

# Picturing Ambiguity: A Visual Twist on the Winograd Schema Challenge

Brendan Park\*, Madeline Janecek\*, Naser Ezzati-Jivan, Yifeng Li, and Ali Emami  
Brock University, St. Catharines, Ontario, Canada  
{bp18ul, mj17th, nezzatijivan, yli2, aemami}@brocku.ca

## Abstract

Large Language Models (LLMs) have demonstrated remarkable success in tasks like the Winograd Schema Challenge (WSC), showcasing advanced textual common-sense reasoning. However, applying this reasoning to multi-modal domains, where understanding text *and* images together is essential, remains a substantial challenge. To address this, we introduce WINOVIS, a novel dataset specifically designed to probe text-to-image models on pronoun disambiguation within multimodal contexts. Utilizing GPT-4 for prompt generation and Diffusion Attentive Attribution Maps (DAAM) for heatmap analysis, we propose a novel evaluation framework that isolates the models' ability in pronoun disambiguation from other visual processing challenges. Evaluation of successive model versions reveals that, despite incremental advancements, Stable Diffusion 2.0 achieves a precision of 56.7% on WINOVIS, showing minimal improvement from past iterations and only marginally surpassing random guessing. Further error analysis identifies important areas for future research aimed at advancing text-to-image models in their ability to interpret and interact with the complex visual world.

## 1 Introduction

The interpretation of ambiguous constructs in language is crucial for assessing common-sense reasoning, with the Winograd Schema Challenge (WSC) (Levesque et al., 2011; Winograd, 1972) significantly influencing the evaluation of natural language understanding models. Advances in transformer-based architectures have led Large Language Models (LLMs) to achieve impressive results on WSC-based tasks, approaching near-human performance (Brown et al., 2020; Sakaguchi et al., 2020; Kocijan et al., 2023).

\*Equal contribution.

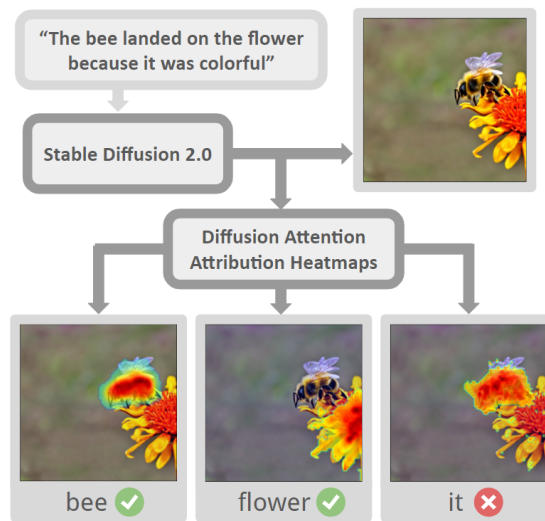


Figure 1: A representative output from Stable Diffusion 2.0 on a WINOVIS instance. The Diffusion Attentive Attribution Maps (DAAM) clarify the model's focus for different terms and the correctness of its interpretation: correctly identifying 'bee' and 'flower' but erroneously associating 'it' with the bee instead of the flower.

Extending common-sense reasoning into multi-modal domains, especially disambiguation tasks, is a persisting challenge. Despite the ability of models like Google's Imagen (Saharia et al., 2022), OpenAI's DALL-E 2 (Ramesh et al., 2022), and Stability AI's recently open-sourced Stable Diffusion (Rombach et al., 2022) to create visually compelling images from text, their interpretability—essential for deciphering the models' reasoning processes—is notably limited (Tang et al., 2023). This gap restricts the development of tools for visuals that match complex texts, reducing model effectiveness when deployed in areas like education and digital media, where text-image integration is essential (Dehouche and Dehouche, 2023; Hattori and Takahara, 2023).

Our response to this challenge is WINOVIS, a dataset aimed at probing text-to-image models' common-sense reasoning capabilities through pronoun disambiguation within multimodal scenar-

ios. WINOVIS not only tests models’ ability to distinguish entities within the generated images, but also examines how these models associate pronouns with the correct referents, a nuanced aspect of common-sense reasoning that has been overlooked. As depicted in the WINOVIS example in Figure 1, while newer Stable Diffusion models can accurately separate entities within an image, they fail to correctly associate the pronoun ‘it’ with the intended referent, revealing the subtleties and potential gaps in multimodal common-sense reasoning.

The development of WINOVIS leveraged the generative power of GPT-4 (OpenAI, 2023; Gilardi et al., 2023), using a methodical approach to create and refine prompts that elicit common-sense reasoning visually. This process included a complete manual review to ensure each scenario’s clarity and relevance for the disambiguation task. Moreover, we introduce a novel evaluation framework that distinguishes between models’ pronoun disambiguation proficiency from their handling of visual processing challenges, such as susceptibility to typographic attacks (Goh et al., 2021) and semantic entanglement (Wu et al., 2023).

Our contributions are summarized as follows:

- **WSC-Adapted Multimodal Dataset (WINOVIS):** A dataset of 500 scenarios for benchmarking text-to-image models’ pronoun disambiguation abilities within a visual context.<sup>1</sup>
- **Novel Evaluation Framework for Multimodal Disambiguation:** Metrics and methods designed to isolate pronoun resolution from other visual processing challenges, advancing the understanding of models’ common-sense reasoning.
- **Insight into Stable Diffusion’s Common-Sense Reasoning:** A critical analysis revealing that even state-of-the-art models like Stable Diffusion 2.0 fall significantly short of human-level performance.

## 2 Background

### 2.1 Latent Diffusion in Image Generation

Latent diffusion models (LDMs) represent a class of generative models designed to synthesize

<sup>1</sup>The dataset has been made available at <https://github.com/bpark2/WinoVis>.

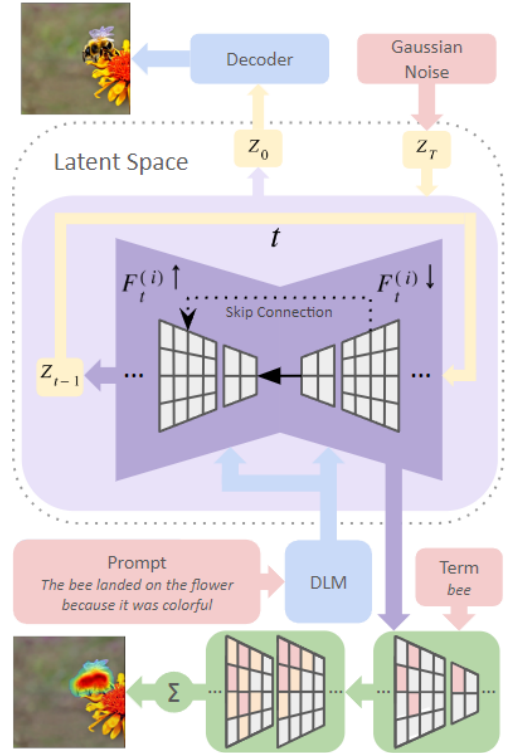


Figure 2: A visual overview of the Stable Diffusion architecture, as well as the Diffusion Attention Attribution Map (DAAM) generation process.

