \frac{1}{\sqrt{1 - \alpha_{t_i}}}(\hat{\ell}_{t_i} + \alpha_{t_i} \epsilon_{\theta}(\hat{\ell}_{t_i}, t_i)) + \sqrt{\alpha_{t_i}} z_{t_i}. \quad (3)$$

In practice, we apply various optimizations to improve the convergence of the above step, like modeling the reverse process as an ODE (Song et al., 2021), but this definition suffices for us. We can additionally condition the latent vectors on text and pass word embeddings $\mathbf{X} := [\mathbf{x}_1; \dots; \mathbf{x}_{l_W}]$ to $\epsilon_{\theta}(\ell, t; \mathbf{X})$. Finally, the VAE decodes the denoised latent $\hat{\ell}_{t_0}$ to an image. For this paper, we use the publicly available weights of the state-of-the-art, 1.1 billion-parameter Stable Diffusion 2.0 model (Rombach et al., 2022), trained on 5 billion caption-image pairs (Schuhmann et al., 2022) and implemented in HuggingFace’s Diffusers library (von Platen et al., 2022).

2.2 Diffusion Attentive Attribution Maps

Given a large-scale latent diffusion model for text-to-image synthesis, which parts of an image does each word influence most? One way to achieve this would be attribution approaches, which are mainly perturbation- and gradient-based (Alvarez-Melis and Jaakkola, 2018; Selvaraju et al., 2017), where saliency maps are constructed either from the first derivative of the output with respect to the input, or from input perturbation to see how the output changes. Unfortunately, gradient methods prove intractable due to needing a backpropagation pass for every pixel for all T time steps, and even minor perturbations result in significantly different images in our pilot experiments.

Instead, we use ideas from natural language processing, where word attention was found to indicate lexical attribution (Clark et al., 2019), as well as the spatial layout of Imagen’s images (Hertz et al., 2022). In diffusion models, attention mechanisms cross-contextualize text embeddings with coordinate-aware latent representations (Rombach et al., 2022) of the image, outputting scores for each token–image patch pair. Attention scores lend themselves readily to interpretation since they are already normalized in $[0, 1]$. Thus, for pixelwise attribution, we propose to aggregate these scores over the spatiotemporal dimensions and interpolate them across the image.

We turn our attention to the denoising network $\epsilon_\theta(\ell, t; \mathbf{X})$ responsible for the synthesis. While the subnetwork can take any form, U-Nets remain the popular choice (Ronneberger et al., 2015) due to their strong image segmentation ability. They consist of a series of downsampling convolutional blocks, each of which preserves some local context, followed by upsampling deconvolutional blocks, which restore the original input size to the output. Specifically, given a 2D latent $\ell_t \in \mathbb{R}^{w \times h}$, the downsampling blocks output a series of vectors $\{\mathbf{h}_{i,t}^\downarrow\}_{i=1}^K$, where $\mathbf{h}_{i,t}^\downarrow \in \mathbb{R}^{\lceil \frac{w}{c^i} \rceil \times \lceil \frac{h}{c^i} \rceil}$ for some $c > 1$. The upsampling blocks then iteratively upscale $\mathbf{h}_{K,t}^\downarrow$ to $\{\mathbf{h}_{i,t}^\uparrow\}_{i=K-1}^0 \in \mathbb{R}^{\lceil \frac{w}{c^i} \rceil \times \lceil \frac{h}{c^i} \rceil}$. To condition these representations on word embeddings, Rombach et al. (2022) use multi-headed cross-attention layers (Vaswani et al., 2017)

$$\mathbf{h}_{i,t}^\downarrow := F_t^{(i)}(\hat{\mathbf{h}}_{i,t}^\downarrow, \mathbf{X}) \cdot (W_v^{(i)} \mathbf{X}), \quad (4)$$

$$F_t^{(i)}(\hat{\mathbf{h}}_{i,t}^\downarrow, \mathbf{X}) := \text{softmax} \left((W_q^{(i)} \hat{\mathbf{h}}_{i,t}^\downarrow)(W_k^{(i)} \mathbf{X})^T / \sqrt{d} \right),$$

where $F_t^{(i)\downarrow} \in \mathbb{R}^{\lceil \frac{w}{c^i} \rceil \times \lceil \frac{h}{c^i} \rceil \times l_H \times l_W}$ and W_k, W_q , and W_v are projection matrices with l_H attention

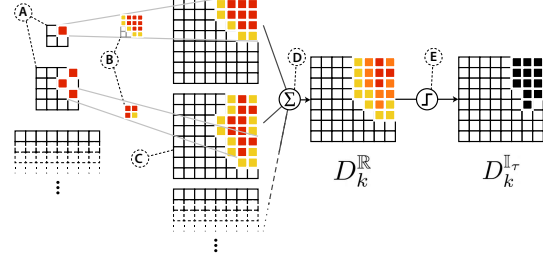


Figure 2: Illustration of computing DAAM for some word: the multiscale attention arrays from Eqn. (5) (see A); the bicubic interpolation (B) resulting in expanded maps (C); summing the heat maps across the layers (D), as in Eqn. (6); and the thresholding (E) from Eqn. (7).

heads. The same mechanism applies when upsampling \mathbf{h}_i^\uparrow . For brevity, we denote the respective attention score arrays as $F_t^{(i)\downarrow}$ and $F_t^{(i)\uparrow}$, and we implicitly broadcast matrix multiplications as per NumPy convention (Harris et al., 2020).

