FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACHBEREICH WIRTSCHAFTS-
WISSENSCHAFTEN

A Linguistic Analysis of Attention Mechanisms in Stable Diffusion Models

Master Thesis

at the Chair of Business Information Systems
(Friedrich-Alexander University of Erlangen-Nuremberg)

Principal Supervisor:	Prof. Dr. Michael Amberg
Associate Supervisor:	Dr. Stefan Arnold

Presented by:	Syed Ikhtiar Ahmed Röthenburger Straße 145 90439 Nürnberg +49 163 9262820 syed.ikhtiar.ahmed@fau.de 23039053

Submission:	17th December 2025

Contents

Figures III
Tables V
1 Introduction 1
1.1 Background and Motivation 1
1.2 Problem Statement 1
1.3 Research Objectives and Questions 2
1.4 Scope and Limitations 2
1.5 Method Overview 3
1.6 Contributions 3
2 Related Work 4
2.1 Diffusion and Stable Diffusion Models 4
2.2 Attention Mechanisms in Transformer-Based Diffusion Models 6
2.3 Interpretability and Visualization of Attention 7
2.4 Linguistic Perspectives on Diffusion Models 8
3 Background Study 9
3.1 Latent Diffusion Model 9
3.2 Attention Mechanisms 13
3.3 Diffusion Attentive Attribution Maps 14
3.4 Stable Diffusion 3 17
4 Our Approach 23
5 Experiments and Analyses 25
5.1 POS Analysis 25
5.1.1 Experimental Setup 25
5.1.2 Observations 25
5.2 DEP Analysis 27
5.2.1 Experimental Setup 27
5.2.2 Observations 28
5.3 PPA Analysis 31
5.3.1 Experimental Setup 31
5.3.2 Observations 32
6 Discussion 36
7 Conclusion 37
Appendix 38
References 51

Figures

1 Illustration of forward and reverse diffusion process. *Source: Ho et al. (2020).* 9
2 Architecture of Latent Diffusion Models *Source: Rombach et al. (2022).* 10
3 U-net architecture. *Source: Ronneberger et al. (2015).* 11
4 Parts of Internal structure and components of the U-Net used in the Stable Diffusion 2.0 base model (Stability AI 2022), available through Hugging Face's Diffusers library (von Platen et al. 2022). 12
5 Illustration of attention computation. *Source: Vaswani et al. (2017).* 13
6 Visualization of the DAAM computation. *Source: Tang et al. (2023).* 15
7 Illustrations of mIoD, mIoH, and various types of overlap. *Source: Tang et al. (2023).* 17
8 Architecture of Diffusion Transformer (DiT). *Source: Peebles and Xie (2023).* 20
9 Architecture of Stable Diffusion 3.0 with MM-DiT block. *Source: Esser et al. (2024).* 21
10 Architecture of Stable Diffusion 3.5 with MMDiT and MMDiT-X block. *Source: Stability AI (2024).* 22
11 Visualization of the attention score matrix including different regions. **A:** Query of the image tokens attend to keys of the image; **B:** Query of the image tokens attend to keys of the text; **C:** Query of the text tokens attend to keys of the image; **D:** Query of the text tokens attend to keys of the text. 23
12 Image generated with the prompt *'A photograph of an astronaut riding a horse.'* and the corresponding attribution maps for the words *'astronaut'*, *'riding'*, and *'horse'*. 24
13 Analysis of pixel coverage and attention values across POS tags. 26
14 The Mean Pixel Count per Layer for all POS Tags. 26
15 The Mean Attention Values per Layer for all POS Tags. 27
16 The mIoU, mIoD, and mIoH scores for all tags across layers. 29
17 The mIoU, mIoD, and mIoH scores for the *det* tag across layers. 29
18 The mIoU, mIoD, and mIoH scores for the *nsubj* tag across layers. 30
19 The difference between mIoD and mean mIoH for all tags across layers. 30
20 Number of correct attachment predictions (out of 80 samples) using quantile thresholding at different threshold values. 31
21 Confusion matrix for true and predicted attachments. 32
22 Total and correct predictions per preposition 33
23 True and predicted number of noun and verb attachments per preposition 33
24 Distribution of IoU scores for all, correct, and incorrect attachments. 34
25 Cohen's d between IoU scores of n_1–n_2 and v–n_2 per Preposition. 35
26 Average IoU per Layer. 35

27 Examples of attribution maps generated for all 10 POS tags, using captions randomly selected from the COCO dataset. Each caption includes at least one instance of every POS tag. 39
28 Image generated with the prompt *'A monkey with a hat walking.'* and the corresponding attribution maps across individual layers for the word *'monkey'*. 40
29 Image generated with the prompt *'Someone pouring tea into a glass cup.'* and the corresponding attribution maps across individual layers for the word *'cup'*. 41
30 Example of attribution maps of head and dependent words for the top 12 tags, along with their overlap statistics. 45
31 Good, bad, and ambiguous examples for each preposition, with corresponding overlap scores and prepositional attachment labels. 51

Tables

1 Overlap statistics for the 12 most frequent head—dependent relationships, where the **Δ** denotes the difference between **mIoD** and **mIoH**. 28

Abstract

Although text-to-image diffusion models are evolving rapidly, research on their internal mechanisms from a syntactic perspective remains limited. In this study, we perform text-to-image attribution analysis on recent Stable Diffusion models that are based on rectified flow transformers. We adapt the Diffusion Attentive Attribution Maps (DAAM) method, which is designed for earlier Stable Diffusion models based on Latent Diffusion Models (LDMs); however, we aggregate attention scores from a specific region of the attention score matrix across all heads, layers, and diffusion steps to generate attribution maps. We analyze the most frequent parts-of-speech tags and find that some tags exhibit broad attention, while others show localized patterns, and identify key layers that contribute more prominently. We further examine head–dependent relationships in generated images and find that dependents dominate heads in most cases, with specific layers contributing most to this dominance. Finally, we investigate prepositional phrase ambiguities and observe a slight noun-attachment bias. Our code is available at https://github.com/patriotic/StableDAAM.

1 Introduction

1.1 Background and Motivation

Text-to-image diffusion models are a class of generative models in which images are synthesized by iteratively removing noise conditioned on textual prompts (Nichol et al. 2022; Ramesh et al. 2022; Rombach et al. 2022). Several models have been developed for high-resolution text-to-image synthesis, such as Imagen (Saharia et al. 2022), DALL-E 2 (Ramesh et al. 2022), and DALL-E 3 (Betker et al. 2023). However, their weights are not publicly available, whereas Stability AI has open-sourced the Stable Diffusion models. Rombach et al. (2022) introduced U-Net based Latent Diffusion Models (LDMs), in which the diffusion process is performed in a latent space rather than a pixel space, reducing computational complexity without degrading performance. All text-to-image models released by Stability AI prior to version 3.0 are based on LDMs. Stable Diffusion 3.0 and 3.5 are the most advanced text-to-image models to date, where an enhanced architecture is introduced that uses rectified flow (Liu et al. 2023) to improve noise sampling techniques and replace the traditional U-Net backbone with a transformer-based architecture (Esser et al. 2024; Stability AI 2024).

The central component in these models that enables the alignment between language and vision is the attention mechanism. Investigating the attention layers provides a way to uncover the internal decision-making process of the model, helping researchers diagnose errors, identify biases, and develop more controllable and reliable text-to-image systems.

In this thesis, we examine the attention mechanisms from a linguistic perspective, with a particular focus on how the model processes syntactic features. This study not only improves our understanding of the interplay between language and vision in Stable Diffusion models 3 and 3.5 but also highlights limitations and areas for improvement.

1.2 Problem Statement

Although recent work—most notably the DAAM by Tang et al. (2023)—has pioneered the interpretability of text-to-image diffusion models from a visuolinguistic perspective, its analysis was limited to Stable Diffusion models based on Latent Diffusion Models (LDMs) (Rombach et al. 2022), where the diffusion process is handled by a U-Net-based architecture. Tang et al. (2023) proposed a novel attribution method to measure how words in a prompt influence different regions of the generated image and provided new insights into how syntactic relationships are mapped in these models.

However, these findings do not extend to newer, state-of-the-art diffusion models like Stable Diffusion 3.0 or 3.5, which use fundamentally different architectures—incorporating rectified

flow and transformer-based backbones (MMDiT or MMDiT-X) (Esser et al. 2024). To date, there has been no systematic investigation of how these advanced models handle syntactic attribution, syntactic relationships, or phrase attachment from a linguistic perspective. This thesis addresses this gap by analyzing attention mechanisms in Stable Diffusion 3.5 (Stability AI 2024), inspired by DAAM but targeting a new generation of diffusion models. This is essential for both machine learning and linguistics to advance the interpretability and reliability of multimodal AI systems.

1.3 Research Objectives and Questions

The objective of this research is to investigate the attention mechanisms of Stable Diffusion 3.5 and to perform a systematic syntactic analysis to interpret the attention scores in this state-of-the-art diffusion model. By examining how the model processes syntactic information, this thesis aims to provide new insights into flow- and transformer-based multimodal generative models (Esser et al. 2024), with relevance for both the machine learning and linguistic communities.

The main research question that guides this thesis is how an input word influences parts of a generated image in Stable Diffusion 3.5. This investigation is inspired by DAAM (Tang et al. 2023), but extends the analysis for more advanced diffusion models. The research focuses on three specific aspects of syntactic analysis: first, how different parts of speech (POS) in the prompt influence the generated image; second, how dependency relationships (DEP) between words are reflected in the visual output; and third, how prepositional phrase attachment (PPA) is resolved in the generated images.

1.4 Scope and Limitations

Certain areas are intentionally left out of scope due to time constraints. This thesis does not cover the semantic analysis of the prompts and the generated images. In particular, it does not address visuosemantic analyzes such as cohyponym or adjectival entanglement experiments presented in the DAAM paper by Tang et al. (2023).

The analysis is conducted exclusively with Stable Diffusion 3.5 (Stability AI 2024), which is selected because it represents the most recent and advanced open-source text-to-image diffusion model available at the time of this research. Further details on the model's architecture and the process of probing it are discussed in later chapters.

1.5 Method Overview

This study employs a custom experimental setup inspired by the implementation of DAAM (Tang et al. 2023) to address the research questions described in Section 1.3. Attention scores are aggregated from the relevant region in all heads, layers, and inference steps of the Stable Diffusion 3.5 model (Stability AI 2024). These scores are used to generate heat maps that visualize the influence of specific words or syntactic structures on different regions of the generated images. The heat maps are further analyzed and evaluated using various quantitative metrics.

The experimental procedure involves extracting syntactic tags from the captions used as prompts. For each caption, an image is synthesized, all attention scores are systematically collected, and heat maps of the corresponding words or tags are generated using the developed setup (see Section 4). This methodology is applied to the three main analyses described in Section 1.3. More details on the experiments and observations are provided in Section 5.

1.6 Contributions

The primary contributions of this thesis are the adaptation of interpretability techniques and systematic syntactic analysis of attention mechanisms in the latest Stable Diffusion 3.0 or 3.5 architecture.

The key contributions of our work are:

(1) We follow the attribution method proposed by Tang et al. (2023); however, we aggregate the attention values from the relevant region of the attention score matrix.

(2) We show that several POS tags attend too broadly, while the attention values for certain POS tags are concentrated in a small region. We also identify the layer that contributes most significantly to the guiding process.

(3) Similar to DAAM (Tang et al. 2023), we investigate the dependency relationships between the head and the dependent words captured in the generated image. We compare our results with those reported in DAAM, which is based on earlier stable diffusion models, and show that several dependency relationships are represented differently in Stable Diffusion 3, with dependents dominating the heads in most cases. We also identify the layers that contribute the most significantly to these dominance relationships.