images by progressively refining random noise. A prominent example is Stable Diffusion (Rombach et al., 2022), a text-to-image LDM optimized to generate images from textual prompts. Stable Diffusion integrates three primary components: a deep language model that extracts semantic embeddings from textual prompts; an encoder-decoder architecture for encoding images into latent space representations and decoding them back; and a neural network that is responsible for mean-prediction (Ho et al., 2020) (denoted as  $\mu_\theta(z, \mathbf{y}, t)$ ), noise-prediction (Ho et al., 2020) (denoted as  $\epsilon_\theta(z, \mathbf{y}, t)$ ), or score-prediction (Song and Ermon, 2019) (denoted as  $s_\theta(z, \mathbf{y}, t)$ ). This network is trained on image and text pairs  $x$  and  $\mathbf{y}$ . During training, which aims to maximize the evidence lower bound (ELBO) (Sohl-Dickstein et al., 2015), the image is initially encoded to  $z_0$ , marking the start of the forward diffusion process, formalized as:

$$\begin{aligned} p(z_t|z_{t-1}) &= \mathcal{N}(z_t|\sqrt{\alpha_t}z_{t-1}, (1 - \alpha_t)\mathbf{I}) \\ &= \mathcal{N}(z_t|\sqrt{1 - \beta_t}z_{t-1}, \beta_t\mathbf{I}), \end{aligned}$$

where  $z_t$  denotes the latent variable at step  $t$ , with  $\beta_t = 1 - \alpha_t$  as the noise schedule hyperparameter, and  $\mathbf{I}$  the identity matrix. The U-Net architecture (Ronneberger et al., 2015), used for denoising, iter-

actively reverses the diffusion through:

$$\begin{aligned} p(\mathbf{z}_{t-1}|\mathbf{z}_t) &= \mathcal{N}(\mathbf{z}_{t-1}|\mu_\theta(\mathbf{z}_t, \mathbf{y}, t), \sigma_t^2 \mathbf{I}) \\ &= \mathcal{N}\left(\mathbf{z}_{t-1} \mid \frac{\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(\mathbf{z}_t, \mathbf{y}, t)}{\sqrt{\alpha_t}}, \sigma_t^2 \mathbf{I}\right) \end{aligned}$$

with  $\sigma_t^2$  as the reverse process noise variance. Cross-attention in the U-Net layers aligns  $\mathbf{z}_t$  with  $\mathbf{y}$ . For conditional generation, the process starts with Gaussian noise  $\mathbf{z}_T$ , conditioned on text  $\mathbf{y}$ , and refines through reverse diffusion, resembling Langevin dynamics (Welling and Teh, 2011). For instance, using the score function, we have

$$\mathbf{z}_{t-1}^{(j)} = \mathbf{z}_{t-1}^{(j-1)} + \frac{\alpha_t}{2} s_\theta(\mathbf{z}_{t-1}^{(j-1)}, \mathbf{y}, t) + \sqrt{\alpha_t} \epsilon_j,$$

where  $j = 1, \dots, J$ ,  $J$  is the number of Langevin steps,  $\mathbf{z}_{t-1}^{(0)} = \mathbf{z}_t$ ,  $\mathbf{z}_{t-1}^{(J)} = \mathbf{z}_{t-1}^{(J)}$ , and  $\epsilon_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . The denoised  $\mathbf{z}_0$  generates the final image, such as the one exemplified in Figure 2 given a WINOVIS instance.

## 2.2 Diffusion Attentive Attribution Maps

The Diffusion Attentive Attribution Map (DAAM) technique facilitates interpretability of the influence that different tokens in a prompt have on the image generated by Stable Diffusion models (Tang et al., 2023). This approach capitalizes on the multi-head cross-attention mechanism (Vaswani et al., 2023), aggregating attention scores from both downsampling and upsampling stages within the U-Net architecture. The attention scores, denoted as  $F_t^{(i)\downarrow}$  for downsampling and  $F_t^{(i)\uparrow}$  for upsampling, link specific words from the prompt to image regions, signified by coordinates  $(x, y)$ , across different heads  $(i)$  and layers  $(l)$ .

To synthesize a comprehensive heatmap from these attention scores, DAAM applies a spatial normalization procedure, scaling the attention scores for the  $k$ -th word to match the original image size and summing them across all attention heads  $(i)$ , layers  $(l)$ , and time steps  $(t)$ :

$$D_k[x, y] = \sum_{i,t,l} \left( F_t^{(i)\downarrow}[x, y, l, k] + F_t^{(i)\uparrow}[x, y, l, k] \right)$$

where  $F_t^{(i)\downarrow}[x, y, l, k]$  and  $F_t^{(i)\uparrow}[x, y, l, k]$  represent the bicubically upsampled attention scores for the downsampling and upsampling pathways, respectively.

DAAM can therefore offer a visual method to evaluate how Stable Diffusion performs pronoun disambiguation, by illustrating where the model concentrates its attention in relation to textual prompts. By examining these visualizations, as demonstrated in Figure 1, we can discern the model’s implicit strategies for linking pronouns with their correct referents.

## 3 Constructing WINOVIS

In this section, we detail the methodology behind the creation of WINOVIS, a novel dataset engineered to assess the pronoun disambiguation capabilities of text-to-image models. The integration of GPT-4 (OpenAI, 2023; Gilardi et al., 2023) into our dataset generation workflow allowed for significant streamlining of the creation process, achieving reductions in both cost and time, enhancing reproducibility, and reducing the incidence of human error. Our Corpus Construction Cycle consists of two main stages: 1) The GPT Prompt Cycle; and 2) The Manual Filter Process. A full visualization of the process is provided in Appendix Figure 10.

### 3.1 Corpus Construction Cycle

**Step 1: GPT Prompt Cycle** In developing WINOVIS, we aimed to adapt the Winograd Schema Challenge (WSC) for visual interpretation. This required avoiding the creation of instances that were visually ambiguous, lacked clear visual contexts, or logically didn’t necessitate both entities. Table 1 showcases problematic examples from the WSC alongside those of WINOVIS.

Our iterative prompting process with GPT-4, as outlined in Appendix Table 5, included both successful and problematic few-shot examples to refine the desired outcomes. This approach, detailed entirely in Appendix Table 6, helped clarify what constitutes an acceptable instance. To enhance logical reasoning, we employed a Chain-of-Thought (CoT) (Wei et al., 2023) strategy, further described in Appendix Table 5 under *CoT*. Querying instances in batches of ten ensured a varied yet coherent collection while minimizing duplicates.

**Step 2: Manual Filter Process** After GPT-4 generated the initial set of instances, a manual review was conducted to filter out instances that failed to meet our study’s criteria:

- **Textual Ambiguity:** If a prompt could not be easily disambiguated by all annotators it was excluded.

WSV	Disparate Entities	The <b>thief</b> stole the <b>diamond</b> because <b>it</b> was valuable. ( $A = \text{diamond}$ )
	Distinct Entities (Age)	The <b>man</b> carried the <b>child</b> because <b>he</b> was tired. ( $A = \text{child}$ )
	Distinct Entities (Role)	The <b>king</b> banished the <b>jester</b> because <b>he</b> was annoying. ( $A = \text{jester}$ )
WSC	Visually Ambiguous	<b>Pete</b> envies <b>Martin</b> because <b>he</b> is very successful. ( $A = \text{Martin}$ )
	Entity Exclusion	<b>Jane</b> knocked on <b>Susan</b> 's door, but there was no answer. <b>She</b> was out. ( $A = \text{Susan}$ )
Filtered	Textually Ambiguous	The <b>dog</b> could not catch the <b>squirrel</b> because <b>it</b> was small. ( $A = ?$ )
	Illogical	The <b>fisherman</b> cast the <b>net</b> because <b>it</b> was full of holes. ( $A = \text{net}$ )
	Visually Indistinctive	The <b>wrestler</b> defeated the <b>opponent</b> because <b>he</b> was weak. ( $A = \text{enemy}$ )
	Redundant Entries	<b>Anthony</b> admired <b>James</b> because <b>he</b> was talented. ( $A = \text{James}$ ) <b>Ryan</b> respected <b>Andrew</b> because <b>he</b> was talented. ( $A = \text{Andrew}$ )

Table 1: Examples from the WINOVIS (WSV) dataset alongside instances from the Winograd Schema Challenge (WSC) dataset and those excluded through manual filtering. In each case, the correct entity is denoted by  $A$ .

- **Illogical Content:** Removed if containing nonsensical or irrelevant information.
- **Visual Indistinctiveness:** Omitted when entities lacked clear visual differentiation, essential for accurate entity-pronoun association.
- **Redundancy:** To ensure a broad range of scenarios, instances that were too similar in content or structure were excluded.

This manual filtering ensured that each prompt included in WINOVIS is well suited for evaluating text-to-image models. Examples of excluded instances for each criterion are provided in Table 1. This review cycle was repeated, refining the selection until achieving a diverse and quality set.

### 3.2 Dataset Characteristics

Each sample of WINOVIS contains a pronoun resolution prompt, a specification of the ambiguous pronoun, an excerpt containing the pronoun, the two referent entity options, the correct referent, and a justification for why the correct entity should be associated with the pronoun.