Spatiotemporal aggregation. $F_t^{(i)\downarrow}[x, y, \ell, k]$ is normalized to $[0, 1]$ and connects the k^{th} word to the intermediate coordinate (x, y) for the i^{th} downsampling block and ℓ^{th} head. Due to the fully convolutional nature of U-Net (and the VAE), the intermediate coordinates locally map to a surrounding affected square area in the final image, the scores thus relating each word to that image patch. However, different layers produce heat maps with varying scales, deepest ones being the coarsest (e.g., $\mathbf{h}_{K,t}^\downarrow$ and $\mathbf{h}_{K-1,t}^\uparrow$), requiring spatial normalization to create a single heat map. To do this, we upscale all intermediate attention score arrays to the original image size using bicubic interpolation, then sum them over the heads, layers, and time steps:

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

where k is the k^{th} word and $\tilde{F}_{t,j,k,\ell}^{(i)\downarrow}[x, y]$ is shorthand for $F_t^{(i)\downarrow}[x, y, \ell, k]$, bicubically upscaled to fixed size (w, h) .¹ Since D_k^R is positive and scale normalized (summing normalized values preserves linear scale), we can visualize it as a soft heat map, with higher values having greater attribution. To generate a hard, binary heat map (either a pixel is influenced or not), we can threshold D_k^R as

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

where $\mathbb{I}(\cdot)$ is the indicator function and $\tau \in [0, 1]$. See Figure 2 for an illustration of DAAM.

¹We show that aggregating across all time steps and layers is indeed necessary in Section A.1.

# Method	COCO-Gen		Unreal-Gen	
	mIoU ⁸⁰	mIoU [∞]	mIoU ⁸⁰	mIoU [∞]
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

Table 1: mIoU of semantic segmentation methods on our synthesized datasets. Best in each section bolded.

3 Attribution Analyses

3.1 Object Attribution

Quantitative evaluation of our method is challenging, but we can attempt to draw upon existing annotated datasets and methods to see how well our method aligns. A popular visuosemantic task is image segmentation, where areas (i.e., segmentation masks) are given a semantically meaningful label, commonly nouns. If DAAM is accurate, then our attention maps should arguably align with the image segmentation labels for these tasks—despite not having been trained to perform this task.

Setup. We ran the Stable Diffusion 2.0 base model using 30 inference steps per image with the DPM (Lu et al., 2022) solver—see the appendix section A.1 for specifics. We then synthesized **one set of images using the validation set of the COCO image captions dataset** (Lin et al., 2014), **representing realistic prompts**, and **another set by randomly swapping nouns in the same set (holding the vocabulary fixed), representing unrealism**. The purpose of the second set was to see how well the model generalized to uncanny prompts, whose composition was unlikely to have been encountered at training time. We named the two sets “**COCO-Gen**” and “**Unreal-Gen**,” each with 100 prompt–image pairs. For ground truth, we extracted all countable nouns from the prompts, then hand-segmented each present noun in the image.

To compute binary DAAM segmentation masks, we used Eqn. 7 with thresholds $\tau \in \{0.3, 0.4, 0.5\}$, for each noun in the ground truth. We refer to these methods as **DAAM- $\langle\tau\rangle$** , e.g., **DAAM-0.3**. For supervised baselines, we evaluated semantic segmen-

tation models trained explicitly on COCO, like Mask R-CNN (He et al., 2017) with a ResNet-101 backbone (He et al., 2016), QueryInst (Fang et al., 2021) with ResNet-101-FPN (Lin et al., 2017), and Mask2Former (Cheng et al., 2022) with Swin-S (Liu et al., 2021), all implemented in MMDetection (Chen et al., 2019), as well as the open-vocabulary CLIPSeg (Lüddecke and Ecker, 2022) trained on the PhraseCut dataset (Wu et al., 2020). We note that CLIPSeg’s setup resembles ours because the image captions are assumed given as well. However, their model is supervised since they additionally train their model on segmentation labels. Our unsupervised baselines consisted of the state-of-the-art STEGO (Hamilton et al., 2021) and PiCIE + H (Cho et al., 2021). As is standard (Lin et al., 2014), we evaluated all approaches using the mean intersection over union (mIoU) over the prediction–ground truth mask pairs. We denote mIoU⁸⁰ when restricted to the 80 COCO classes that the supervised baselines were trained on (save for CLIPSeg) and mIoU[∞] as the mIoU without the class restriction.