(4) We investigate how the model resolves syntactic ambiguities arising from prepositional phrases and show that it exhibits a small tendency toward noun attachment rather than verb attachment.

2 Related Work

2.1 Diffusion and Stable Diffusion Models

The theoretical basis of diffusion models is derived from nonequilibrium thermodynamics, particularly the study of diffusion processes in physics. These concepts were initially put forth by Sohl-Dickstein et al. (2015) for generative modeling, demonstrating how new samples could be generated by reversing a diffusion process that eventually destroys data structures. This physically inspired approach provided a principled alternative to existing generative models, with two primary advantages: flexibility in selecting model architecture and tractable training objectives based on variational bounds. In contrast to other methods that required careful architectural constraints or adversarial training, their framework enabled any parametric model to learn the reverse diffusion process while preserving stable and well-defined training objectives.

With Denoising Diffusion Probabilistic Models (DDPMs), Ho et al. (2020) popularized this framework and showed that high-quality images could be produced by parameterizing the reverse diffusion process. The generative process is a discrete-time Markov chain in the DDPM formulation, which uses variational inference to learn how to reverse a fixed forward noising process.

A major theoretical advancement came from Song et al. (2021b), who presented a stochastic differential equation (SDE) perspective that unifies score-based generative modeling and diffusion probabilistic modeling. This framework revealed that the forward diffusion process could be described by an SDE, and the reverse generative process corresponds to solving a reverse-time SDE. Most importantly, they showed that there is an ordinary differential equation (ODE) that can sample from the same distributions as any diffusion SDE. The capabilities of diffusion models are significantly enhanced by this probability flow ODE, making deterministic generation, exact likelihood computation, and improved sampling efficiency possible.

In addition to that, more flexible sampling procedures became feasible through the SDE/ODE framework. Song et al. (2021a) introduced denoising diffusion implicit models (DDIMs), showing that faster sampling than DDPM can be achieved using the probability flow ODE. Dhariwal and Nichol (2021) investigated contemporary diffusion models and demonstrated that diffusion models surpass GANs in image synthesis quality. Further studies investigated different sampling techniques to speed up the generation process without affecting quality (Lu et al. 2022; Zhang and Chen 2023).

Sequential evaluation of the image generation process in diffusion models was computationally expensive because it operated in a high-dimensional space. This challenge was addressed by Rombach et al. (2022) using an autoencoder that transforms data from the pixel space to the

low-dimensional latent space and performs the diffusion process there. This departure made both training and sampling significantly more efficient while preserving image quality. By operating in compressed latent space, Latent Diffusion Models (LDMs) achieved high-resolution image synthesis at lower computational cost. Additionally, they incorporated a conditioning mechanism based on cross-attention into the U-Net backbone that enables multi-modal training.

There has been extensive research aimed at minimizing the cost of transporting data between two distributions for generative modeling and this can be efficiently modeled using ordinary differential equations (ODEs). Normalizing Flows (NFs) use a sequence of discrete invertible and differentiable transformations, whereas Continuous Normalizing Flows (CNFs) model this transformation as a continuous-time process defined by an ODE Chen et al. (2018). However, training CNFs is computationally expensive, memory intensive, and can be unstable because it requires repeatedly solving ODEs. These issues were addressed by Lipman et al. (2023), which significantly improves training scalability and stability, using flow matching to directly regress the velocity vector fields instead of solving ODEs during training. Liu et al. (2023) proposed a rectification of the flow, learning the velocity fields during training of CNFs that follow straight paths between data and noise for faster sampling.

The evolution from Stable Diffusion 1.x / 2.x to 3.0 involved several fundamental changes (Esser et al. 2024). First, it achieved superior performance for text-to-image synthesis by replacing the traditional diffusion process with rectified flow (Liu et al. 2023), which learns the direct transport between noise and data distributions. Second, it replaced the cross-attention-based U-Net backbone with a joint-attention-based transformer architecture (MMDiT). This separates text and image modalities while enabling a bidirectional flow of information between text and image tokens within the network for text-to-image generation. Subsequently, Stable Diffusion 3.5 was released by Stability AI (2024) with architectural refinements (MMDiT-X), where a self-attention module was added to each MMDiT block in the first 13 layers, enhancing multi-resolution generation, increasing overall image coherence, and making training more stable.

Most of the recent research has focused on the generation quality and efficiency metrics of these diffusion models. Karras et al. (2022) organized the design choices of the diffusion models and presented a unified design space. Some notable early work replaced the CNN-based backbone in diffusion models with transformers and analyzed their scalability and performance (Dosovitskiy et al. 2021; Peebles and Xie 2023; Bao et al. 2023). However, the ability to process linguistic or textual information was not investigated in these studies.

2.2 Attention Mechanisms in Transformer-Based Diffusion Models

The attention mechanism in neural network architectures was first introduced in neural machine translation by Bahdanau et al. (2015) to address the limitation of encoder-decoder models with fixed-length vectors, which struggled to retain long-range dependencies in sequence tasks. They proposed an architecture to jointly learn alignment and translation, allowing the model to focus on relevant parts of the input while predicting each target token, thereby preserving context in longer sentences. Subsequently, the idea of self-attention, particularly scaled dot-product attention, was introduced with the novel architecture Transformer and extended with multi-head attention, enabling parallelization and reducing training time, which led to better performance in sequence modeling or translation tasks (Vaswani et al. 2017).

Research on diffusion models has increasingly incorporated attention mechanisms into their architectures. Early works used self-attention blocks at fixed resolutions within pixel-space U-Nets (Ho et al. 2020), while Dhariwal and Nichol (2021) improved on this by increasing the number of attention heads and using multiple resolutions to capture global context more efficiently. For conditional image generation, cross-attention has enabled diffusion models to integrate text conditioning effectively in both pixel-space (Saharia et al. 2022) and latent-space U-Nets (Rombach et al. 2022).

The architecture of computer vision models was largely based on CNNs; however, the Vision Transformer (ViT) (Dosovitskiy et al. 2021) introduced a transformer-based architecture that splits images into patches and processes them with self-attention, achieving strong results in image classification tasks when trained on large datasets. Subsequently, the CNN-based U-Net backbone in diffusion models was replaced by a transformer that operates on latent patches instead of image patches (Peebles and Xie 2023; Bao et al. 2023). Furthermore, Peebles and Xie (2023) compared four different conditioning strategies in the DiT design space and found that AdaLN-Zero outperforms the other conditioning mechanisms, including cross-attention.

Instead of injecting conditional information at each step using mechanisms such as cross-attention (Rombach et al. 2022; Peebles and Xie 2023), the Multimodal Diffusion Transformer (MMDiT) maintains separate streams for text and latent image (Esser et al. 2024). Both modalities flow simultaneously through the architecture, and within each MMDiT block, the sequences are joined for the attention operation, enabling bidirectional exchange of information between text and image tokens. However, how these design choices affect the linguistic properties of the model during image generation remains largely unexplored.

2.3 Interpretability and Visualization of Attention

Interpretation and visualization of attention help to understand the behavior of a model. In particular, they indicate which parts of the input the model focuses on when making decisions. Early works explored the interpretation of attention weights, including soft alignment patterns in translation tasks (Bahdanau et al. 2015), the capture of long-range dependencies, sentence structure, and anaphora resolution in self-attention Vaswani et al. (2017). However, some researchers questioned the reliability of interpreting attention weights, stating that they are weakly correlated with feature importance for model predictions (Jain and Wallace 2019; Serrano and Smith 2019), while others argued that attention can still provide useful information (Wiegreffe and Pinter 2019).

Analyzing the attention weights of transformer-based models indicates the strength of a token's contribution to the model's decision-making process. Clark et al. (2019) analyzed attention and showed that specific heads specialize in certain syntactic roles. Further tools were developed to visualize attention at different levels, including individual neurons to understand the query–key contribution, head-level views to observe patterns, and model-level views to obtain an overall perspective across layers and heads (Vig 2019), with later work introducing a global view based on joint query–key embeddings (Yeh et al. 2024). Other studies expanded attention analysis: Abnar and Zuidema (2020) captured global information flow by computing how attention propagates across layers using rollout and attention flow, and Hao et al. (2021) proposed self-attention attribution, which combines attention weights with gradients to interpret interactions between tokens.

In transformer-based vision models, image patches are used instead of text tokens. The average attention distance captures how each patch attends to others within a head, indicating whether the head focuses on local or global context, while global attention maps are created using attention rollout and visualized as heatmaps over the image to show which regions most influence the model's prediction (Dosovitskiy et al. 2021). However, Chefer et al. (2021b) proposed combining attention with gradient-based propagation of relevance scores that incorporate information from both attention layers and skip connections.

Chefer et al. (2021a) introduced a generalized method to explain attention in multimodal models by using attention layers to generate relevance maps that capture interactions between input modalities. In latent diffusion models (LDMs), cross-attention–based word-to-pixel attribution maps explain how each text token influences specific image regions during generation. These maps are created by aggregating normalized attention scores across all layers and timesteps (Tang et al. 2023). Recent studies investigated methods to precisely localize textual concepts in images with saliency maps that outperformed visualizations based on cross-attention (Helbling et al. 2025).

2.4 Linguistic Perspectives on Diffusion Models

Investigating text-to-image multimodal diffusion models to examine how language is processed during image synthesis can lead to more faithful image generation. From a syntactic perspective, the analysis focuses on how the structure of words, phrases, and sentences is conveyed in the generated image, while the semantic lens examines how meanings are reflected.

With the goal of understanding how a particular word is captured in the image space, Tang et al. (2023) investigated the attention-based attribution quality of LDM in six parts of speech, while Shahariar et al. (2024) explored the impact of adversarial attacks using those POS tags by perturbing the prompt. Another analysis was carried out to investigate how dependency relationships between words are captured in images generated by LDM (Tang et al. 2023). Some other works examined how verb information is encoded in the internal representations of multimodal transformers (Hendricks and Nematzadeh 2021; Beňová et al. 2025). While exploring the modifiers, Tang et al. (2023) found that feature entanglement exists in the generated image with LDM and later on, Rassin et al. (2023) proposed synGen to address syntactic misalignment between entities and its modifiers. However, these studies are still unexplored for stable diffusion 3.0/3.5.

Linguistic ambiguities are an inherent phenomenon in natural language due to multiple interpretations of the same expression (Mehrabi et al. 2023). Nakashole and Mitchell (2015) provided a method to resolve syntactic ambiguities arising from prepositional phrases by utilizing semantic knowledge. Elsharif et al. (2025) carried out an exploratory analysis of linguistic ambiguity in LDM using prompts, covering lexical, syntactic and figurative ambiguities. Wang et al. (2025) introduced a benchmark for evaluating cross-modal disambiguation. Other researchers proposed frameworks for improving the resolution of ambiguities to generate images that better reflect the intended meaning (Mehrabi et al. 2023; Wang et al. 2025). However, Stable Diffusion 3.0/3.5 has not yet been investigated to understand linguistic ambiguities.

The researchers analyzed transformer heads (Clark et al. 2019) and layers (Tenney et al. 2019) to understand how linguistic information propagates through the network and where different types of information are captured. In text-to-image diffusion models, Chefer et al. (2024) decomposed the generated image to study the internal representations of textual concepts; Toker et al. (2024) introduced the Diffusion Lens to analyze the intermediate layerwise representation of concepts in T5 (Raffel et al. 2020) and CLIP (Radford et al. 2021) text encoders used in diffusion models; and Park et al. (2024) investigated how concepts are prioritized at each denoising step. Furthermore, Kim et al. (2025) studied how visual semantic information is captured across different layers and steps in LDM using k-sparse autoencoders (k-SAE). Nevertheless, to the best of our knowledge, no layerwise analysis of image synthesis models based on Rectified Flow Transformers has been conducted.