**Disparate and Distinct Entities** The instances within WINOVIS fall into two broad categories: *Disparate Entities* and *Distinct Entities*. Disparate Entities encompass scenarios with significantly different subjects, such as those across species or object classes (e.g., a person vs. a dog, or a car vs. a tree). Distinct Entities, while sharing some similarities, are visually distinguishable by attributes like age (a mother and child), role (a cop and a thief), or other descriptors, posing more nuanced challenges for pronoun resolution.

WINOVIS is primarily designed to evaluate a model’s *common-sense* reasoning capabilities, rather than to pose a significant challenge. Consequently, a substantial portion of its instances

(84.2%) involve disparate entities. To assess the model in a more demanding context, the remaining 15.8% of the instances feature distinct entities.

**Context Types** To further examine the comprehensiveness of WINOVIS as a commonsense reasoning benchmark, we categorized each prompt based on the contextual details it provides to link the correct referent to the pronoun. The four contextual categories present in WINOVIS are:

- **Visually Tangible:** These entries contain descriptions that should have a clear visual impact on the associated referent.
- **Emotional:** These entries describe the emotional or mental state of the referent, which, although more subtle, would still affect the referent’s appearance.
- **Characteristic:** These entries include details pertaining to a referent’s personality or nature. While less visually tangible, these details may affect the associated referent’s finer details.
- **Visually Intangible:** These entries involve attributes with minimal to no visual impact on the referent, such as taste, speed, or sound. These entries assess the model’s understanding of purely textual input.

We argue that proficiency in pronoun disambiguation requires the capacity to effectively leverage all four context types. Therefore, WINOVIS includes prompts from each category, providing a comprehensive assessment of a model’s capabilities. Examples and the distribution of each category within WINOVIS are shown in Appendix Table 7.

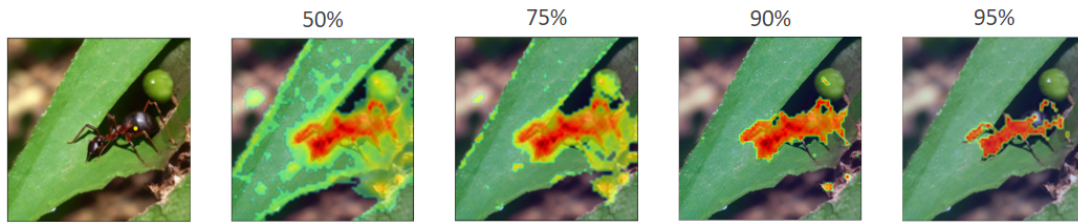


Figure 3: The results of different heatmap thresholds for the prompt “The ant could not carry the leaf because it was too weak” and the term ‘it’.

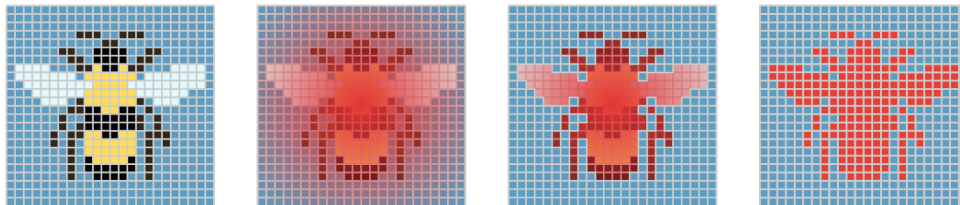


Figure 4: Illustrative example of thresholding on attention maps, progressing through stages to apply a 90<sup>th</sup> percentile threshold, resulting in a binary mask that accentuates key attention regions.

#### 4 Evaluating Pronoun Disambiguation in WINOVIS

This section outlines our systematic pipeline to evaluate the capability of Stable Diffusion models to accurately disambiguate pronouns within the context of WINOVIS. Our pipeline comprises four stages: 1) Filtering out captioned images to remove visual representations that include embedded text; 2) Enhancing the clarity of distributed attention attribution maps through noise reduction; 3) Excluding images with significant heatmap overlap between referent entities from our analysis; and 4) Determining the model’s final pronoun association by establishing a decision boundary.

**Step 1: Caption Filtering** Text-to-image LDMs sometimes generate images where prompt text appears visually, resulting in ‘captioned’ images. These images erroneously direct a term’s attribution to this text, complicating the assessment of the model’s visual common-sense reasoning. An example of this is shown in Appendix Figure 11.

We therefore specifically excluded captioned images from the analysis set of a studied model, prioritizing those yielding visuals strictly relevant to common-sense interpretation. This exclusion is based on the premise that visual common-sense reasoning should be assessed purely on the model’s ability to interpret and generate relevant visual content, without the confounding influence of embedded text. Data on the frequency of prompts resulting in captioned images is detailed in Table 2.

#### Step 2: Noise Reduction in Attention Maps

To ensure attention heatmaps clearly reflect the model’s focus, we apply a 90<sup>th</sup> percentile thresholding technique to the heatmaps generated from WINOVIS prompts. This approach filters out the bottom 90% of attention scores, considered as noise, and retains only the highest-intensity areas indicative of the model’s primary interest. This 90<sup>th</sup> percentile threshold was chosen after extensive testing with various thresholds. It was found to be the most effective in balancing the elimination of irrelevant noise while preserving the focal points crucial for understanding the model’s interpretation of the prompt. Thresholds below the 90<sup>th</sup> percentile included too much noise, while higher thresholds risked omitting significant details.

Following this, areas surpassing the threshold are converted into binary masks, delineating significant attention (‘1’) from the rest (‘0’). This representation simplifies the evaluation of the model’s attention distribution, facilitating a more straightforward comparison of its responses to various prompts, thus setting a clearer stage for analyzing how the model associates pronouns with their referents. The impact of this thresholding and the utility of binary masks in enhancing map interpretability are visualized in Figures 3 and 4, respectively.

**Step 3: Heatmap Overlap Filtering** Building on the binary masks created from the previous step, we next employ the Intersection over Union (IoU) metric to further dissect the model’s pronoun disambiguation capabilities. The IoU metric, widely



Figure 5: Instances of heatmap overlap generated by Stable Diffusion 2.0 using the WINOVIS dataset: On the left, two entities lead to nearly identical heatmaps, while on the right, two visually distinct entities show significant heatmap overlap.

recognized in computer vision for evaluating object detection accuracy (He et al., 2017; Szeliski, 2022; Takikawa et al., 2019; Zhu et al., 2019), measures the overlap between two areas. It is commonly applied to assess the precision of detected objects against ground truth, by comparing their respective binary masks. The IoU calculation is as follows:

$$\text{IoU} = \frac{\text{Area of Overlap between the binary masks}}{\text{Area of Union of the binary masks}}$$

This yields a value from 0 (no overlap) to 1 (complete overlap), indicating the strength of association between two terms.

For our purposes, a high IoU score between a pronoun and an entity suggests a correct pronoun-to-entity linkage by the model, while high scores between both entities indicate ‘heatmap overlap’—a state where the model fails to distinguish entity associations, leading to potential misattribution of the pronoun. Refer to Figure 5 for examples of this phenomenon.

Heatmap overlap complicates pronoun disambiguation, as it reflects a failure to distinguish between entities in the first place. To identify an optimal overlap threshold for detecting such errors, we manually inspected 50 WINOVIS instances, evaluating heatmap overlays from Stable Diffusion 2.0. A consensus emerged favoring an IoU threshold of 0.4, which yielded full agreement with classifications made by our team, as depicted in Figure 6.

WINOVIS instances with entity pairs with IoU scores exceeding this threshold are therefore considered invalid, warranting exclusion from further

Version	Captioned	Overlapped	Evaluable
1.0	178	24	298
1.5	135	36	329
2.0	160	71	269
XL	2	73	425

Table 2: The number of images generated by Stable Diffusion versions from WINOVIS prompts, categorized by suitability for pronoun disambiguation analysis.