Results. We present results in Table 1. The COCO-supervised models (rows 1–3) are constrained to COCO’s 80 classes (e.g., “cat,” “cake”), while DAAM (rows 5–7) is open vocabulary; thus, DAAM outperforms them by 22–28 points in mIoU[∞] and underperforms by 20 points in mIoU⁸⁰. CLIPSeg (row 4), an open-vocabulary model trained on semantic segmentation datasets, achieves the best of both worlds in mIoU⁸⁰ and mIoU[∞], with the highest mIoU[∞] overall and high mIoU⁸⁰. However, its restriction to nouns precludes it from generalized segmentation (e.g., verbs). DAAM largely outperforms both unsupervised baselines (rows 6–7) by a margin of 4.4–29 points (see rows 7–10), likely because we assume the prompts to be provided. Similar findings hold on the unrealistic Unreal-Gen set, showing that DAAM is resilient to nonsensical texts, confirming that DAAM works when Stable Diffusion has to generalize in composition.

As for τ , 0.4 works best on all splits, though it isn’t too sensitive, varying by 3–6 points in mIoU. We also show that all layers and time steps contribute to DAAM’s segmentation quality, shown in Section A.1. Overall, DAAM forms a strong baseline of 57.9–64.8 mIoU⁸⁰. We conclude that it is empirically sane, which we further support for all parts of speech in the next section.

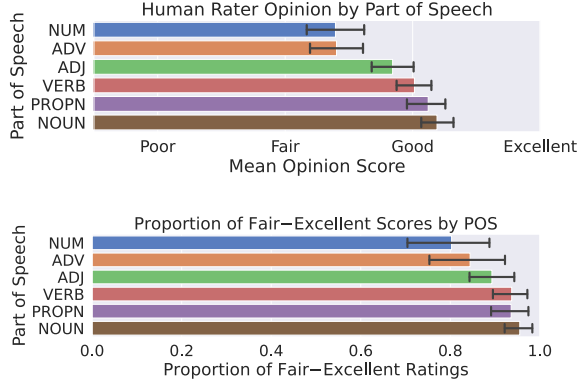


Figure 3: On the top, mean opinion scores grouped by part of speech, with 95% confidence interval bars; on the bottom, proportion of fair–excellent scores, grouped by part-of-speech.

3.2 Generalized Attribution

We extend our veracity analyses beyond nouns to all parts of speech, such as adjectives and verbs, to show that DAAM is more generally applicable. A high-quality, reliable analysis requires human annotation; hence, we ask human raters to evaluate the attribution quality of DAAM maps, using a five-point Likert scale.

This setup generalizes that of the last section because words in general are not visually separable, which prevents effective segmentation annotation. For example, in the prompt “people running,” it is unclear where to visually segment “running.” Is it just the knees and feet of the runners, or is it also the swinging arms? On the contrary, if annotators are instead given the proposed heat maps for “running,” they can make a judgement on how well the maps reflect the word.

Setup. To construct our word–image dataset, we first randomly sampled 200 words from each of the 14 most common part-of-speech tags in COCO, extracted with spaCy, for a total of 2,800 unique word–prompt pairs. Next, we generated images alongside DAAM maps for all pairs, varying the random seed each time. To gather human judgements, we built our annotation interface in Amazon MTurk, a crowdsourcing platform. We presented the generated image, the heat map, and the prompt with the target word in red, beside a question asking expert workers to rate how well the highlighting reflects the word. They then selected a rating among one of “bad,” “poor,” “fair,” “good,” and “excellent”, as well as an option to declare the image itself as too poor or the word too abstract to inter-

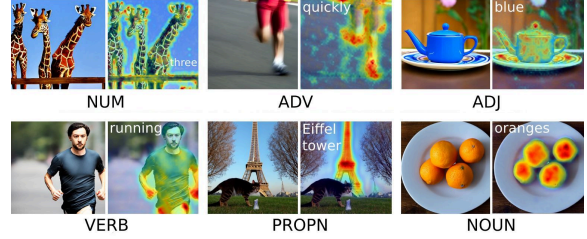


Figure 4: Example generations and DAAM heat maps from COCO for each interpretable part-of-speech.

pret. For quality control, we removed annotators failing attention tests. For further robustness, we assigned three unique raters to each example. We provide further details on the user interface and annotation process in the appendix section A.2.

Results. Our examples were judged by a total of fifty raters, none producing more than 18% of the total number of annotations. We filtered out all word–image pairs deemed too abstract (e.g., “the”), when any one of the three assigned raters selected that option. This resulted in six interpretable part-of-speech tags with enough judgements—see the appendix for detailed statistics. To compute the final score of each word–image pair, we took the median of the three raters’ opinions.

We plot our results in Figure 3. In the top sub-plot, we show that DAAM maps for adjectives, verbs, nouns, and proper nouns attain close to or slightly above “good,” whereas the ones for numerals and adverbs are closer to “fair.” This agrees with the generated examples in Figure 4, where numerals (see the giraffes’ edges) and adverbs (feet and ground motion blur) are less intuitively highlighted than adjectives (blue part of teapot), verbs (fists and legs in running form), and nouns. Nevertheless, the proportion of ratings falling between fair and excellent are above 80% for numerals and adverbs and 90% for the rest—see the bottom of Figure 3. We thus conclude that DAAM produces plausible maps for each interpretable part of speech.

One anticipated criticism is that different heat maps may explain the same word, making a qualitative comparison less meaningful. In Figure 4, “quickly” could conceivably explain “running” too. We concede to this, but our motivation is not to compare *quality* but rather to demonstrate *plausibility*. Without these experiments, the DAAM maps for words like “running” and “blue” could very well have been meaningless blotches.