3 Background Study

3.1 Latent Diffusion Model

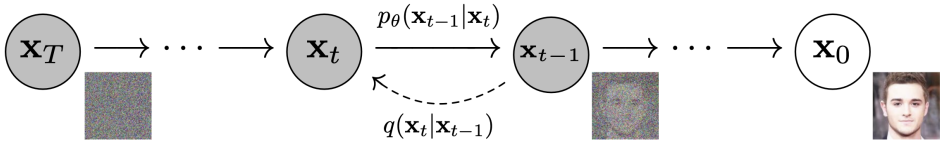


Figure 1 Illustration of forward and reverse diffusion process.
Source: Ho et al. (2020).

Diffusion probabilistic models gradually add Gaussian noise to an image over many steps using a fixed Markov chain. The model learns a parameterized reverse Markov chain that predicts the noise added at each step. During inference, the model starts with pure noise and removes the noise step by step to generate an image (Ho et al. 2020).

In diffusion models, Gaussian noise is gradually added to an image x over T timesteps during the forward process, where the amount of noise at each step is controlled by α_t (Ho et al. 2020), as

$$x_t = \sqrt{\bar{\alpha}_t}\, x + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1), \quad t = 1, \ldots, T \tag{3.1}$$

A neural network is trained to predict the added noise at a random timestep t using the objective (Ho et al. 2020; Rombach et al. 2022):

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2] \tag{3.2}$$

After training, the model can generate images by starting from pure noise and iteratively removing noise using the network (Ho et al. 2020).

Latent diffusion models perform the diffusion process in a compressed latent space instead of a pixel space. Training is performed in two stages, where the image is first compressed to a low-dimensional latent space by training an autoencoder that preserves perceptually important features while removing imperceptible details, and then the diffusion model is trained in that latent space. This reduces complexity and significantly minimizes computational costs. Additionally, a conditioning mechanism based on cross-attention is incorporated into the architecture to control the generation process (Rombach et al. 2022).

The image is compressed to a perceptually equivalent representation by a trained autoencoder consisting of an encoder and a decoder. The encoder $\mathcal{E}$ maps the image $x \in \mathbb{R}^{H\times W\times 3}$ into a latent representation $z \in \mathbb{R}^{h\times w\times c}$ as $z = \mathcal{E}(x)$ using a downsampling factor of $f = 2^m$, $m \in \mathbb{N}$. The decoder $\mathcal{D}$ reconstructs the image from the latent z as $\tilde{x} = \mathcal{D}(\mathcal{E}(x))$ (Rombach et al. 2022).

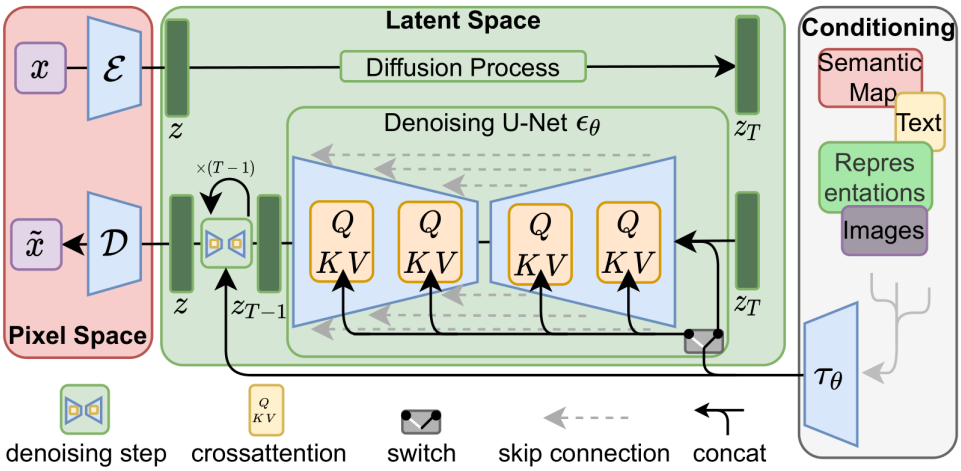


Figure 2 Architecture of Latent Diffusion Models
Source: Rombach et al. (2022).

Rombach et al. (2022) experimented with the autoencoder using six different downsampling factors, $f \in \{1, 2, 4, 8, 16, 32\}$, and found that LDMs with downsampling factors of 4 and 8 produced the best results for faithful image generation. Moreover, they investigated with two types of regularization, KL-reg and VQ-reg, to prevent arbitrarily high variance in the latent representations.

Subsequently, the diffusion process is applied in the latent space z_t (Rombach et al. 2022), as

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t)\|^2] \tag{3.3}$$

The U-Net architecture, based on the convolutional neural networks (CNNs), achieved excellent performance in image segmentation tasks (Ronneberger et al. 2015). It consists of 23 convolutional layers, as depicted in Figure 3. Inspired by this, Ho et al. (2020) proposed U-Net with self-attention in diffusion probabilistic models to generate high-quality images. Subsequently, Dhariwal and Nichol (2021) conducted ablation studies on U-Net–based diffusion models and suggested the Ablated Diffusion Model (ADM) and its guided variant ADM-G, identifying the most effective architectural configurations for image synthesis. Later, Rombach et al. (2022) further modified this ablated U-Net and incorporated it as a denoising network into the latent diffusion models (LDMs). The self-attention layers of the ablated U-Net are replaced with transformer blocks, where each block consists of alternating layers of self-attention, position-wise MLP, and cross-attention, with the cross-attention layer specifically used to incorporate conditioning information.

The internal structure of the UNet used in LDM consists of 4 downsampling blocks, 4 upsampling blocks, and one mid block, as shown in Figure 4. The first three downsampling blocks and the last three upsampling blocks include Transformer models, ResNet blocks, and down-

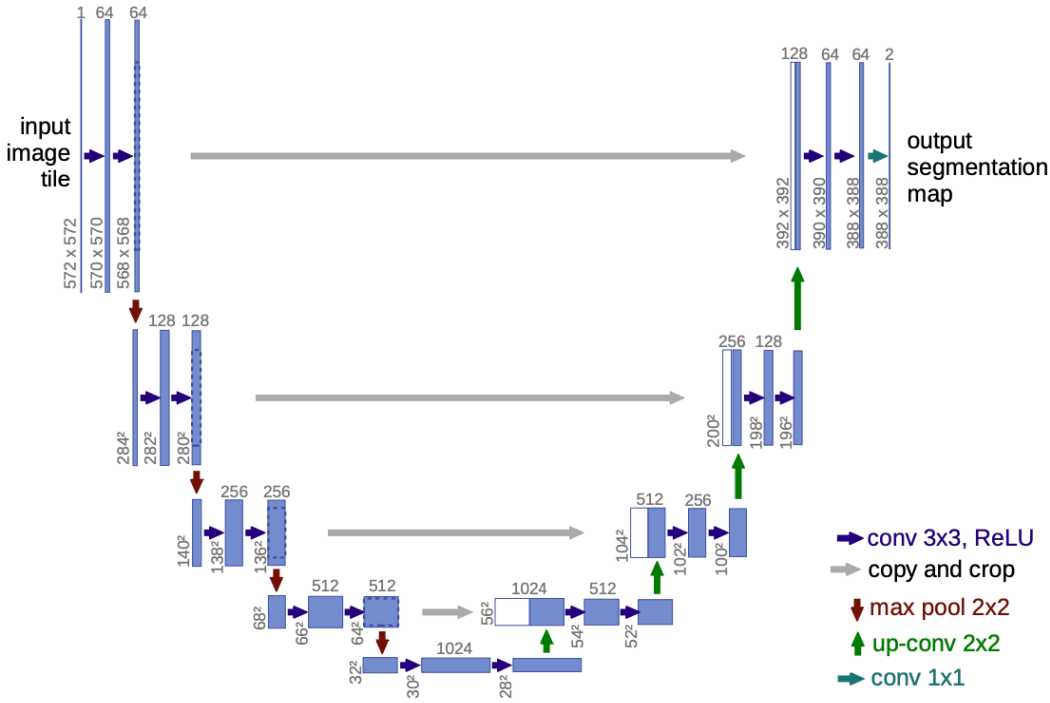


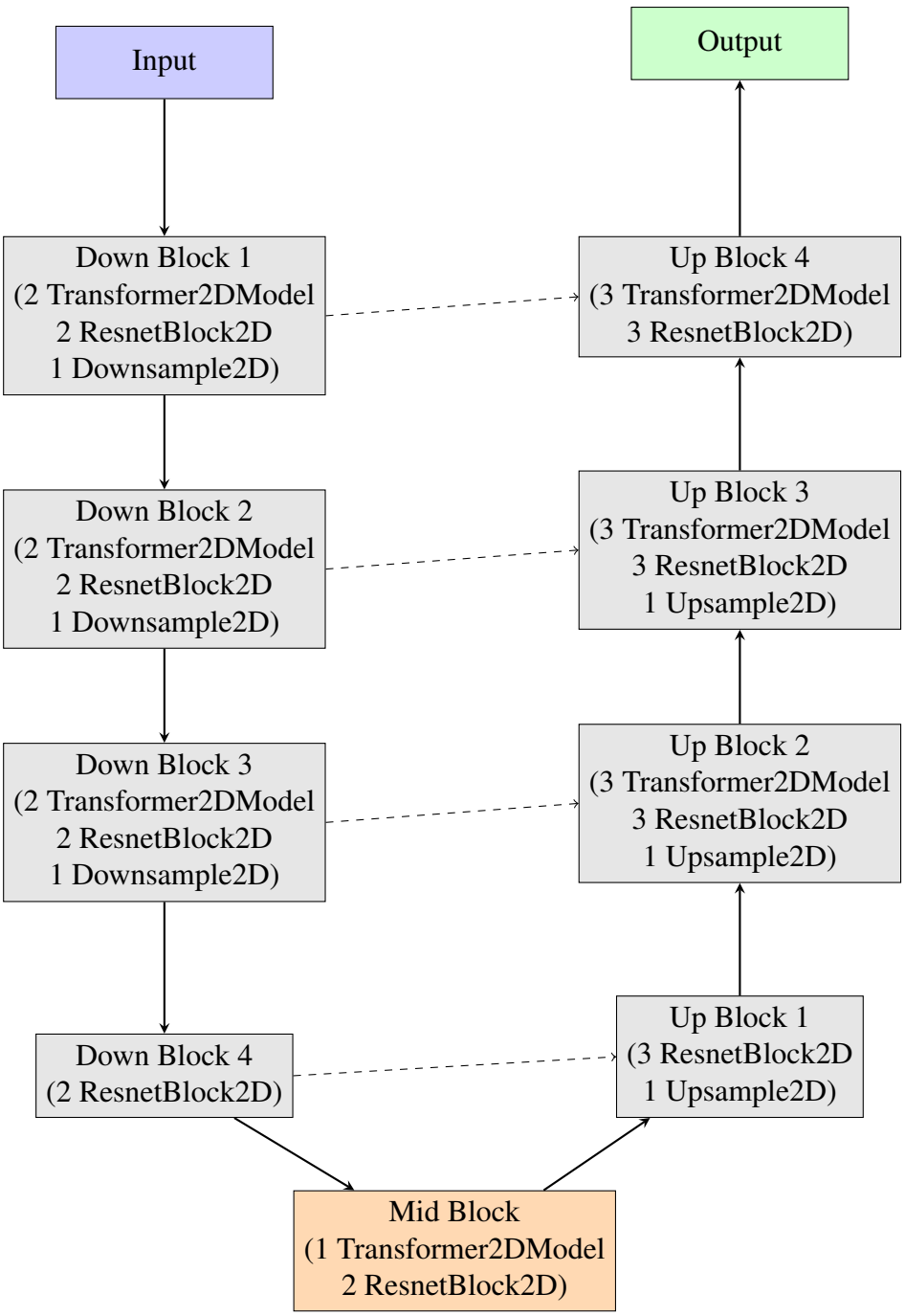
Figure 3 U-net architecture.
Source: Ronneberger et al. (2015).

samplers or upsamplers. The last downsampling block and the first upsampling block do not contain Transformer models. The mid block contains both transformer models and ResNet layers. Each Transformer model includes one Basic Transformer Block with self-attention (attn1), cross-attention (attn2), and a feed-forward (ff) layer.