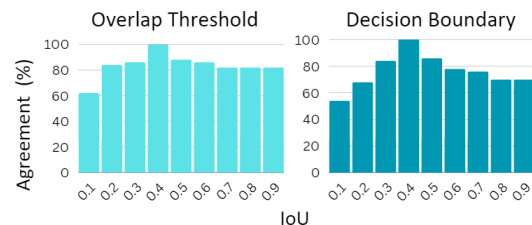


Figure 6: Depicts the level of agreement between the manual decisions and different IoU values for the overlap threshold (left) and decision boundary (right).

analysis to ensure a focus on clear cases of pronoun disambiguation. This filtering process’s impact on the dataset, segmented by model versions is detailed in Table 2.

**Step 4: Making the Final Decision** In this final step, we utilize the IoU metric once more to establish a decision boundary for evaluating the model’s proficiency in pronoun disambiguation. This process involved another comparative analysis conducted by our team of annotators, who manually assessed 50 images generated by SD 2.0 from

SD Version	#Correct	#Incorrect	#Neither	Precision	Recall	F1-Score	Certainty
1.0	24	24	250	50.0	8.8	14.9	16.1
1.5	38	31	260	55.1	12.8	20.7	21.0
2.0	55	42	172	<b>56.7</b>	<b>24.2*</b>	<b>34.1*</b>	<b>36.1*</b>
XL	1	0	424	N/A	N/A	N/A	0.24

Table 3: Comparative performance of Stable Diffusion (SD) models 1.0, 1.5, 2.0, and Stable Diffusion XL (SDXL) (Podell et al., 2023) on the WINOVIS dataset. Metrics are presented as percentages, with \* indicating a statistically significant difference for best model (2.0) from second best (1.5) based on a Z-test for two independent proportions ( $p < 0.01$ ).

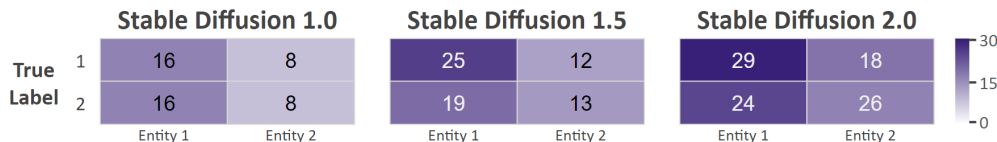


Figure 7: Confusion matrices showing raw count performances of Stable Diffusion models on WINOVIS. Each matrix provides the counts of predictions for Entity 1 and Entity 2 against their true labels.

WINOVIS instances. Each image was reviewed with its corresponding heatmap to determine the presence of a definitive pronoun-to-entity association. Remarkably, the IoU threshold that aligned with manual assessments was identified again at 0.4, mirroring the overlap threshold. This consistency underscores the threshold’s robustness in distinguishing between clear and ambiguous entity associations (Figure 6 illustrates this agreement).

An IoU score exceeding this threshold signals a strong association between the pronoun and a specific entity, as interpreted by the model. This scenario unfolds in two ways:

- If **only one** referent entity’s IoU score with the pronoun surpasses this threshold, it directly informs the model’s prediction, indicating a clear pronoun-to-entity association.
- If **both** referent entities’ IoU scores exceed the threshold, the entity with the higher score is considered the model’s chosen referent.

Predictions are categorized as either *correct* or *incorrect* based on their alignment with the WINOVIS instance’s intended meaning. Cases where neither entity meets the IoU criterion are labeled as *neither*, suggesting the model’s failure to disambiguate the pronoun altogether.

## 5 Experiments & Results

### 5.1 Experimental Setup

#### Dataset Generation and GPT-4 Configuration

For dataset generation, we used GPT-4 (gpt-4-0613;

OpenAI et al., 2023)) with temperature and nucleus sampling (*top-p*) settings optimized to enhance output diversity while adhering to the specific task structure detailed in the prompts. After evaluating temperature values within the range [0,2] with a fixed *top-p* of 1.0, we determined a temperature of 0.8 as the optimal balance for maintaining both dataset integrity and diversity. The WINOVIS images were then generated using Stable Diffusion (SD) versions 1.1, 1.5, 2.0 and Stable Diffusion XL (SDXL), through HuggingFace’s Diffusers library (von Platen et al., 2022), with each model configured to use 50 diffusion steps.

**Diffusion Steps Analysis** An analysis of image generation quality across different diffusion step settings (20, 50, and 100 steps) was performed to identify the optimal configuration for producing WINOVIS images. The evaluation criteria included image quality and the presence of unintended captioning. Fifty steps were found to provide the best balance between image quality and computational efficiency, with no significant quality improvements observed at 100 steps.

**Main Experiments** Using the WINOVIS dataset, we prompted SD versions 1.1, 1.5, 2.0 and XL to generate corresponding images. Throughout this process, heatmaps for both entities and the pronoun were extracted.<sup>2</sup> These prepared heatmaps enabled the application of the IoU metric, as elaborated in Steps 3 and 4 of Section 4.

<sup>2</sup>The heatmap extraction method is based on code available under the MIT License at <https://github.com/castorini/daam>.

**Evaluation Metrics** We measure model performance using the following metrics, adapted for pronoun disambiguation tasks:

- **Certainty:** The frequency with which the model makes a clear pronoun-to-entity association as opposed to its associations being marked as ‘neither’.
- **Precision:** The proportion of the model’s pronoun-to-entity associations that are correct out of all associations made.
- **Recall:** The model’s ability to correctly associate pronouns with entities, where ‘neither’ responses are treated as missed opportunities for correct associations (i.e., false negatives).<sup>3</sup>
- **F1-Score:** The harmonic mean of precision and recall, providing an overall measure of the model’s disambiguation performance.

## 5.2 Results

Table 3 presents the performance of models on the WINOVIS dataset. Key insights include:

**Model Progression and Certainty:** SD 2.0 demonstrates superior precision, recall, and F1-scores, alongside a reduced rate of neither predictions, indicating both progress in pronoun disambiguation and decisiveness. Despite advancements, all models still show a *significant* need for development, with persistent challenges highlighted by the notable proportion of ‘neither’ outcomes and modest precision scores.

The confusion matrices depicted in Figure 7 show the raw count performance of models on the WINOVIS dataset’s pronoun disambiguation problems. Notably, the matrices indicate a gradual decrease in the confusion between entities as the model version increases, with SD 2.0 showing a more distinct separation between the two entities. This suggests an improvement in the models’ ability to discern between entities over iterations.

**Dismal SDXL Performance:** SDXL’s attention maps almost always did not meet the IoU threshold set out for a viable prediction on WINOVIS. Specifically, the heatmaps attributed to the pronoun were often widely dispersed across the image, resulting

<sup>3</sup>An alternative evaluation approach, not penalizing models for “neither” predictions, treats the problem as a multi-class classification. Metrics are computed for each entity class and then averaged. For this analysis, see Appendix Table 4.

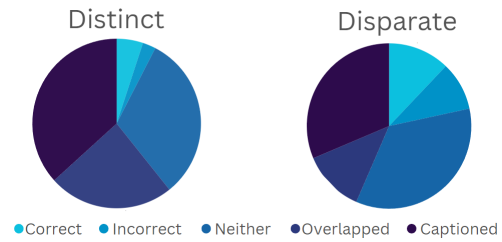


Figure 8: A comparison of the proportion of correct, incorrect, neither, overlapped, and captioned images when SD 2.0 is given distinct versus disparate entities.

in a neither prediction. An example of this can be seen in Appendix Figure 15.

The culprit for this may be SDXL’s consideration of a large context for high-resolution generation. Effectively, this may dilute the attention weights of ambiguous tokens and the extra refiner component would impact the generation of attention heatmaps altogether. At the same time, it was intriguing that this issue occurs exclusively for the token corresponding to the ambiguous pronoun (i.e., in Appendix Figure 15, both the ant and the leaf result in heatmaps that SDXL correctly identifies). This may suggest a tradeoff between image generation quality and pronoun disambiguation – larger, more capable models may come with a pronounced cost to interpretability, resulting compromised performance on benchmarks such as WinoVis.

## 6 Error Analysis

In this section, we further examine the performance of the most effective model iteration, SD 2.0. We compare the results across our dataset categories outlined in Section 3.2, namely disparate and distinct entities. The proportions of correct, incorrect, overlapped, neither, and captioned instances for both categories are visualized in Figure 8.