#	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	<u>12.</u>	55.4
9	nsubj	76.4	63.9	<u>12.</u>	52.2
10	amod	62.4	77.6	<u>15.</u>	51.1
11	nmod:of	73.5	57.9	<u>16.</u>	47.5
12	obj	75.6	46.3	<u>29.</u>	55.4
14	Coreferent word pairs	84.8	77.4	7.4	66.6

Table 2: Head-dependent DAAM map overlap statistics across the ten most common relations in COCO. Bolded are the dominant maps, where the absolute difference Δ between mIoD and mIoH exceeds 10 points. All bolded numbers are significant ($p < 0.01$).

4 Visuosyntactic Analysis

Equipped with DAAM, we now study how syntax relates to generated pixels. We characterize pairwise interactions between head-dependent DAAM maps, augmenting previous sections and helping to form hypotheses for further research.

Setup. We randomly sampled 1,000 prompts from COCO, performed dependency parsing with CoreNLP (Manning et al., 2014), and generated an image for each prompt and DAAM maps for all words. We constrained our examination to the top-10 most common relations, resulting in 8,000 head-dependent pairs. Following Section 3.1, we then binarized the maps to quantify head-dependent interactions with set-based similarity statistics. We computed three statistics between the DAAM map of the head and that of the dependent: first, the mean visual intersection area over the union (mIoU), i.e., $\frac{|A \cap B|}{|A \cup B|}$; second, the mean intersection over the dependent (mIoD; $\frac{|A \cap B|}{|B|}$); and third, the intersection over the head (mIoH; $\frac{|A \cap B|}{|A|}$). mIoU measures similarity, and the difference between mIoD and mIoH quantifies dominance. If $\text{mIoD} > \text{mIoH}$, then the head contains (dominates) the dependent more, and vice versa—see Appendix B for a visual tutorial.

Results. We present our quantitative results in Table 2 and examples in Figure 5. For baselines, we computed overlap statistics for unrelated pairs of words and all head-dependent pairs. Unsurprisingly, both baselines show moderate similarity and no dominance (43–48 mIoU, $\Delta \leq 1$; rows 1–2). For syntactic relations, we observe no dominance for noun compounds (row 3), which is ex-

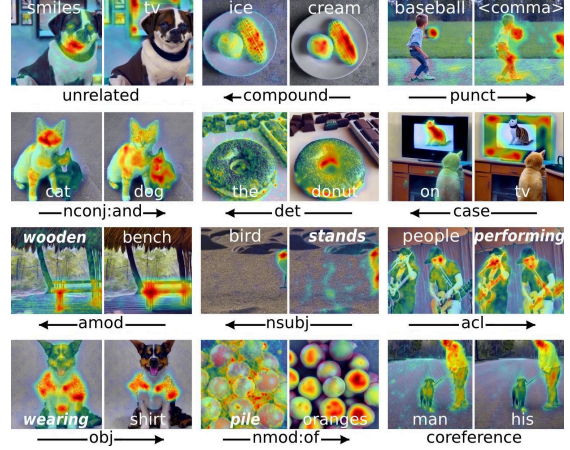


Figure 5: Twelve example pairs of DAAM maps, with the dominant word in bold, if present for the relation. Note that the visualization scale is normalized for each image since our purpose is to study the *spatial locality* of attribution *conditioned on the word*. For example, the absolute magnitude for the comma above is weak.

pected since the two nouns complement one another (e.g., “ice cream”). Punctuation and articles (punct, det; rows 4 and 6) also lack dominance, possibly from having little semantic meaning and attending broadly across the image (Figure 5, top right). This resembles findings in Kovaleva et al. (2019), who note BERT’s (Devlin et al., 2019) punctuation to attend widely. For nouns connected with “and” (row 5), the maps overlap less (38.7 mIoU vs. 50+), likely due to visual separation (e.g., “cat and dog”). However, the overlap is still far above zero, which we attribute partially to feature entanglement, further explored in Section 5.1.

Starting at row 8, we arrive at pairs where one map dominates the other. A group in core arguments arises (nsubj, obj), where the head word dominates the noun subject’s or object’s map (12–29-point Δ), perhaps since verbs contextualize both the subject and the object in its surroundings—see the middle of and bottom left of Fig. 5. We observe another group in nominal dependents (nmod:of, amod, acl), where nmod:of mostly points to collective nouns (e.g., “pile of oranges”), whose dominance is intuitive. In contrast, adjectival modifiers (amod) behave counterintuitively, where descriptive adjectives (dependents) visually dominate the nouns they modify ($\Delta \approx 15$). We instead expect objects to contain their attributes, but this is not the case. We again ascribe this to entanglement, elucidated in Section 5.2. Lastly, coreferent word pairs exhibit the highest overlap out of all relations (66.6 mIoU), indicating attention to the same referent.