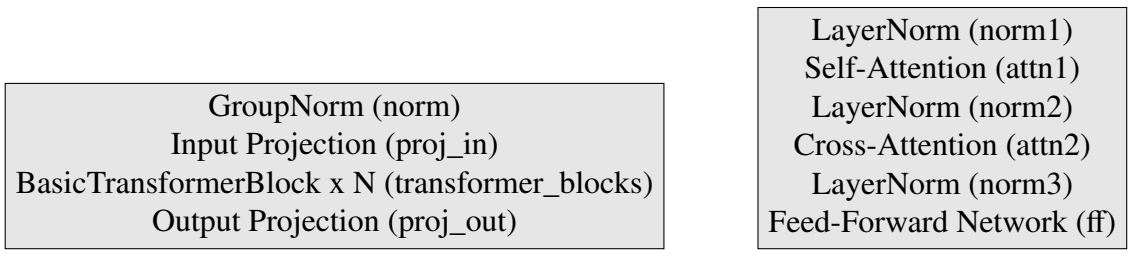
A guiding mechanism is used to control the generation process using a conditional input y, which is first converted to an intermediate representation through a domain-specific encoder τ_θ. Both the denoising network ϵ_θ and the encoder τ_θ are jointly optimized so that the model learns to generate images conditioned on y. The objective function can be rewritten (Rombach et al. 2022):

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|^2] \tag{3.4}$$

The conditioner τ_θ is implemented by Rombach et al. (2022) using Transformer blocks composed of multi-head self-attention layers, layer normalization, and position-wise MLPs. They experimented with two types of conditioning: text-to-image and layout-to-image synthesis. For text-to-image generation, the text is first tokenized, encoded into a latent representation by a Transformer encoder (with a maximum sequence length of 77 tokens, hidden dimension of 1280, and depth of 32 layers), and then used to condition the U-Net via cross-attention.



(a) Core component of the UNet2DConditionModel. Dashed line (- -) indicates the skip connections.



(b) Component of the Transformer2DModel.

(c) Basic Transformer Block

Figure 4 Parts of Internal structure and components of the U-Net used in the Stable Diffusion 2.0 base model (Stability AI 2022), available through Hugging Face's Diffusers library (von Platen et al. 2022).

3.2 Attention Mechanisms

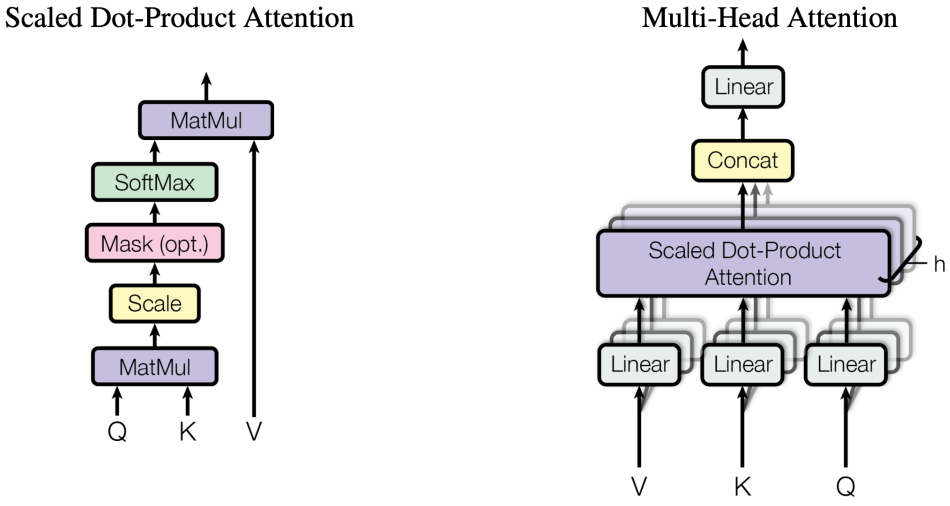


Figure 5 Illustration of attention computation.
Source: Vaswani et al. (2017).

Self-attention helps capture long-range dependencies and contextual relationships of an element, such as a word (Vaswani et al. 2017) or an image patch (Dosovitskiy et al. 2021), with other elements of the same type in a sequence (Vaswani et al. 2017). In scaled dot-product attention, queries Q, keys K, and values V of dimensions d_k, d_k, and d_v, respectively, are computed from the input embedding X using the projection matrices W^Q, W^K, and W^V:

$$Q = W^Q X, \quad K = W^K X, \quad V = W^V X \tag{3.5}$$

The dot product between queries and keys produces similarity scores between tokens. These scores are scaled by $\sqrt{d_k}$ to prevent the softmax function from saturating when the dot products are large, which would otherwise lead to vanishing gradients. The softmax then normalizes the scores into a probability distribution over all keys for each query. Finally, the outputs are multiplied with the value vectors V, producing a weighted combination that preserves the input information and yields the context vector (Vaswani et al. 2017).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.6}$$

Multiple heads in multi-head attention allow the model to perform attention computations in parallel and capture different types or aspects of relationships simultaneously. The outputs of all heads are then concatenated and projected using a weight matrix W^O (Vaswani et al. 2017).

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{3.7}$$

When it comes to multimodal models, two different sequences are combined using cross-attention. In Latent Diffusion Models (LDMs), Rombach et al. (2022) employed cross-attention to guide the diffusion process during image generation. For text-to-image generation, the queries are mapped from the noise latent, while the keys and values are mapped from the text embeddings.

In the Stable Diffusion 2.0 base model (Stability AI 2022) for a 512×512 pixel image, there are 16 cross-attention blocks (see Figure 4). The latent representation has a initial spatial size of 64×64, which, after flattening, results in 4096 tokens. These form the queries $Q \in \mathbb{R}^{4096 \times d_k}$. The text embedding has a sequence length of 77, which forms the keys $K \in \mathbb{R}^{77 \times d_k}$. Therefore, the attention score matrix then results in $QK^T \in \mathbb{R}^{4096 \times 77}$. The spatial dimensions change during sampling; however, the dimensionality of the keys remains unchanged.

In terms of Stable Diffusion (SD) 3.0, conditional information, — timestep embeddings and coarse-grained text representations — was injected into the attention mechanism by modulating the hidden representations of text and image separately for each modality, allowing the model to align visual and textual features more effectively. The sequences from both modalities were then concatenated to form a shared representation used for the attention operation (see Section 3.4; Esser et al. 2024).

3.3 Diffusion Attentive Attribution Maps

Many researchers have investigated the processing of linguistic information in latent diffusion models (LDMs) for text-to-image generation in various ways, as discussed in related work. However, Tang et al. (2023) interpreted Stable Diffusion by aggregating normalized cross-attention scores, generating attribution maps, and analyzing the syntactic and semantic properties processed by the LDM. They proposed an attribution method, Diffusion Attentive Attribution Maps (DAAM), to investigate the influence of individual words on the generated image.

From Equation 3.4, the denoising network is given by $\epsilon_\theta(z_t, t, \tau_\theta(y))$, where conditioning is applied using a domain-specific encoder $\tau_\theta(y)$. The output of this encoder is the embeddings $X \in \mathbb{R}^{M \times d_\tau}$. These embeddings, corresponding to words or tokens, are fed into each transformer block of the U-Net. Each downsampling or upsampling block can be represented as $h_{i,t}^{\downarrow}$ or $h_{i,t}^{\uparrow}$, respectively, where i denotes the block index and t denotes the time step. The output of the pre-

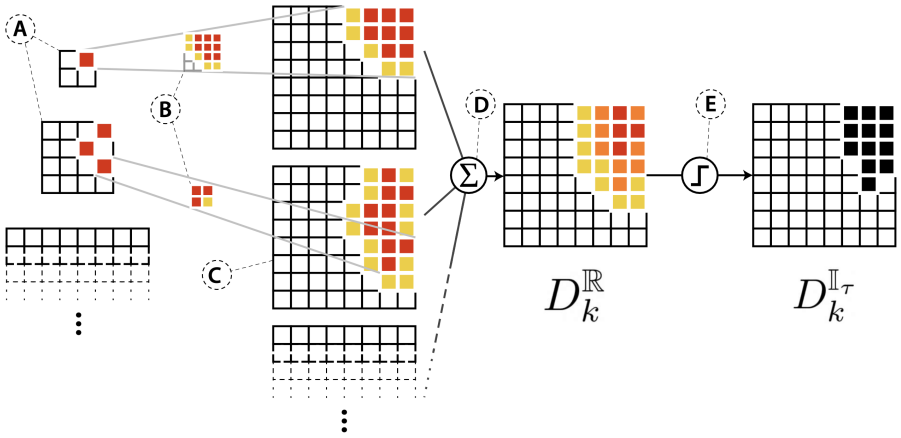


Figure 6 Visualization of the DAAM computation.
Source: Tang et al. (2023).

vious block or layer is denoted by $\hat{h}_{i,t}^{\downarrow}$ or $\hat{h}_{i,t}^{\uparrow}$. Finally, normalized scaled cross-attention scores are computed as follows (Tang et al. 2023; Rombach et al. 2022):

$$F_t^{(i)}(\hat{h}_{i,t}^{\downarrow}, X) := \text{softmax}\left(\frac{(W_q^{(i)}\hat{h}_{i,t}^{\downarrow})(W_k^{(i)}X)^T}{\sqrt{d}}\right) \tag{3.8}$$

and the attention output is given by

$$h_{i,t}^{\downarrow} := F_t^{(i)}(\hat{h}_{i,t}^{\downarrow}, X)(W_v^{(i)}X) \tag{3.9}$$

Here, W_q, W_k, and W_v are the projection matrices for the queries Q, keys K, and values V, respectively. In cross-attention, the queries from the latent patches attend to the keys of the input tokens. The resulting normalized scaled scores are then multiplied with the value vectors of the text tokens to generate a context vector. For the k-th word, these scores are interpolated to the original image size $[x, y]$ using bicubic interpolation and are aggregated across all heads l, blocks i, and denoising time steps t_j in the U-Net of the LDM (3.10). The index j serves as a constraint that is later used for ablation studies (Tang et al. 2023).

$$D_k^{\mathbb{R}}[x, y] := \sum_{i,j,l}\left(\tilde{F}_{t_j,k,l}^{(i)\downarrow}[x, y] + \tilde{F}_{t_j,k,l}^{(i)\uparrow}[x, y]\right) \tag{3.10}$$

A soft heat map can be generated from $D_k^{\mathbb{R}}$; however, Tang et al. (2023) converted these aggregated scores into a hard or binary heat map using an indicator function $D_k^{\mathbb{I}_\tau}$ and a threshold $\tau \in [0, 1]$, based on the maximum score value (see 3.11; Figure 6).

$$D_k^{\mathbb{I}_\tau}[x, y] := \mathbb{I}\left(D_k^{\mathbb{R}}[x, y] \geq \tau \max_{i,j} D_k^{\mathbb{R}}[i, j]\right) \tag{3.11}$$

The image is generated using the prompt and the Stable Diffusion 2.0 base model (Stability AI 2022) with 30 inference steps (Tang et al. 2023). Simultaneously, the normalized scaled attention scores are collected and stored in tuples ($factor, layer, head$), where $factor$ represents the downsampling or upsampling factor of the latent patch in a particular U-Net block, *layer* indicates the U-Net block, and *head* refers to the cross-attention head within that layer. These scores are summed across each denoising time step. After image generation is completed, the scores are interpolated to the original image size. The resulting matrix contains the interpolated attention scores for all 77 tokens, which is the maximum sequence length of the model's text encoder. The first N tokens in the matrix correspond to the first N tokens in the prompt. Then a threshold τ is applied to generate the DAAM maps for the tokens $n \in N$ of interest.