**Disparate Entities:** In general, SD 2.0 performed the best when working with disparate entities (recall that these were the “easier” problems). Over half of the images were evaluable, with the other 43.4% containing captioning or heatmap overlap. Among the evaluable instances, 31.4% had neither entity chosen, 9.5% were incorrect, and 12.1% were correct. Figure 9 (left) shows SD 2.0’s incorrect pronoun attribution in a WINOVIS scenario involving disparate entities.

**Distinct Entities:** SD 2.0 struggled the most with distinct entities. The majority of instances were not evaluable, with 60.8% of the items containing captioning or heatmap overlap. Among the evaluable





Figure 9: Examples of *incorrect* pronoun associations for *disparate* entities (left) and *distinct* entities (right).

instances, it displayed notable difficulties in making the correct association: in 31.6% of instances neither entity was chosen, while in 2.5% of cases, the incorrect entity was chosen. Only 5.1% of instances resulted in the correct entity being chosen. Figure 9 (right) depicts an example of two distinct entities, a child and a woman. Interestingly, in this image the pronoun ‘she’ is more strongly attributed to the child instead of the woman, even when the child’s gender is not specified.

## 7 Related Work

**Multimodal Reasoning** The recent surge in popularity of generative models has underscored the necessity for explainable creativity (Llano et al., 2020), leading to a significant body of research investigating the determinants of high-quality prompts for image generation (Wang et al., 2023b; Oppenlaender, 2023; Pavlichenko and Ustalov, 2023). Despite these advancements, the evaluation of how vision models actually interpret prompts is largely underexplored. Most studies focus on the models’ semantic understanding of terms (Tang et al., 2023; Parcalabescu et al., 2022; Thrush et al., 2022) or susceptibility to bias (Wang et al., 2023a). These evaluations often involve direct, unambiguous prompts, sidestepping more nuanced challenges. WINOVIS addresses this gap by evaluating the common-sense reasoning of models through the lens of pronoun resolution. This challenge not only expands the scope of assessment for generative models but also sets a new benchmark for understanding their capabilities in interpreting complex linguistic structures.

**WSC-Style Tasks** The Winograd Schema Chal-

lenge (WSC) (Levesque et al., 2011) has catalyzed the development of various datasets aimed at advancing pronominal coreference resolution, each enriching the field by addressing distinct facets of the challenge. Datasets such as Winogrande (Sakaguchi et al., 2020) and KnowRef (Emami et al., 2019) expand on the WSC by tackling its limited size, whereas WinoGender (Rudinger et al., 2018), WinoBias (Zhao et al., 2018), and KnowRef-60k (Emami et al., 2020) study model biases. Further enhancements and crowd-sourcing efforts (Wang et al., 2018; Trichelair et al., 2018; Kocijan et al., 2019; Elazar et al., 2021; Zahraei and Emami, 2024; Sakaguchi et al., 2020) have continually refined the WSC task’s scope and methodology. WINOVIS uniquely adapts the WSC for text-to-image model evaluation, focusing on multimodal common-sense reasoning. It introduces the challenge of visually disambiguating pronouns, filling a crucial gap in multimodal evaluation.

## 8 Conclusion

This paper presented WINOVIS, a new approach to test how well text-to-image models like Stable Diffusion handle pronoun disambiguation. Our work reveals significant gaps in these models’ abilities to interpret ambiguous scenarios accurately. Central to our contribution is a novel evaluation framework designed to isolate common-sense reasoning in pronoun disambiguation from well-studied challenges such as typographic attacks and semantic entanglement. Future research should build on our groundwork to develop models that not only generate visually compelling images but also accurately understand the narratives and relationships within them.

## Limitations

**Entity Separation:** Stable Diffusion models encounter challenges with distinguishing between two semantically similar entities. This can be seen in either heatmap overlap or entanglement, both of which result in a significant proportion of generated images being unsuitable for pronoun disambiguation. Entanglement is particularly pronounced in images generated from prompts featuring semantically similar entities. Since sentences from WINOVIS often employ such entities to introduce ambiguity, resolving entanglement could improve the model’s ability to distinguish individual entities and expand the range of Winograd-like prompts that Stable Diffusion can visualize for our analysis.

**Model Diversity:** Due to its open-source nature, Stable Diffusion facilitated the creation of heatmaps using DAAM, a capability not available in closed-source LDMs. Currently, DAAM is the only framework which enables the interpretation of such models and is specifically designed for Stable Diffusion. Future research should investigate methods to enhance interpretability across a wider range of LDMs and multi-modal diffusion models (and more open-source ones, as they become increasingly available), enabling their assessment in pronoun disambiguation using WINOVIS.

**Bias Analysis:** Our study does not explicitly address potential biases in Stable Diffusion that might influence its decision-making processes. Instances of incorrect pronoun resolution, such as the woman-child example depicted in Figure 9, hint at underlying biases. Future work should rigorously explore these biases and their effects on model performance. Investigating whether Stable Diffusion exhibits systematic preferences in resolving ambiguities could uncover patterns in its reasoning strategies, guiding efforts to mitigate biases and enhance multimodal pronoun disambiguation capabilities.

**Dataset Diversity:** Although efforts were made to maximize dataset diversity during the generation of samples for WINOVIS, opportunities for enhancement remain. Further refinement could entail creating samples that exhibit greater complexity and encompass a broader spectrum of circumstances, entities, and instances of ambiguous pronouns.

**Filtering Limitations:** Although our filtering process aimed to minimize the impact of model weaknesses on our analysis, exceptions exist. In

certain cases, semantic entanglement eluded detection through heatmap overlap measures (see Appendix Figure 14 for an example). Future research should investigate alternative detection methods to better mitigate the influence of such model flaws on our analysis of WINOVIS.

## Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada, the New Frontiers in Research Fund, and the Canada Research Chairs Program.

## References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.
- Nassim Dehouche and Kullathida Dehouche. 2023. What’s in a text-to-image prompt? The potential of stable diffusion in visual arts education. *Heliyon*.
- Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. [Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ali Emami, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. An analysis of dataset overlap on Winograd-style tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5855–5865.
- Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. [The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

- Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30.
- Shun Hattori and Madoka Takahara. 2023. A study on human-computer interaction with text-to/from-image game AIs for diversity education. In *International Conference on Human-Computer Interaction*, pages 471–486. Springer.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems*.
- Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. [A surprisingly robust trick for the Winograd schema challenge](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842, Florence, Italy. Association for Computational Linguistics.
- Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2023. [The defeat of the Winograd Schema Challenge](#). *Artificial Intelligence*, 325:103971.
- Hector Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd Schema Challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.
- Maria Teresa Llano, Mark d’Inverno, Matthew Yee-King, Jon McCormack, Alon Ilisar, Alison Pease, and Simon Colton. 2020. [Explainable computational creativity](#). In *International Conference on Computational Creativity 2020, ICCO’20*, pages 334–341. Association for Computational Creativity (ACC). International Conference on Computational Creativity 2020, ICCO 2020 ; Conference date: 07-09-2020 Through 11-09-2020.
- OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madeleine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rameez Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayarvigiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong,

- Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. [Gpt-4 technical report](#).
- OpenAI. 2023. [Gpt-4 technical report](#). *ArXiv*, abs/2303.08774.
- Jonas Oppenlaender. 2023. [A taxonomy of prompt modifiers for text-to-image generation](#). *Behaviour & Information Technology*, pages 1–14.
- Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. [VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.
- Nikita Pavlichenko and Dmitry Ustalov. 2023. [Best prompts for text-to-image models and how to find them](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 2067–2071, New York, NY, USA. Association for Computing Machinery.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. [Sdxl: Improving latent diffusion models for high-resolution image synthesis](#).
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.
- 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*, pages 10684–10695.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. [U-Net: Convolutional networks for biomedical image segmentation](#).
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. [Photo-realistic text-to-image diffusion models with deep language understanding](#). *Advances in Neural Information Processing Systems*, 35:36479–36494.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. [WinoGrande: An adversarial Winograd schema challenge at scale](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.
- Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. [Deep unsupervised learning using nonequilibrium thermodynamics](#). In *International Conference on Machine Learning*.
- Yang Song and Stefano Ermon. 2019. [Generative modeling by estimating gradients of the data distribution](#). In *Neural Information Processing Systems*.
- Richard Szeliski. 2022. *Computer vision: Algorithms and applications*. Springer Nature.
- Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. 2019. [Gated-SCNN: Gated shape CCNs for semantic segmentation](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5229–5238.
- Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2023. [What the DAAM: Interpreting stable diffusion using cross attention](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659, Toronto, Canada. Association for Computational Linguistics.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. [Winoground: Probing vision and language models for visio-linguistic compositionality](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238.
- Paul Trichelair, Ali Emami, Jackie Chi Kit Cheung, Adam Trischler, Kaheer Suleman, and Fernando Diaz. 2018. [On the evaluation of common-sense reasoning in natural language understanding](#). In *Critiquing and Correcting Trends in Machine Learning NeurIPS 2018 Workshop*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention is all you need](#).
- 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](#). <https://github.com/huggingface/diffusers>.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the*