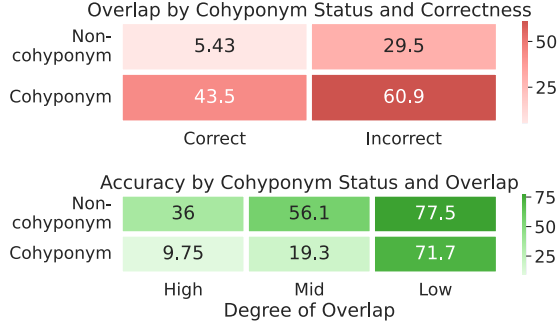


Figure 6: Above: DAAM map overlap in mean IoU, subdivided by cohyponym status and correctness; below: generation accuracy, subdivided by cohyponym status and amount of overlap.

5 Visuosemantic Analyses

5.1 Cohyponym Entanglement

To further study the large $nconj$ and overlap found in Section 4, we hypothesize that semantically similar words in a prompt have worse generation quality, where only one of the words is generated in the image, not all.

Setup. To test our hypothesis, we used WordNet (Miller, 1995) to construct a hierarchical ontology expressing semantic fields over COCO’s 80 visual objects, of which 28 have at least one other cohyponym across 16 distinct hypernyms (as listed in the appendix). Next, we used the prompt template, “a(n) <noun> and a(n) <noun>,” depicting two distinct things, to generate our dataset. Using our ontology, we randomly sampled two cohyponyms 50% of the time and two non-cohyponyms other times, producing 1,000 prompts from the template (e.g., “a **giraffe** and a **zebra**,” “a **cake** and a **bus**”). We generated an image for each prompt, then asked three unique annotators per image to select which objects were present, given the 28 words. We manually verified the image–label pairs, rejecting and republishing incorrect ones. Finally, we marked the overall label for each image as the top two most commonly picked nouns, ties broken by submission order. We considered generations correct if both words in the prompt were present in the image. For more setup details, see the appendix.

Results. Overall, the non-cohyponym set attains a generation accuracy of 61% and the cohyponym set 52%, statistically significant at the 99% level according to the exact test, supporting our hypothesis. To see if DAAM assists in explaining these effects, we compute binarized DAAM maps ($\tau = 0.4$, the best value from Sec. 3.1) for both words and quan-

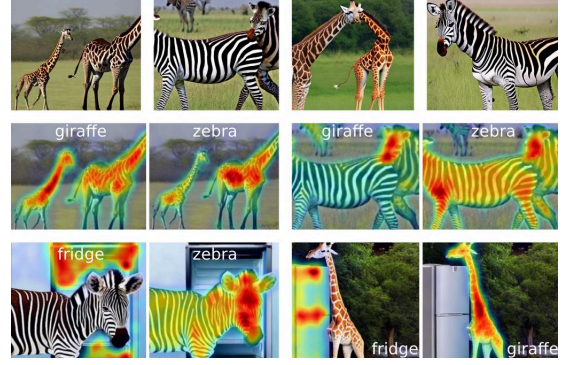


Figure 7: Rows starting from the top: generated images for cohyponyms “a giraffe and a zebra,” heat maps for the first two images, and heat maps for non-cohyponymic zebra–fridge and giraffe–fridge prompts.



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.”

tify the amount of overlap with IoU. We find that the mIoU for cohyponyms and non-cohyponyms are 46.7 and 22.9, suggesting entangled attention and composition. In the top of Figure 6, we further group the mIoU by cohyponym status and correctness, finding that incorrectness and cohyponymy independently increase the overlap. In the bottom subplot, we show that the amount of overlap (mIoU) differentiates correctness, with the low, mid, and high cutoff points set at ≤ 0.4 , $0.4-0.6$, and ≥ 0.6 , following statistics in Section 4. We observe accuracy to be much better on pairs with low overlap (71.7–77.5%) than those with high overlap (9.8–36%). We present some example generations and maps in Figure 7, which supports our results.

5.2 Adjectival Entanglement

We examine prompts where a noun’s modifying adjective attends too broadly across the image. We start with an initial seed prompt of the form, “a <adj> <noun> <verb phrase>,” then vary the adjective to see how the image changes. If there is no entanglement, then the background *should*



Figure 9: A DAAM map and generated images for “a <adj> car driving down the streets,” above images of the cropped background, saturated for visualization.

not gain attributes pertaining to that adjective. To remove scene layout as a confounder, we fix all cross-attention maps to those of the seed prompt, which Hertz et al. (2022) show to equalize layout.

Our first case is, “a {rusty, metallic, wooden} shovel sitting in a clean shed,” “rusty” being the seed adjective. As shown in Figure 8, the DAAM map for “rusty” attends broadly, and the background for “rusty” is surely not clean. When we change the adjective to “metallic” and “wooden,” the shed changes along with it, becoming grey and wooden, indicating entanglement. Similar observations apply to our second case, “a {bumpy, smooth, spiky} ball rolling down a hill,” where “bumpy” produces rugged ground, “smooth” flatter ground, and “spiky” blades of grass. In our third case, we study color adjectives using “a {blue, green, red} car driving down the streets,” presented in Figure 9. We discover the same phenomena, with the difference that these prompts lead to *quantifiable* notions of adjectival entanglement. For, say, “green,” we can conceivably measure the amount of additional green hue in the background, with the car cropped out—see bottom row. A caveat is that entanglement is not necessarily unwanted; for instance, rusty shovels likely belong in rusted areas. It strongly depends on the use case of the model.