The faithfulness of the method was quantitatively evaluated using an image segmentation task (Tang et al. 2023). The images were generated from prompts collected from the COCO dataset, and corresponding DAAM maps were created for the countable nouns in each prompt. The ground truth masks were obtained by manually segmenting the objects corresponding to the countable nouns in the images. The mean Intersection over Union (mIoU) score was then calculated on N prompt-image pairs as

$$\text{mIoU} = \frac{1}{N} \sum_{n=1}^{N} \frac{|A_n \cap B_n|}{|A_n \cup B_n|}, \tag{3.12}$$

DAAM was compared with other methods by Tang et al. (2023), and it was shown that the method performs best with a threshold value of $\tau = 0.4$. The analysis was extended to verify whether DAAM can be generalized to other parts of speech. Six interpretable part-of-speech tags were selected in the analysis and a human evaluation was conducted to assess the quality of the attribution. After establishing the plausibility of DAAM, visual syntactic and semantic analyzes were examined to understand the impact of words on the generated maps.

$$\text{mIoH} = \frac{1}{N} \sum_{n=1}^{N} \frac{|A_n \cap B_n|}{|A_n|}, \tag{3.13}$$

$$\text{mIoD} = \frac{1}{N} \sum_{n=1}^{N} \frac{|A_n \cap B_n|}{|B_n|}, \tag{3.14}$$

The dependency relationships between the head and the dependent words were investigated in the syntactic analysis (Tang et al. 2023). Three overlap statistics were used: mean Intersection over Union (mIoU) (3.12), mean Intersection over Head (mIoH) (3.13), and mean Intersection over Dependent (mIoD) (3.14). A_n and B_n represent the binarized map of the head word and the dependent word for the n-th pair among N head–dependent pairs. The mIoU score indicates

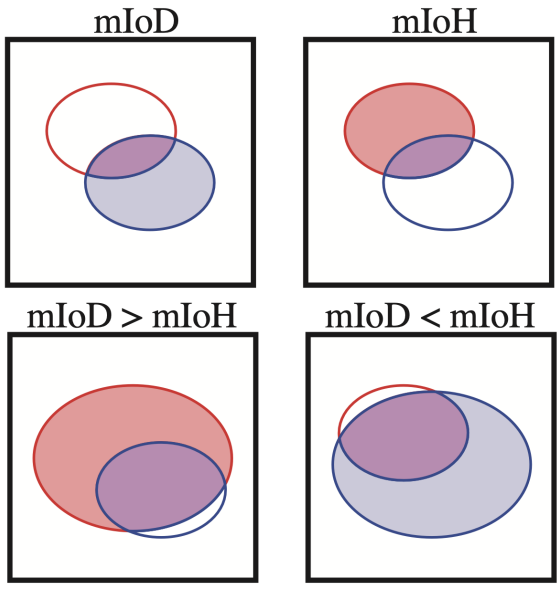


Figure 7 Illustrations of mIoD, mIoH, and various types of overlap.
Source: Tang et al. (2023).

how much pixels overlap between the binarized maps of the head and the dependent words. The mIoD and mIoH scores indicate the proportion of pixels in the dependent and head maps, respectively, that fall within their intersected region. If mIoD > mIoH, the head map overlaps the dependent map more than the dependent map overlaps the head. Conversely, if mIoH > mIoD, the dependent map covers the head map more than the head map covers the dependent (see Figure 7).

3.4 Stable Diffusion 3

Flow-based generative models are a class of deep generative models that can model complex, high-dimensional probability densities by learning a composition of invertible mappings between a simple distribution (such as Gaussian noise) and complex data distributions, as illustrated by NICE (Dinh et al. 2015), RealNVP (Dinh et al. 2017), Glow (Kingma and Dhariwal 2018), Flow++ (Ho et al. 2019). This framework allows for exact log-likelihood computation using the change-of-variables formula and enables efficient sampling by applying the inverse transformations. A flow model f maps observed data x to latent noise $z = f(x)$ and the inverse f^{-1} maps z back to the data. The overall transformation is constructed as a sequence of invertible mappings, as illustrated below:

$$x \overset{f_1}{\leftrightarrow} h_1 \overset{f_2}{\leftrightarrow} h_2 \cdots \overset{f_k}{\leftrightarrow} z \tag{3.15}$$

Flow-based models can efficiently generate samples and compute exact densities because each f_i has a tractable inverse and a tractable Jacobian. The composition of such invertible transfor-

mations is known as a Normalizing Flow (Rezende and Mohamed 2015). Continuous Normalizing Flows (CNFs) extend normalizing flows (NFs) into continuous time by parameterizing the transformation from noise to data as a deterministic path that is formulated as an ordinary differential equation (ODE) (Chen et al. 2018):

$$\frac{dz}{dt} = f_\theta(z(t), t), \tag{3.16}$$

The resulting trajectory can be solved with an ODE solver, but generating a sample requires simulating the entire forward path. Flow Matching (FM) addresses this by providing a simulation-free training procedure, where a neural network is trained to regress directly to the ground truth vector field u_t along a probability path p_t, rather than requiring simulation of data or noise trajectories over time. Its objective function is as follows (Lipman et al. 2023):

$$\mathcal{L}_{FM} = \mathbb{E}_{t,\, p_t(z)} \| v_\theta(z, t) - u_t(z) \|^2 \tag{3.17}$$

However, marginalization of p_t makes it intractable to compute u_t. Therefore, a tractable objective with explicit conditioning is applied in Conditional Flow Matching (CFM), as follows (Lipman et al. 2023; Esser et al. 2024):

$$\mathcal{L}_{CFM} = \mathbb{E}_{t,\; p_t(z|\epsilon)\, p(\epsilon)} \| v_\theta(z, t) - u_t(z|\epsilon) \|^2 \tag{3.18}$$

The loss was then converted by Esser et al. (2024) to align with the general form of the diffusion process

$$z_t = a_t x_0 + b_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \tag{3.19}$$

which defines a probability path from data x_0 to noise ϵ. They reparameterized the CFM objective (3.18) to express it in noise prediction form, introduced a time-dependent weighting function, and provided a unified objective:

$$\mathcal{L}_w(x_0) = -\frac{1}{2} \mathbb{E}_{t \sim U(t),\, \epsilon \sim \mathcal{N}(0,I)} \Big[w_t \, \lambda_t' \, \| \epsilon_\theta(z_t, t) - \epsilon \|^2 \Big] \tag{3.20}$$

They analyzed four different generative training objectives: Rectified Flow (RF) (Liu et al. 2023), EDM (Karras et al. 2022), Cosine (Nichol and Dhariwal 2021), and Latent Diffusion Models (LDM) (Rombach et al. 2022). Among these, Rectified Flow (RF) defines a straight path that transports the data distribution to noise (and vice versa) (Liu et al. 2023; Esser et al. 2024):

$$z_t = (1 - t) x_0 + t \epsilon \tag{3.21}$$

and the corresponding weighting function is (Esser et al. 2024):

$$w_t^{RF} = \frac{t}{1 - t} \tag{3.22}$$

The uniform distribution $\mathcal{U}(t)$ is then changed to a distribution with timestep densities $\pi(t)$ to consider intermediate steps (Esser et al. 2024):

$$w_t^{\pi} = \frac{t}{1-t}\pi(t) \tag{3.23}$$

In Latent Diffusion Models (LDMs), the neural network ϵ_θ is trained to predict the noise directly. In contrast, the network in Stable Diffusion (SD) 3.0 is supposed to regress the velocity field u_t under the Flow Matching (FM) objective. However, in Rectified Flow (FM), velocity is simply the difference between data and noise, $x_0 - \epsilon$, which is equivalent to predicting noise with linear transformation. Therefore, the neural network in SD 3.0 empirically predicts the noise under the flow matching and rectified flow objective, rather than using the standard diffusion objective.

Similar to Latent Diffusion Models (LDMs), Stable Diffusion 3.0 trained its text-to-image generation model in the latent space (Esser et al. 2024). The RGB image $X \in \mathbb{R}^{H\times W\times 3}$ was first mapped to a latent representation $x \in \mathbb{R}^{h\times w\times d}$ using the encoder $\mathcal{E}$ of the pretrained autoencoder, $x = \mathcal{E}(X)$, with a downsampling factor $f = 8$. They experimented with three different values of latent channels, $d \in \{4, 8, 16\}$, and observed that reconstruction performance improved as the number of latent channels increased. Therefore, they chose $d = 16$ as they decided to scale to larger model sizes. The decoder $\mathcal{D}$ reconstructs the image from the latent as $X \approx \mathcal{D}(x)$. The forward process was applied in the latent space using Equation 3.19. Sampling a representation was performed using Equation 3.16, and the latent was normalized by its mean and standard deviation similar to the LDM. This was then divided into 2×2 patches, each patch was flattened into a vector, positional encodings were added, and the resulting vectors were projected into patch embeddings that were propagated through the network.

Text representations were obtained during training using three pretrained and frozen text encoders (Esser et al. 2024): CLIP-L/14 (Radford et al. 2021), OpenCLIP bigG/14 (Cherti et al. 2023), and T5-v1.1-XXL (Raffel et al. 2020). At inference time, any subset of these encoders can be utilized; however, empirical results indicate that using all three leads to improved performance on complex prompts. Removing the T5 encoder during inference does not affect aesthetic quality and has only a minor impact on prompt adherence. In contrast, including T5 improves the model's ability to generate written text within images.

More specifically, the captions c were processed by three text encoders to obtain a vector representation c_{vec} and a context representation c_{ctxt} (Esser et al. 2024). The pooled outputs of CLIP-L/14 (dimension 768) and OpenCLIP-bigG/14 (dimension 1280) were concatenated to form $c_{\text{vec}} \in \mathbb{R}^{2048}$. For the context representation, the outputs of all three encoders, each of sequence length 77, were considered. The output of CLIP L/14, $c_{ctxt}^{CLIP_{CLIP\ L/14}} \in \mathbb{R}^{77\times 768}$, and OpenCLIP biG/14, $c_{ctxt}^{CLIP_{OpenCLIP\ biG/14}} \in \mathbb{R}^{77\times 1280}$, were concatenated along the channel axis to form CLIP representations, $c_{ctxt}^{CLIP} \in \mathbb{R}^{77\times 2048}$. The zero-padding was then applied along the channel axis of this CLIP representation to align its dimensions with the output of T5-v1.1-XXL, $c_{ctxt}^{T5} \in \mathbb{R}^{77\times 4096}$.

Finally, $c_{\text{ctxt}}^{\text{CLIP}}$ and $c_{\text{ctxt}}^{\text{T5}}$ were concatenated along the sequence axis, resulting in the final context representation $c_{\text{ctxt}} \in \mathbb{R}^{154\times4096}$.