- 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Jialu Wang, Xinyue Liu, Zonglin Di, Yang Liu, and Xin Wang. 2023a. [T2IAT: Measuring valence and stereotypical biases in text-to-image generation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2560–2574, Toronto, Canada. Association for Computational Linguistics.
- Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023b. [DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–911, Toronto, Canada. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#).
- Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688. Citeseer.
- Terry Winograd. 1972. [Understanding natural language](#). *Cognitive Psychology*, 3(1):1–191.
- Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. 2023. [Uncovering the disentanglement capability in text-to-image diffusion models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910.
- Pardis Sadat Zahraei and Ali Emami. 2024. [WSC+: Enhancing the Winograd Schema Challenge using tree-of-experts](#). *arXiv preprint arXiv:2401.17703*.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. [Gender bias in coreference resolution: Evaluation and debiasing methods](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.
- Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. 2019. [Improving semantic segmentation via video propagation and label relaxation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8865.

## A Appendix

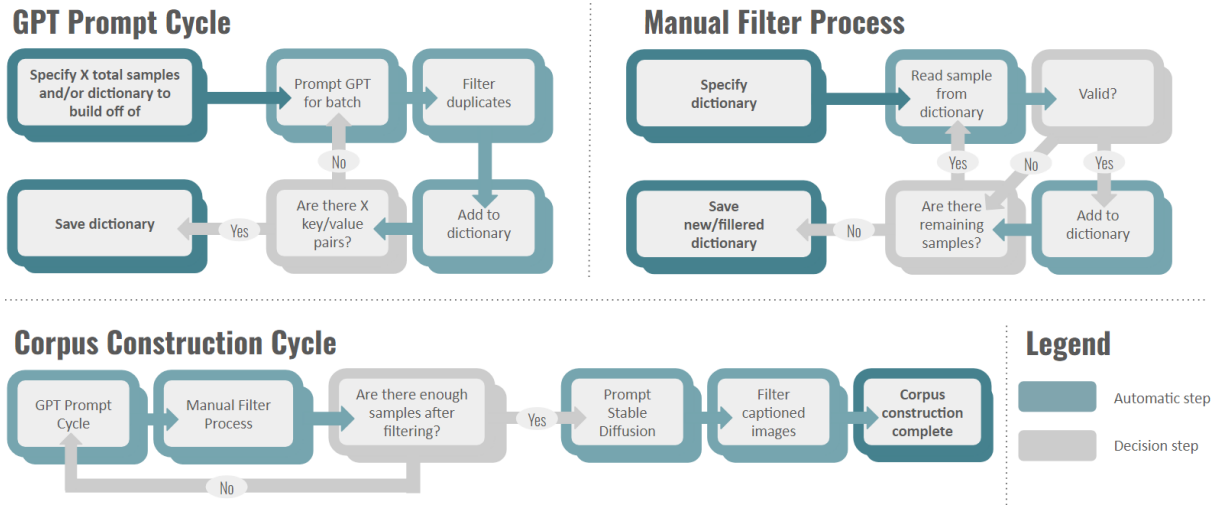


Figure 10: A visual overview of our Corpus Construction Cycle.



Figure 11: An example of image captioning. In this case, the prompt “The customer returned the product because it was unsatisfied” produced an image that includes the word ‘customer’. The attribution heatmap for the term ‘customer’ focuses on this text.

Model	Correct Attribution		Incorrect Attribution		Accuracy	Precision	Recall	F1-Score
	Entity 1	Entity 2	Entity 1	Entity 2				
1.0	16	8	16	8	50.0	50.0	50.0	50.0
1.5	25	13	19	12	55.1	54.4	54.1	54.3
2.0	29	26	24	18	<b>56.7</b>	<b>56.9</b>	<b>56.9</b>	<b>56.9</b>

Table 4: An alternate evaluation of Standard Diffusion models that treats the problem as a multi-class classification task. The reported Precision and Recall scores are computed by taking the average of both entity classes.

Component	Prompt Content
Setup	<p>A Winograd schema sentence is a sentence that contains an ambiguity and requires world knowledge and reasoning for its resolution. For example: The city councilmen refused the demonstrators a permit because they feared violence.</p> <p>Here, “they” presumably refers to the city council; because city councils are typically responsible for maintaining order and avoiding violence in their city. It is more plausible that a city council would fear violence than actively advocate for it. In this example we get the answer based on our world knowledge that tells us city councils generally wish to preserve order, while protest movements sometimes embrace confrontation and violence to achieve political aims. This matches the logical referents in the schema.</p>
Criteria	<p>Winograd schema sentences must abide by five rules:</p> <ol style="list-style-type: none"> <li>1. Be easily disambiguated by the reader;</li> <li>2. Not be solvable by simple techniques such as selectional restrictions;</li> <li>3. The “snippet” must directly refer to the entity specified by the “answer”</li> <li>4. Neither of the “options” should be found in the “snippet”.</li> <li>5. The “pronoun” must be applicable to both “options”. For example, two men could share the pronoun “he” or “him”. Furthermore, a person with an occupation such as an athlete or doctor and a non-human entity cannot share the pronouns “he” or “she” but may share “it”. If a plural pronoun is used such as “they” then both “options” should also be plural. For example, coaches instead of coach and players instead of player.</li> </ol>
Examples	<p>Here is an example of some sentences which match the format of the Winograd schema: (using output with reason examples)</p> <p><b>INSERT WSC SAMPLES</b></p> <p>An example of an invalid pair is:</p> <p>Sentence1:</p> <pre>{ “statement”: “The boy kicked the ball because it was deflated.”, “pronoun”: “it”, “snippet”: “it was deflated”, “options”: [“the boy”, “the ball”], “answer”: 1, “reason”: “If ‘deflated’ is used, it implies the ball was deflated.” }</pre>
CoT	<p>Without skipping any, come up with <b>BATCH_SIZE</b> new valid sentences starting at sentence one. Think step by step for each new sentence by following these steps:</p> <ol style="list-style-type: none"> <li>1. Come up with two entities or objects which share a pronoun.</li> <li>2. Think of a pronoun that seems just as semantically compatible with the two antecedent options, but can be disambiguated using common sense reasoning and not at all with distributional cues between the antecedents and the rest of the sentence.</li> <li>3. Come up with a completely new sentence that follows the principles of the example sentences and follows the rules listed above.</li> </ol> <p>Repeat this process for all the sentences you generate. The sentences should be original and diverse in the topics that they cover.</p>

Table 5: The prompt used in the Corpus Construction Cycle broken down into distinct sections.