6 Related Work and Future Directions

The primary area of this work is in understanding neural networks from the perspective of computational linguistics, with the goal of better informing future research. A large body of relevant papers exists, where researchers apply textual perturbation (Wallace et al., 2019), attention visualization (Vig, 2019; Kovaleva et al., 2019; Shimaoka et al., 2016), and information bottlenecks (Jiang et al., 2020) to relate important input tokens to the outputs of large language models. Others explicitly test for linguistic constructs within models, such as Hendricks and Nematzadeh’s (2021) probing

of vision transformers for verb understanding and Ilinykh and Dobnik’s (2022) examination of visual grounding in image-to-text transformers. Our distinction is that we carry out an attributive analysis in the space of generative diffusion models, as the pixel output relates to syntax and semantics. As a future extension, we plan to assess the unsupervised parsing ability of Stable Diffusion with syntactic–geometric probes, similar to Hewitt and Manning’s (2019) work in BERT.

The intersection of text-to-image generation and natural language processing is certainly substantial. In the context of enhancing diffusion models using prompt engineering, Hertz et al. (2022) cement cross-attention maps for the purpose of precision-editing generated images using text, and Woolf (2022) proposes negative prompts for removing undesirable, scene-wide attributes. Related as well are works for generative adversarial networks, where Karras et al. (2019) and Materzyńska et al. (2022) disentangle various features such as style and spelling. Along this vein, our work exposes more entanglement in cohyponyms and adjectives. A future line of work is to disentangle such concepts and improve generative quality.

Last but not least are semantic segmentation works in computer vision. Generally, researchers start with a backbone encoder, attach decoders, and then optimize the model in its entirety end-to-end on a segmentation dataset (Cheng et al., 2022), unless the context is unsupervised, in which case one uses contrastive objectives and clustering (Cho et al., 2021; Hamilton et al., 2021). Toward this, DAAM could potentially provide encoder features in a segmentation pipeline, where its strong raw baseline numbers suggest the presence of valuable latent representations in Stable Diffusion.

7 Conclusions

In this paper, we study visiolinguistic phenomena in diffusion models by interpreting word–pixel cross-attention maps. We prove the correctness of our attribution method, DAAM, through a quantitative semantic segmentation task and a qualitative generalized attribution study. We apply DAAM to assess how syntactic relations translate to visual interactions, finding that certain maps of heads inappropriately subsume their dependents’. We use these findings to form hypotheses about feature entanglement, showing that cohyponyms are jumbled and adjectives attend too broadly.

Acknowledgments

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, companies sponsoring the Vector Institute, and the HuggingFace team. In particular, we would like to thank Aleksandra (Ola) Piktus, who helped us get a community grant for our public demonstration on HuggingFace spaces.

References

- David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the robustness of interpretability methods. *arXiv:1806.08049*.
- Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. 2019. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv:1906.07155*.
- Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. 2021. PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT’s attention. In *Proceedings of BlackboxNLP*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.
- Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. 2021. Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snaveley, and William T. Freeman. 2021. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*.
- Charles R. Harris, K. Jarrod Millman, Stéfan J. Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, et al. 2020. Array programming with NumPy. *Nature*.
- David J. Hauser and Norbert Schwarz. 2016. Attentive turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv:2208.01626*.
- John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.
- Nikolai Ilinykh and Simon Dobnik. 2022. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*.
- Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. 2020. Inserting information bottlenecks for attribution in transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.
- Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv:1312.6114*.
- Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

- Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *arXiv:2206.00927*.
- Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the Association for Computational Linguistics: System Demonstrations*.
- Joanna Materzyńska, Antonio Torralba, and David Bau. 2022. Disentangling visual and written concepts in CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- George A. Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv:2205.11487*.
- Christoph Schuhmann, Romain Beaumont, Cade W Gordon, Ross Wightman, Theo Coombes, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Sonse Shimaoka, Pontus Stenetorp, Kentaro Inui, and Sebastian Riedel. 2016. Neural architectures for fine-grained entity type classification. *arXiv preprint arXiv:1606.01341*.
- Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- Jesse Vig. 2019. BertViz: A tool for visualizing multi-head self-attention in the BERT model. In *ICLR Workshop: Debugging Machine Learning Models*.
- Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. [Diffusers: State-of-the-art diffusion models](#).
- Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*.
- Max Woolf. 2022. [Stable Diffusion 2.0 and the importance of negative prompts for good results](#).
- Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. PhraseCut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2022. Diffusion models: A comprehensive survey of methods and applications. *arXiv:2209.00796*.



Figure 10: Example ground-truth segmentation masks from four prompt-image pairs.

A Supplements for Attribution Analyses

A.1 Object Attribution

Generation setup. For all generated images in the paper, we ran the Stable Diffusion 2.0 base model (512 by 512 pixels) with 30 inference steps, the default 7.5 classifier guidance score, and the state-of-the-art DPM solver. We automatically filtered out all offensive images, against which the 2.0 model has both training-time and after-inference protection. We also steered clear of offensive prompts, which were absent to start with in COCO. Our computational environment consisted of PyTorch 1.11.0 and CUDA 11.4, running on Titan RTX and A6000 graphics cards.