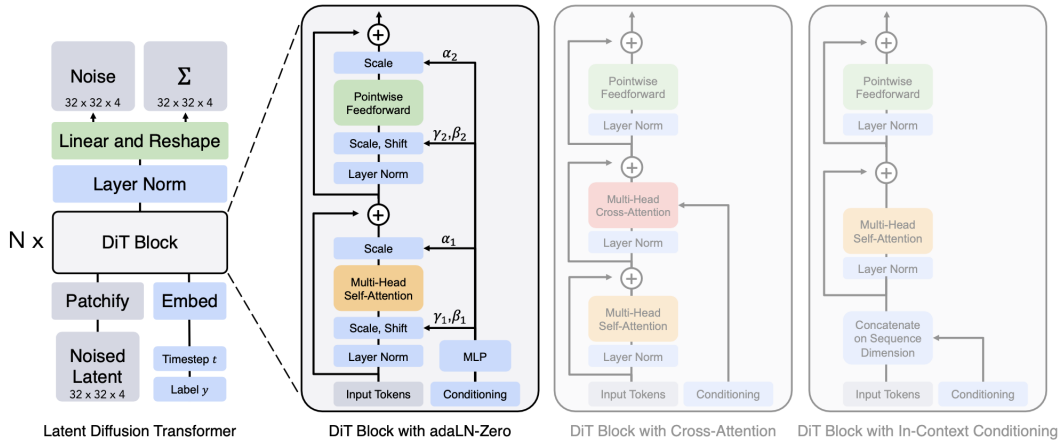
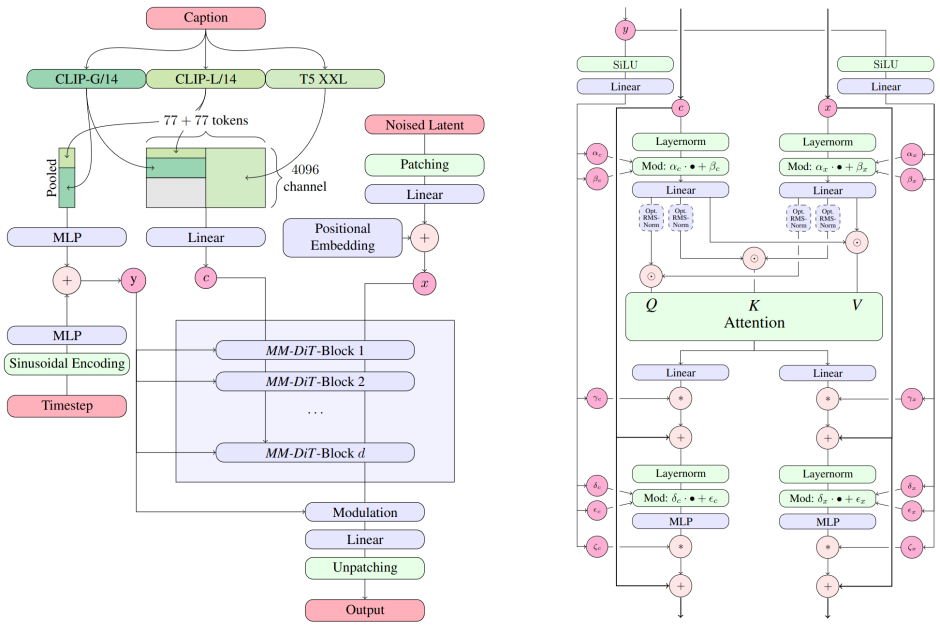


Figure 8 Architecture of Diffusion Transformer (DiT).
Source: Peebles and Xie (2023).

In Latent Diffusion Models (LDMs), a fixed text representation was fed directly into the U-Net blocks (Rombach et al. 2022). Diffusion Transformers (DiTs) followed LDM but replaced the U-Net with a transformer (Peebles and Xie 2023). Conditional information was incorporated into the DiT blocks by combining timestep embeddings and the class label, and different conditioning strategies were investigated: adaptive layer normalization (adaLN), cross-attention, and in-context conditioning. The adaLN was extended with a zero-initialization strategy (adaLN-Zero), where an extra scaling parameter α was applied prior to any residual connections (see Figure 8), and AdaLN-Zero was found to outperform other variants.

The architecture of Stable Diffusion (SD) 3.0 was based on the DiT architecture (see Figure 9) (Esser et al. 2024). A two-way flow of information was propagated through each Multimodal Diffusion Transformer (MMDiT) block, allowing bidirectional mixing between text and image token streams within the network. Separate modulation signals were applied at various points in the MMDiT block, similar to the AdaLN-Zero mechanism. Different learnable modulation parameters for scaling and shifting operations were computed from a combined representation, which was created by adding the timestep embedding t to the pooled text representation c_{vec}. Both scaling and shifting operations were performed before and after attention operations for each modality. Specifically, the context representation c_{ctxt} and the latent image representation x were modulated by applying $\alpha_c \cdot h + \beta_c$ before attention and using $\delta_c \cdot h + \epsilon_c$ after attention. Additionally, only the scaling parameters γ_c and ζ_c were applied before each residual connection (see Figure 9).

Each MMDiT block contained a single attention block with depth denoted by d (Esser et al. 2024). The dimensionality of the hidden representations in the attention blocks was set to



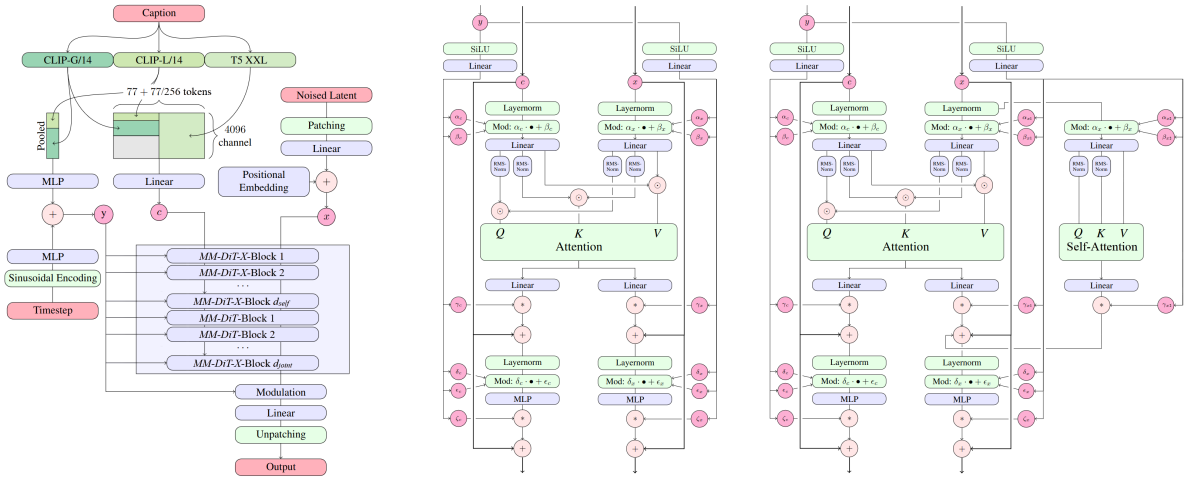
(a) Overview of all components. (b) One *MM-DiT* block

Figure 9 Architecture of Stable Diffusion 3.0 with MM-DiT block.
Source: Esser et al. (2024).

$64 \cdot d$. The MLP within each block expanded this dimension by a factor of 4, resulting in $4 \cdot 64 \cdot d$ channels. The number of attention heads in each block was equal to d, keeping the head dimension constant at 64. Scaling studies were conducted by varying the model size through different values of the depth d.

Stability AI released a subsequent version, Stable Diffusion (SD) 3.5 (Stability AI 2024), in which they extended the MMDiT block of SD 3.0 (Stability AI 2022) to MMDiT-X (see Figure 10) by introducing an additional self-attention module in the first 13 blocks (or layers) of the transformer. Normalizing the query and key embeddings before calculating the attention matrix was found to prevent the growth and instability of the attention logits (Esser et al. 2024); therefore, QK-Normalization was implemented using RMSNorm (Zhang and Sennrich 2019) in SD 3.5. Multi-resolution image generation performance was enhanced through staged training with gradually increasing latent resolutions, including training on images of varying scales at the final stage, extending the positional embedding space, and applying random crop augmentation on the positional embeddings (Stability AI 2024). However, the modulation techniques and processing of the text representations remained unchanged from SD 3.0.

After analyzing the Stable Diffusion 3.5 Medium model (Stability AI 2024) for 512×512 images, it was found that the network contains a total of 24 MMDiT-X blocks, corresponding to a depth of $d = 24$. Each block consists of a single joint attention module, while the first 13 blocks additionally include a self-attention module applied to the image stream. The sequence lengths of the hidden text and image representations are 154 and 1024, respectively, with both



(a) Overview of all components. (b) One *MM-DiT* block (c) One *MM-DiT-X* block

Figure 10 Architecture of Stable Diffusion 3.5 with MMDiT and MMDiT-X block. *Source: Stability AI (2024).*

having a hidden dimensionality of $64 \times 24 = 1536$. The implementation of attention processor is based on the Joint Attention Processor 2.0 from the Diffusers library developed by Hugging Face [1].

Queries Q, keys K, and values V are projected from the hidden representations using learned weight matrices. The hidden text representation is denoted as $c_{ctxt} \in \mathbb{R}^{154\times1536}$, and the hidden image representation as $x \in \mathbb{R}^{1024\times1536}$. The outputs of Q, K, and V maintain the same dimensionality as their corresponding input representations.

Each of these is then reshaped into multiple attention heads, with a constant head dimension of 64, resulting in Q, K, and V tensors of shape $\mathbb{R}^{154\times64}$ for text and $\mathbb{R}^{1024\times64}$ for image, for each head. The hidden representations of text are then concatenated to the image representations along the sequence dimension, forming a joint representation of $\mathbb{R}^{1178\times64}$ for each of Q, K, and V.

The attention scores are computed as $QK^\top \in \mathbb{R}^{1178\times1178}$, allowing every token from text or image to attend to all others across both modalities, enabling bidirectional mixing between modalities. These attention scores are then multiplied with V to generate the context tensor of shape $\mathbb{R}^{1178\times64}$, which is reshaped back to $\mathbb{R}^{1178\times1536}$.

Finally, this combined context vector is split according to the sequence lengths of the text and image representations and propagated to the next transformer layer. For the additional self-attention modules, attention scores and outputs are computed only over the image representations.

[1] Hugging Face, *Diffusers Library – Joint Attention Processor 2.0*, available at: https://github.com/huggingface/diffusers/blob/v0.35.1/src/diffusers/models/attention_processor.py#L1420.

4 Our Approach

In Diffusion Attentive Attribution Maps (DAAM) (Rombach et al. 2022), as discussed in Section 3.3, attention scores are aggregated across all heads, layers, and steps of cross-attention–based Stable Diffusion models. These scores are then interpolated to the size of the image to generate attribution maps. The attribution maps can be computed for specific words, helping to identify the influence of those words on particular regions of the generated image. However, DAAM does not support the state-of-the-art Stable Diffusion 3 or 3.5 models, because these models use a joint representation of image and text for the attention operation instead of a cross-attention–based approach (see Section 3.4).