<p>The athlete left the game because it was [risky/exhausting].</p> <p>a:</p> <pre>{ "statement": "The athlete left the game because it was risky.",   "pronoun": "it",   "snippet": "it was risky",   "options": ["athlete", "game"],   "answer": 1,   "reason": "If 'risky' is used, it implies the game was risky, causing the athlete to leave." }</pre> <p>b:</p> <pre>{ "statement": "The athlete left the game because it was exhausting.",   "pronoun": "it",   "snippet": "it was exhausting",   "options": ["athlete", "game"],   "answer": 0,   "reason": "If 'exhausting' is used, it implies the athlete was exhausted, causing him to leave the game." }</pre> <p>Explanation: The "snippet" refers to the game's impact on the athlete when it should refer to the "athlete" itself. To correct this sample, the term used should be exhausted instead of exhausting.</p>
<p>The boy kicked the ball because it was [deflated/inflated].</p> <p>a:</p> <pre>{ "statement": "The boy kicked the ball because it was deflated.",   "pronoun": "it",   "snippet": "it was deflated",   "options": ["the boy", "the ball"],   "answer": 1,   "reason": "If 'deflated' is used, it implies the ball was deflated." }</pre> <p>b:</p> <pre>{ "statement": "The boy kicked the ball because it was inflated.",   "pronoun": "it",   "snippet": "it was inflated",   "options": ["the boy", "the ball"],   "answer": 1,   "reason": "If 'inflated' is used, it implies the ball was inflated, prompting the boy to kick it." }</pre> <p>Explanation: In a Pair, a and b must not have the same "answer". If Pair2.a's "answer" is 0, Pair2.b's "answer" should be 1 and vice-versa.</p>

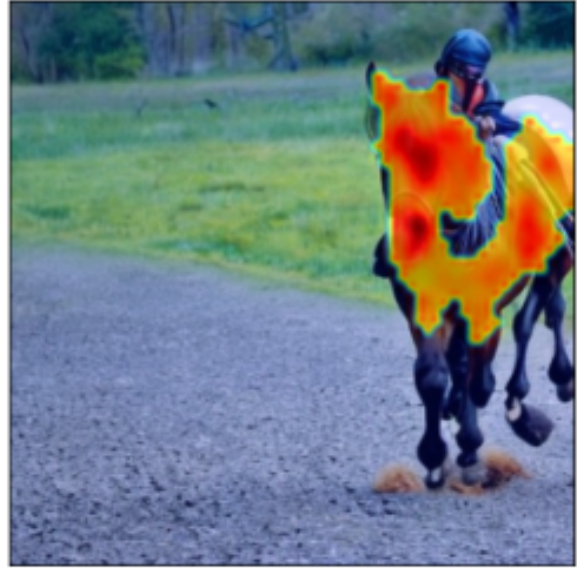
Table 6: Examples of invalid instances that were included in the prompt used in the Corpus Construction Cycle.

Context Type	% of WSV	Example
Visually Tangible	38.6	The <b>plumber</b> had to replace the <b>pipe</b> because <b>it</b> was <i>rusty</i> .
Emotional	15.0	The <b>dog</b> chased the <b>car</b> because <b>it</b> was <i>excited</i> .
Characteristic	29.2	The <b>king</b> did not trust the <b>advisor</b> because <b>he</b> was <i>deceitful</i> .
Visually Intangible	17.2	The <b>cat</b> is afraid of the <b>vacuum cleaner</b> because it is <i>loud</i> .

Table 7: Examples taken from the WINOVIS (WSV) dataset exhibiting the four different context types.



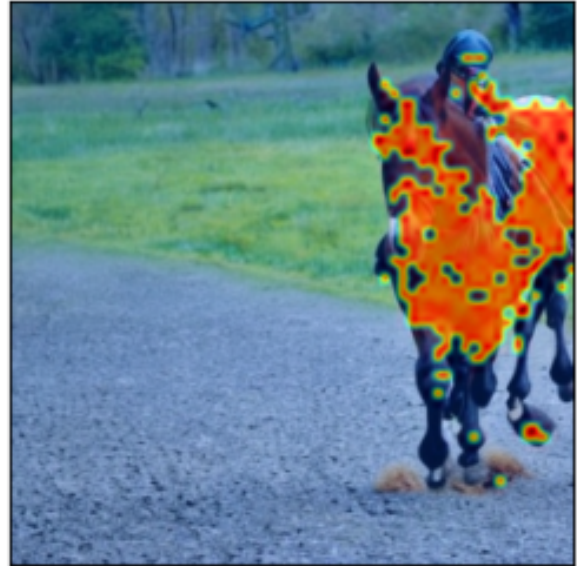
The **horse** outran the **dog** because **it** was faster



horse



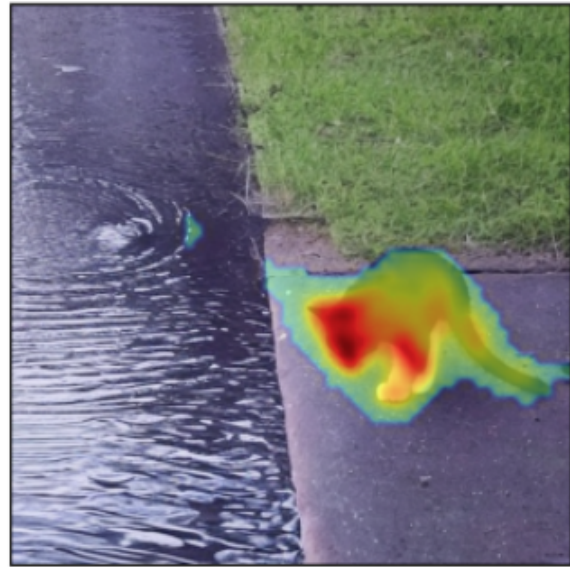
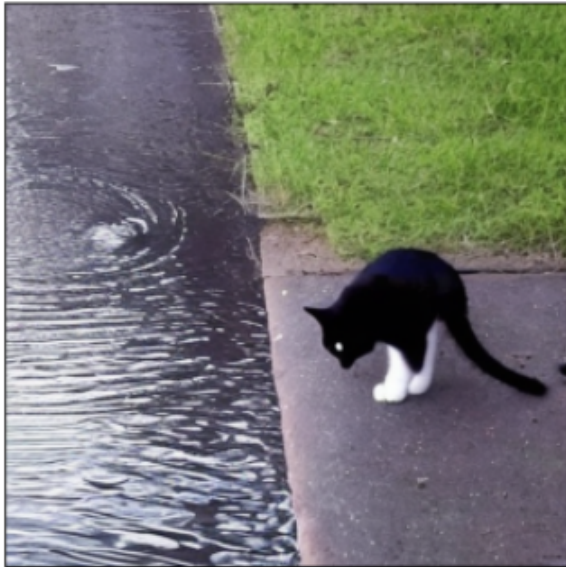
dog



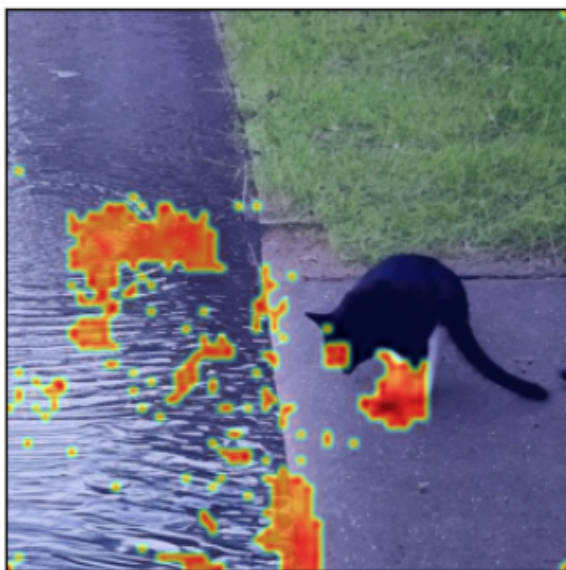
it

Figure 12: An example of a generated image containing only one of the entities from the prompt. While the horse is visible, the dog is not.