Segmentation process. To draw the ground-truth segmentation masks, we used the object selection tool, the quick selection tool, and the brush from Adobe Photoshop CC 2022 to fill in a black mask for each area corresponding to a present noun. We then exported each mask (without the background image) as a binary PNG mask and attached it to the relevant noun—see Figure 10 for some examples. Two trained annotators worked on the total set of 200 image-prompt pairs, with one completing 75 on each dataset and the other 25 on each.

Layer and time step ablation. We conducted ablation studies to see if summing across *all* time steps and layers, as in Eqn. 6, is necessary. We searched both sides of the summation: for one study, we restricted DAAM to $j \leq j^*$, as $j^* = 1 \rightarrow T$; for its dual study, we constrained $j \geq j^*$. We applied the same methods to layer resolution,

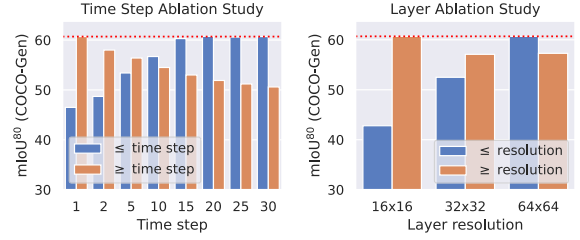


Figure 11: On the left, taking the first n or last n time steps; on the right, the equivalent for layer resolution.

i.e., c^i . We present our results in Fig. 11, which suggests that all time steps and layers contribute positively to segmentation quality.

A.2 Generalized Attribution

Annotation process. We designed our annotation user interfaces (UIs) for Amazon MTurk, a popular crowdsourcing platform, where we submitted a batch job requiring three unique annotators at the master level to complete each task. We presented the UI pictured in Figure 12, asking them to rate the relevance of the red word to the highlighted area in the image. If the image was too poor or if the word was missing, they could also choose options 6 and 7.

To filter out low-quality or inattentive annotators, we randomly asked workers to interpret punctuation, such as periods. Since these tokens are self-evidently too abstract and missing in the image, we removed workers who didn’t select one of those two options. However, we found overall attention to be high, having a reject rate of less than 2% of the tasks, consistent with Hauser and Schwarz’s (2016) findings that MTurk users outperform subject pool participants. We show response statistics in Figure 13, where adpositions, coordinating conjunctions, participles, punctuation, and articles have high non-interpretable rates.

B Supplements for Syntactic Analyses

Measures of overlap. We use three measures of overlap to characterize head-dependent map interactions: mean intersection over union (mIoU), intersection over the dependent (mIoD), and intersection over the head (mIoH). When mIoU is high, the maps overlap greatly; when mIoD is high but mIoH is low, the head map occupies more of the dependent than the dependent does the head; when the opposite is true, the dependent occupies more.

How well does the **highlighted parts** of the image reflect the meaning of the **red word** in the sentence?

presents surround a **christmas tree** in front of a fireplace.

Don't be afraid to click "Word too abstract" or "Image itself is poor" if it's actually the case.

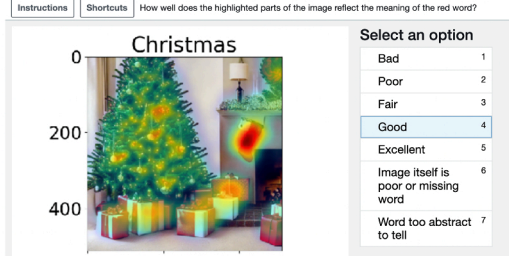


Figure 12: Annotation UI for generalized attribution.

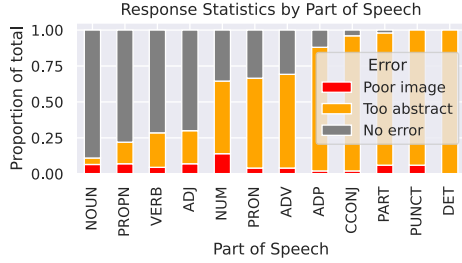


Figure 13: Response statistics by part of speech.

Concretely, given a sequence of binarized DAAM map pairs $\{(D_{(i1)}^{\mathbb{I}_\tau}, D_{(i2)}^{\mathbb{I}_\tau})\}_{i=1}^n$, where $i1$ are **dependent** indices and $i2$ **head** indices, we compute mIoU as

$$\frac{1}{n} \sum_{i=1}^n \frac{\sum_{(x,y)} D_{(i1)}^{\mathbb{I}_\tau}[x,y] \wedge D_{(i2)}^{\mathbb{I}_\tau}[x,y]}{\sum_{(x,y)} D_{(i1)}^{\mathbb{I}_\tau}[x,y] \vee D_{(i2)}^{\mathbb{I}_\tau}[x,y]}, \quad (8)$$

where \wedge is the logical-and operator, returning 1 if both sides are 1, 0 otherwise, and \vee the logical-or operator, returning 1 if at least one operand is 1, and 0 otherwise. Let the top part of the inner fraction be the intersection, or INT for short. Define mIoD as

$$\frac{1}{n} \sum_{i=1}^n \frac{\text{INT}}{\sum_{(x,y)} D_{(i1)}^{\mathbb{I}_\tau}[x,y]}, \quad (9)$$

and mIoH as

$$\frac{1}{n} \sum_{i=1}^n \frac{\text{INT}}{\sum_{(x,y)} D_{(i2)}^{\mathbb{I}_\tau}[x,y]}, \quad (10)$$

We visually present our mIoD and mIoH statistics in Figure 14.