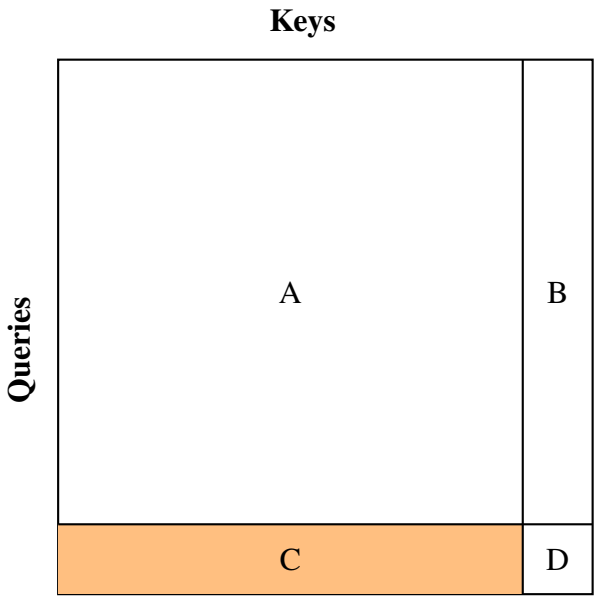
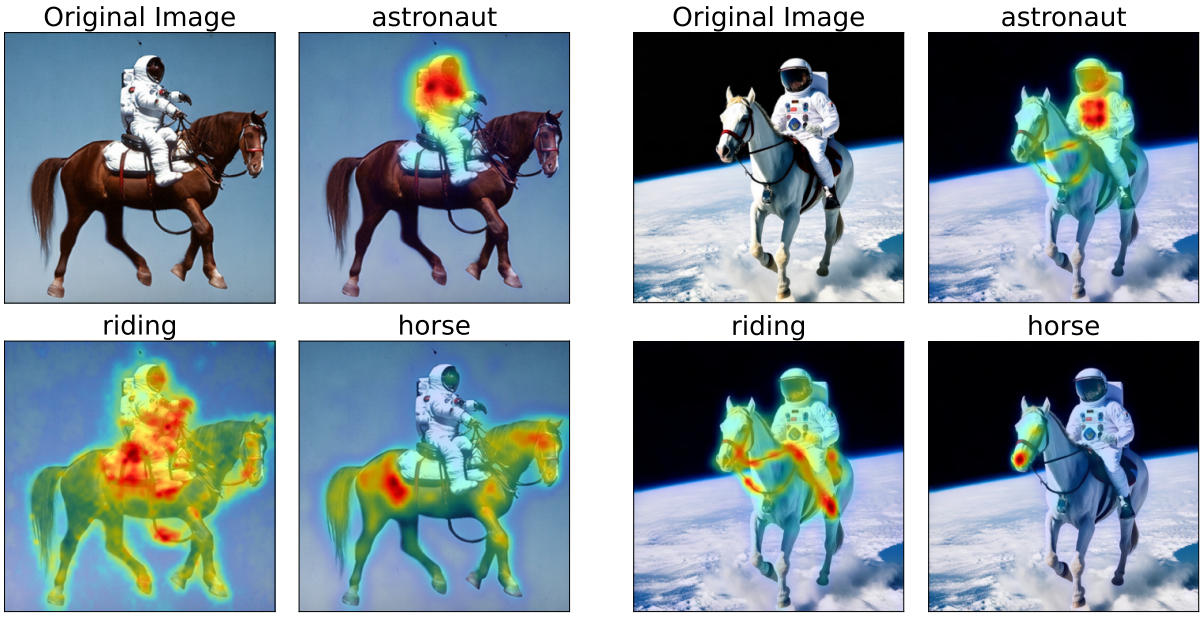


Figure 11 Visualization of the attention score matrix including different regions. **A:** Query of the image tokens attend to keys of the image; **B:** Query of the image tokens attend to keys of the text; **C:** Query of the text tokens attend to keys of the image; **D:** Query of the text tokens attend to keys of the text.

In this thesis, attention scores from each MMDiT block of Stable Diffusion 3 are aggregated following the DAAM methodology. Since the text representation is concatenated with the image representation along the sequence axis, the attention scores are collected only from the relevant region rather than aggregating all scores. The attention score matrix is divided into four regions, as illustrated in Figure 11. Among these regions, the last part of the queries, representing text tokens, attends to the first part of the keys, representing image tokens. Attention scores from this region—indicated as region C in Figure 11—are collected across all heads, layers, and steps. Since only a particular region of the attention scores is considered, it is not normalized, similar to DAAM. The subsequent process follows the DAAM approach, in which scores are interpolated to image size, attribution maps are generated, and syntactic experiments are carried out (see Section 5).

All images for the experiments were generated using the Stable Diffusion 3.5 Medium model from the Hugging Face diffusers library (Stability AI 2024). The image resolution was set to 512×512 pixels, and 30 inference steps were used for each generation. The T5 encoder (Raffel et al. 2020) was not utilized during inference, as it does not affect the aesthetic quality of the generated images (Esser et al. 2024, see Section 3.4). Captions from the COCO dataset (Lin et al. 2014) were used as text prompts.



(a) Attribution maps from the Stable Diffusion 2.0 Base model using DAAM.

(b) Attribution maps from the Stable Diffusion 3.5 Medium model using our approach.

Figure 12 Image generated with the prompt *'A photograph of an astronaut riding a horse.'* and the corresponding attribution maps for the words *'astronaut'*, *'riding'*, and *'horse'*.

5 Experiments and Analyses

5.1 POS Analysis

5.1.1 Experimental Setup

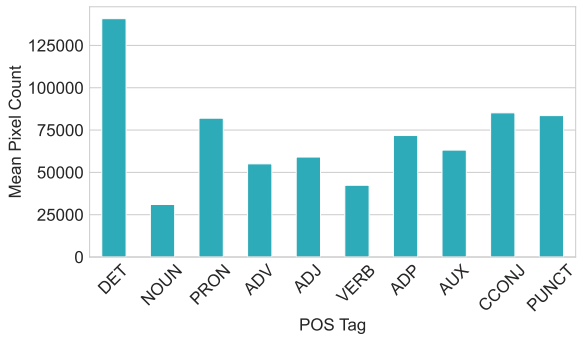
An experiment was conducted to investigate how important part-of-speech (POS) tags are represented in different parts of the generated images. This experiment was later extended to identify the layers that are influential in the generation process for specific POS tags.

POS tag frequencies were extracted using `spaCy` from the COCO dataset. A total of 591,753 captions were processed, and the top 10 POS tags were selected. For each caption, a single word corresponding to one of the selected POS tags was considered to introduce variation across tags. A total of 100 captions per POS tag, resulting in 1,000 distinct captions, were used for the experiments.

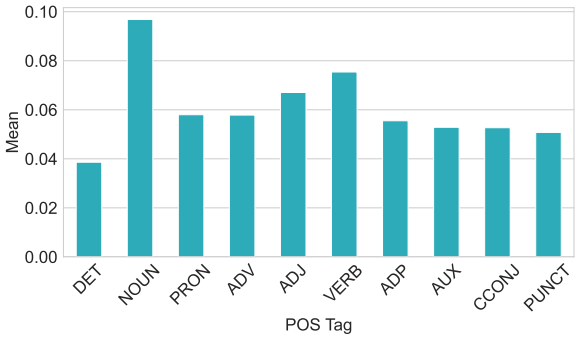
Using these captions, images were generated to examine the relationship between POS tags and attention patterns. For each generation, attention scores were aggregated, interpolated to match the image size, and binarized using max-based thresholding. A threshold value of 0.4 was used, similar to DAAM-0.4, as Tang et al. (2023) found it to perform best. The pixel count and mean attention values were then calculated across all layers as well as for individual layers. Additional examples are provided in Appendix A.1.

5.1.2 Observations

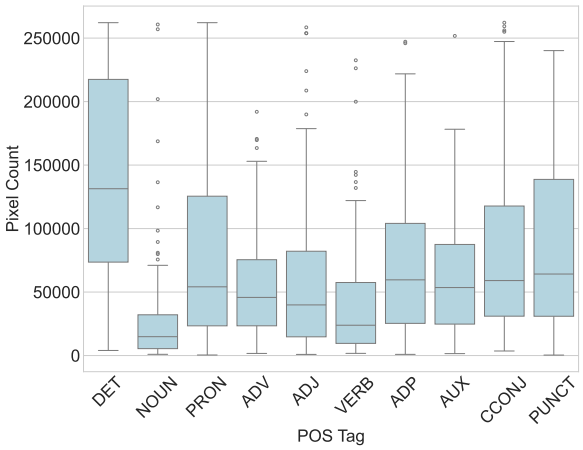
Overall, the average of high attention scores across all POS tags is around 0.06, which covers an average region of nearly 71K pixels. The *DET* tag was found to have the highest spread in terms of pixel count and covers a larger region (see Figure 13). Almost half of the image is captured by this tag—140K out of 262K(512 * 512) pixels. However, the attention scores for this tag are low, with a mean value less than 0.04. In other words, small attention values are spread over a large area. All high attention scores are concentrated in a small region for *Noun*, with *Verb* showing a similar but slightly less intense pattern. Other POS tags show a consistent pattern, with total pixel counts ranging from 50K to 86K, and corresponding mean attention scores falling between 0.05 and 0.07.



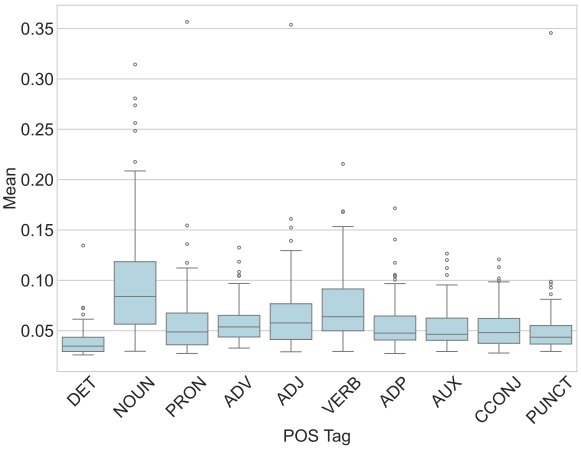
(a) Mean Pixel Count per POS Tag.



(b) Mean Attention Values per POS Tag.



(c) Distribution of Pixel Counts per POS Tag.



(d) Distribution of Mean Attention Values per POS Tag.

Figure 13 Analysis of pixel coverage and attention values across POS tags.

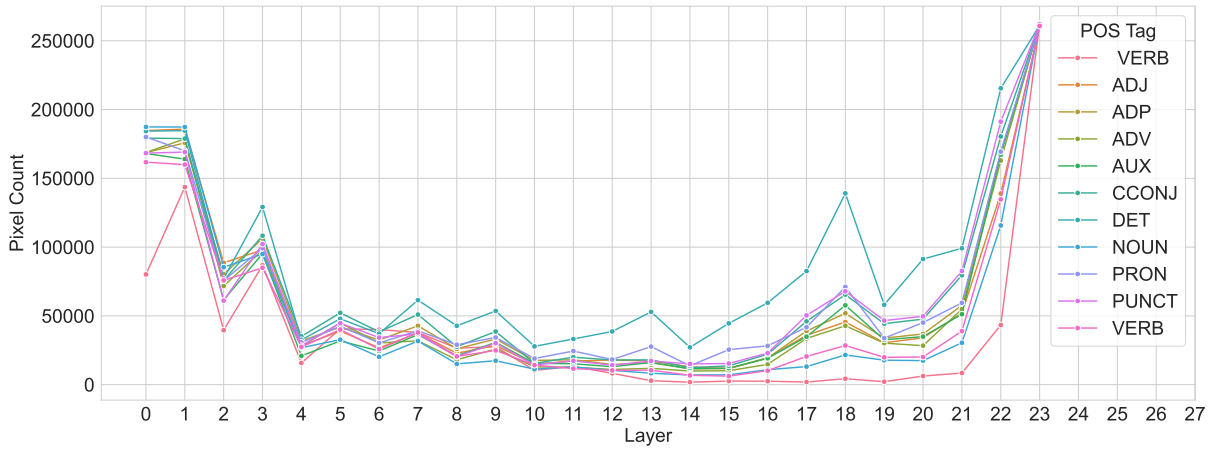


Figure 14 The Mean Pixel Count per Layer for all POS Tags.