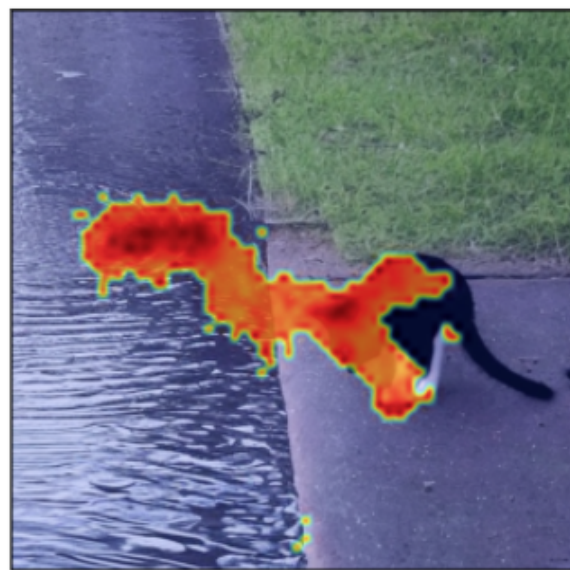
The **cat** avoided the **water** because **it** was scared



cat



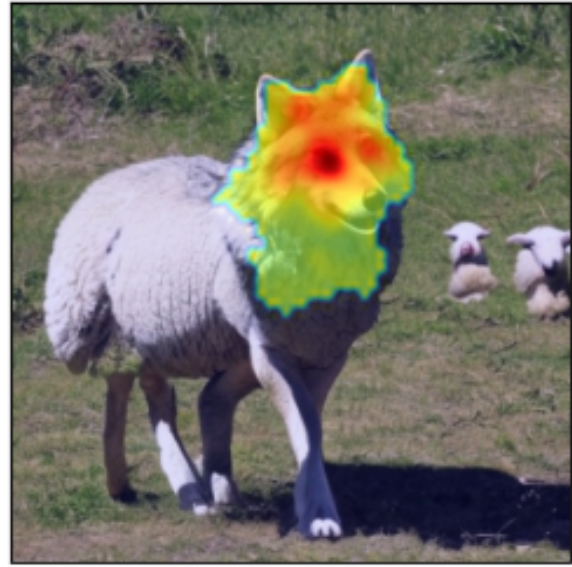
water



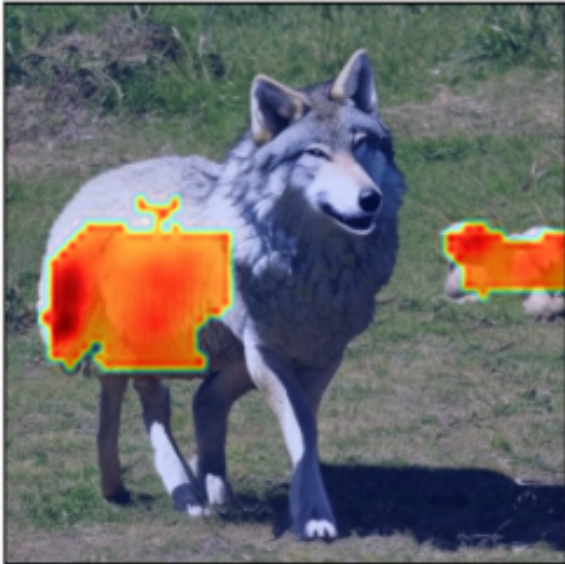
it

Figure 13: An example of the case where the DAAM heatmap for the pronoun does not clearly indicate a decision made by the model. Rather than overlapping with either the 'cat' or the 'water', the heatmap for 'it' appears to overlap slightly with both while also encompassing some space not seen in either of the entities' heatmaps.

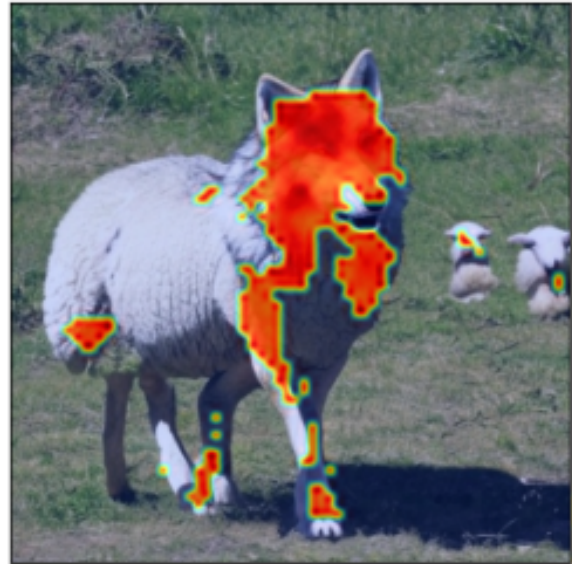
The **wolf** attacked the **sheep** because **it** was hungry



wolf



sheep



it

Figure 14: An example of an image that was not automatically filtered via measurement of heatmap overlap. While the two entities are semantically entangled their heatmaps are distinct (non-overlapped).

The **ant** could not carry the **leaf** because **it** was weak

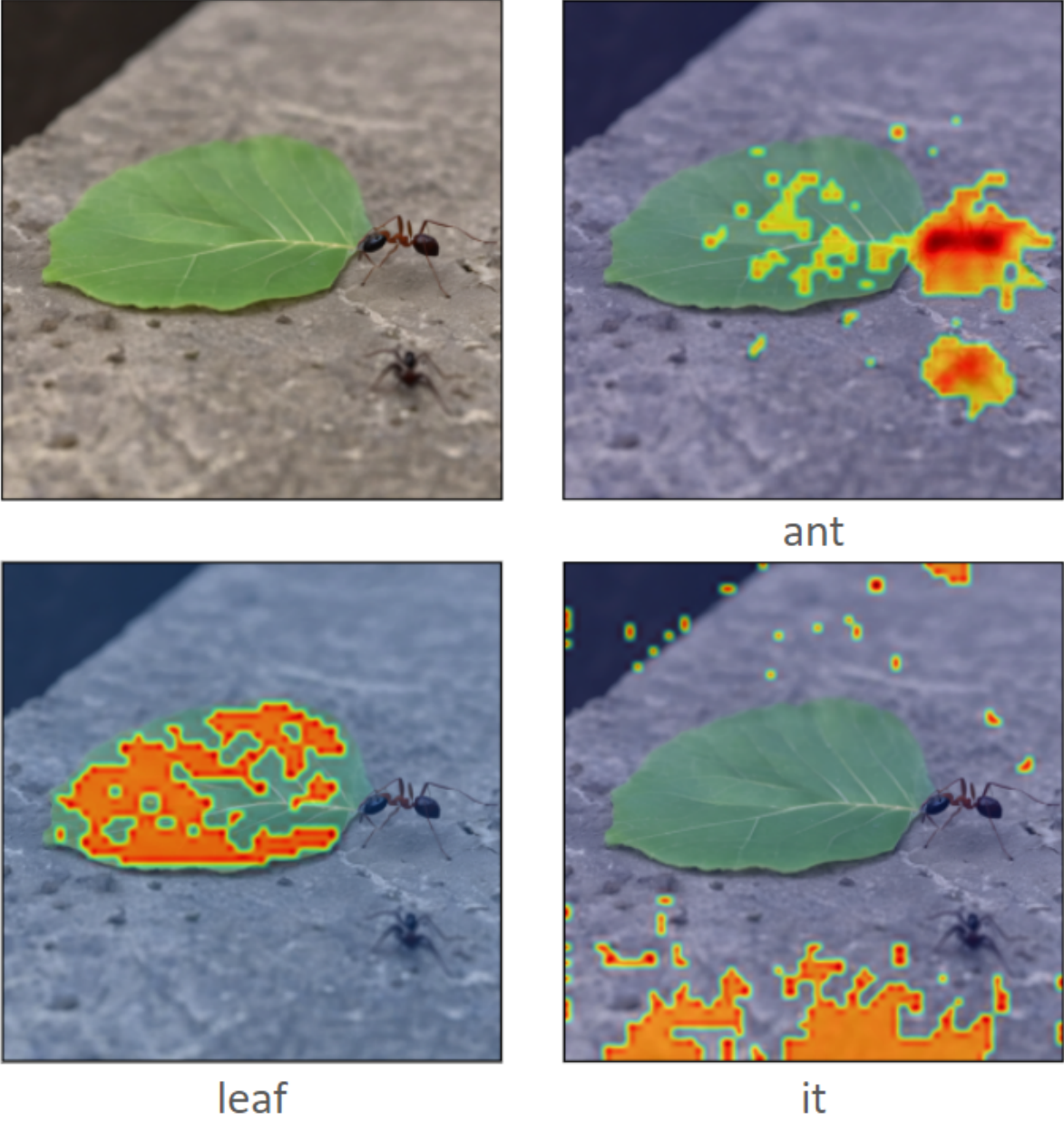


Figure 15: An example of an image generated by SDXL. Here, both entity heatmaps overlap correctly with their respective visual representations. However, the heatmap for the ambiguous pronoun is distributed across the image showing a lack of certainty in the model’s decision.