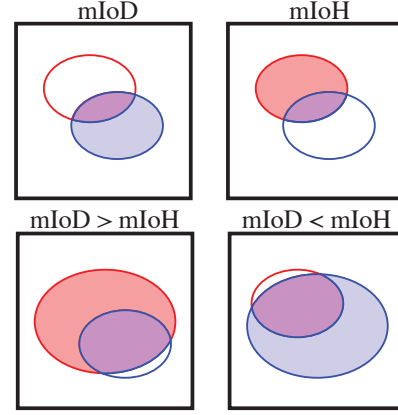
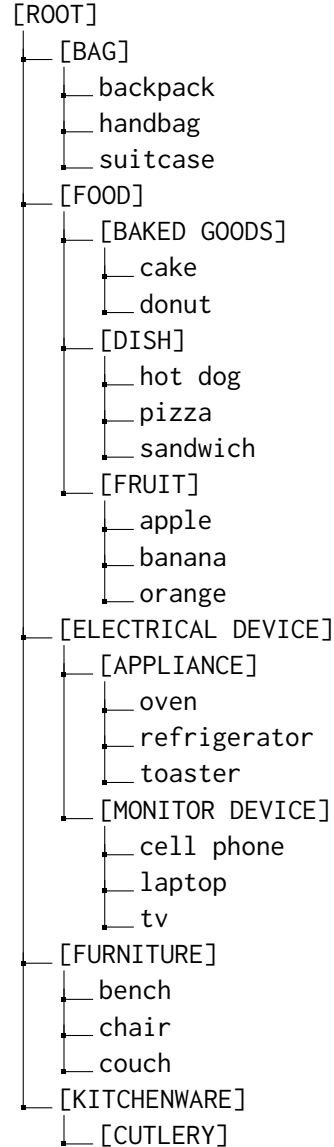


Figure 14: Portrayals of mIoD, mIoH, and different forms of overlap.

C Supplements for Semantic Analyses

Semantic relation ontology. We present our relation ontology below, continued on the next page:



Please identify all objects in the image. There might be multiple. Please read the labels carefully.

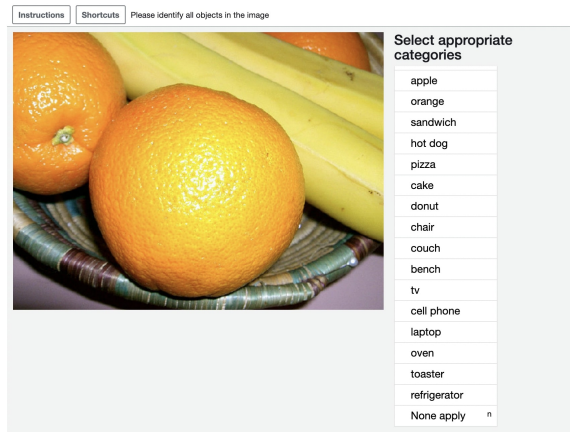
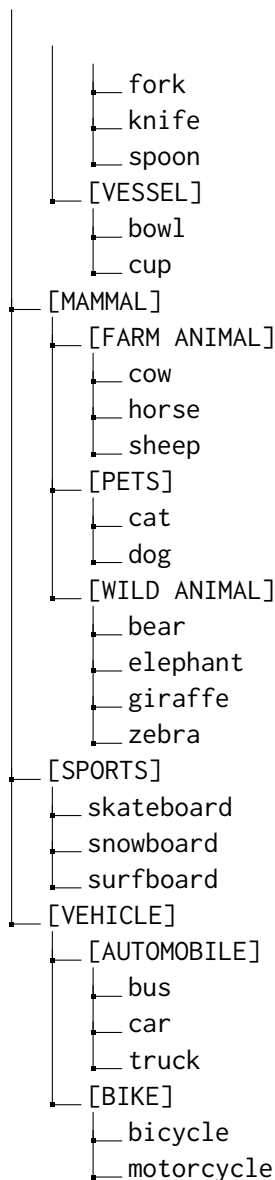


Figure 15: The annotation UI for cohyponym entanglement, asking annotators to pick the present objects.



Cohyponym annotation process. Similar to the generalized attribution annotation process, we de-

signed our annotation UIs for Amazon MTurk. We submitted a job requiring three unique annotators at the master level to complete each task. We presented to them the UI shown in Figure 15. We manually verified each response, removing workers whose quality was consistently poor. This included workers who didn't include all objects generated. Overall, the worker quality was exceptional, with a reject rate below 2%. Out of a pool of 30 workers, no single worker annotated more than 16% of the examples.