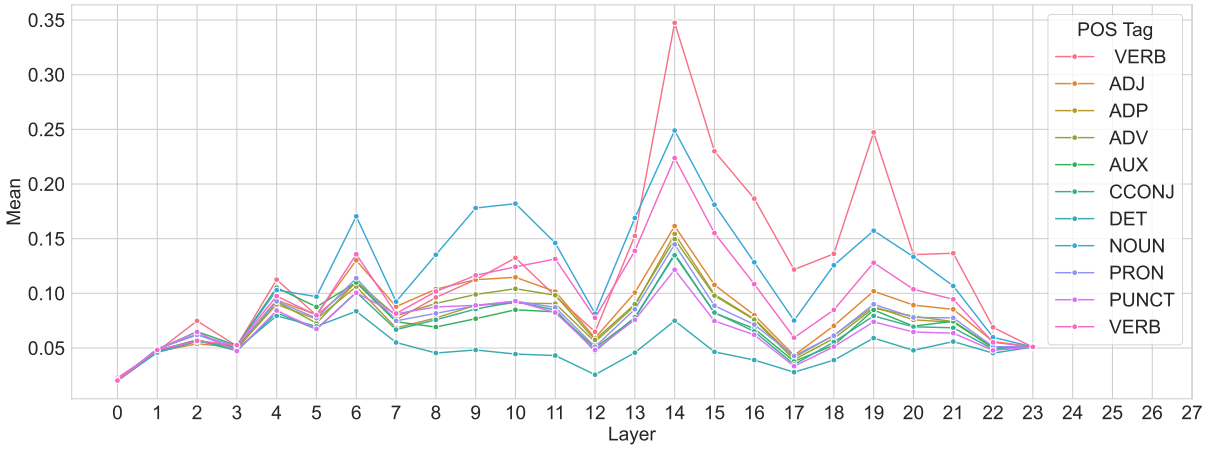


Figure 15 The Mean Attention Values per Layer for all POS Tags.

From a layer-wise perspective, the pixel count shows a U-shaped pattern (see Figure 14), where the middle layers (4–21) indicate that the smaller regions are utilized during the image generation process. In contrast, mean attention scores were expected to follow an inverse U-shaped trend, but this was not observed (see Figure 15). Except for layer 14, the scores fluctuate across layers, although a generally consistent pattern is observed among all POS tags.

Layer 14 shows the highest mean attention values and the lowest pixel counts in nearly all tags, suggesting that it plays a critical role in guiding the generation process. Notably, *Nouns* and *Verbs* exhibit a significant concentration of attention scores within smaller regions in this layer. Similarly, layers 12 and 17 show comparatively lower mean attention scores.

5.2 DEP Analysis

5.2.1 Experimental Setup

This experiment investigated how the dependency relationships (DEP) between words are captured in the generated images and conducted a layer-wise analysis to identify the significant layers.

All 591,753 captions from the COCO training dataset were processed to extract the top 12 dependency relationships using `spaCy`. A total of 1,000 captions were randomly sampled. Each caption was processed to extract dependency tags for all words, and only those belonging to the top 12 relationships were selected. Following the approach in the previous experiment (see Section 5.1.1), an image was generated for each caption, and the corresponding attention scores were aggregated, interpolated to the size of the image, and masks were created using max-based thresholding (DAAM-0.4). Similar to DAAM(Tang et al. (2023), see Section 3.3),

the IoU, IoD, and IoH scores were calculated using the head and dependent words of a tag, and then these scores were averaged per distinct tag.

For the layer-wise experiment, 100 captions were sampled from the 1,000 captions. The IoU, IoD, and IoH scores for each tag were calculated for each of the 24 layers. These scores were then averaged per distinct tag for each layer.

5.2.2 Observations

DEP Tag	Frequency	mIoU	mIoD	mIoH	Δ
det	2110	0.282	0.381	0.807	0.426
aux	157	0.299	0.415	0.717	0.303
punct	866	0.221	0.333	0.632	0.299
nsubj	401	0.254	**0.624**	0.411	0.214
cc	252	0.351	0.504	0.702	0.198
acl	397	0.255	0.410	0.585	0.174
prep	1686	0.346	0.512	0.686	0.174
amod	749	0.380	0.582	0.680	0.099
conj	292	0.239	**0.469**	0.404	0.066
pobj	1661	0.284	0.527	0.538	0.011
dobj	327	0.220	0.444	0.453	0.009
compound	513	0.341	0.579	0.587	0.008

Table 1 Overlap statistics for the 12 most frequent head—dependent relationships, where the **Δ** denotes the difference between **mIoD** and **mIoH**.

In general, mIoU scores are approximately half of those reported in DAAM (Tang et al. 2023, see Table 1). In most cases (10 out of 12), the dependent dominates the head more than vice versa — except for *nsubj* and *conj*.

The *punct* and *det* tags show little dominance in DAAM (Tang et al. 2023); however, our findings reveal the opposite pattern, with a strong dominance observed by the dependent. The *punct*, *det*, and *aux* attend too broadly for the dependent word and contains most of it in the head maps (see Figure 30). The largest difference between mIoD and mIoH was observed for the *det* tag, where the dependent map covers the head map more extensively than the head map covers the dependent.

Similar to DAAM (Tang et al. 2023), low dominance was observed for the *compound* tag, since both nouns complement each other. Furthermore, a low dominance was found for *pobj* and *dobj* in our investigation, although the strongest dominance for object by head word was reported in

DAAM. In both DAAM and our approach, the heads of *nsubj* and *conj* contain more of the dependents than the opposite, but the overlap is almost twice that observed in DAAM.

In terms of the coordinating conjunction (*cc*), the previous conjunct is considered as the head word by `spaCy`. For example, for the caption *'A plate filled with assorted veggies and rice.'* the dependent word is *'and'* and the head word is *'veggies'*. The attribution maps for the head overlap the dependent more than the dependent does for the head.

As expected, the verb dominates the noun for the *acl* tag and the preposition modifies the noun or verb for the *prep* tag. The adjectival modifier (*amod*) dominates the nouns it modifies; however, the difference Δ is slightly smaller than that observed by DAAM, by around 5 points.

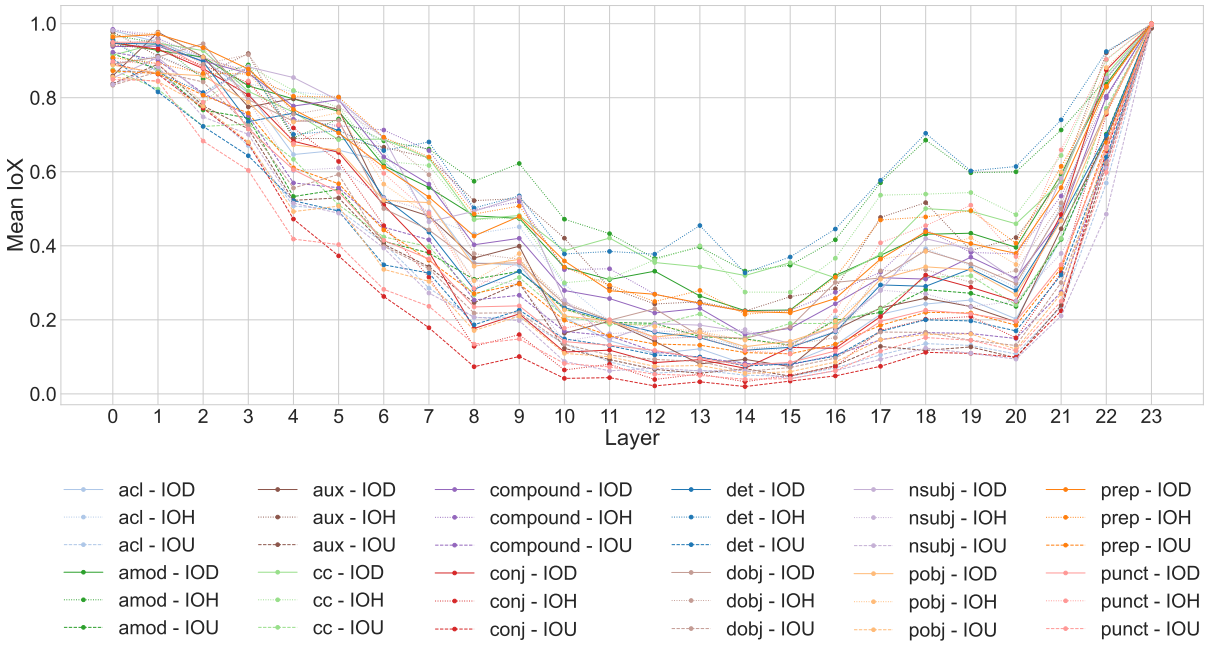


Figure 16 The mIoU, mIoD, and mIoH scores for all tags across layers.

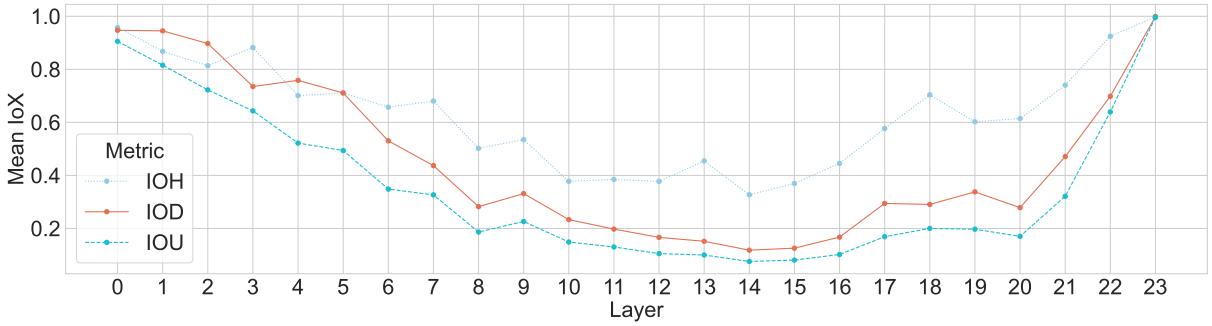


Figure 17 The mIoU, mIoD, and mIoH scores for the *det* tag across layers.

From a broader perspective, the mIoX (mIoU, mIoD, and mIoH) for all tags show a U-shaped pattern across 24 layers (see Figure 16). Figure 17 illustrates the det tag across all layers,

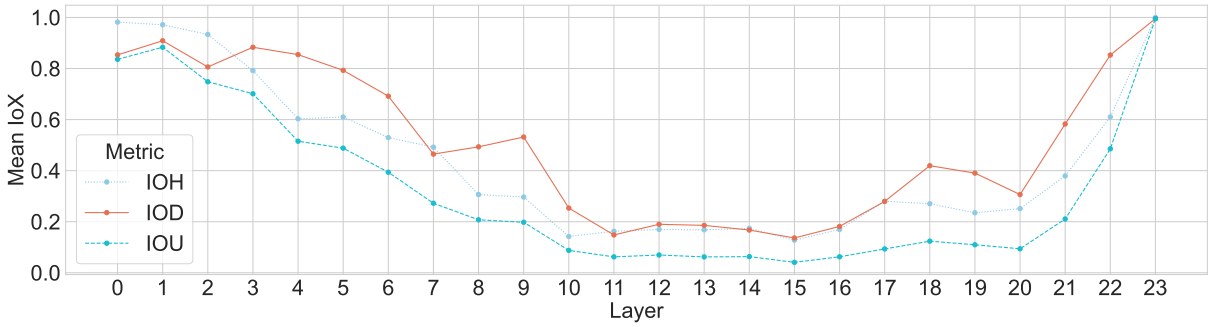


Figure 18 The mIoU, mIoD, and mIoH scores for the *nsubj* tag across layers.

where the dependent dominates the head, and Figure 18 illustrates the nsubj tag, where the head dominates the dependent.

The actual magnitude of the dominance relationship per tag across layers is captured from the difference between mIoD and mIoH (see Figure 19). The differences are lower in the beginning and end layers, but the IoX scores (see Figure 16) are higher, indicating a broad attention pattern. In the middle layers, the differences are also low, but the IoX scores are low as well. The differences are found to be higher around layer 5 and layer 18, suggesting that these layers contribute more significantly to the dominance relationship between head and dependent, regardless of the dependency tag. In particular, around layer 18 the dependency dominance information is more captured because of lower IoX scores but large differences compared with the layers around 5.

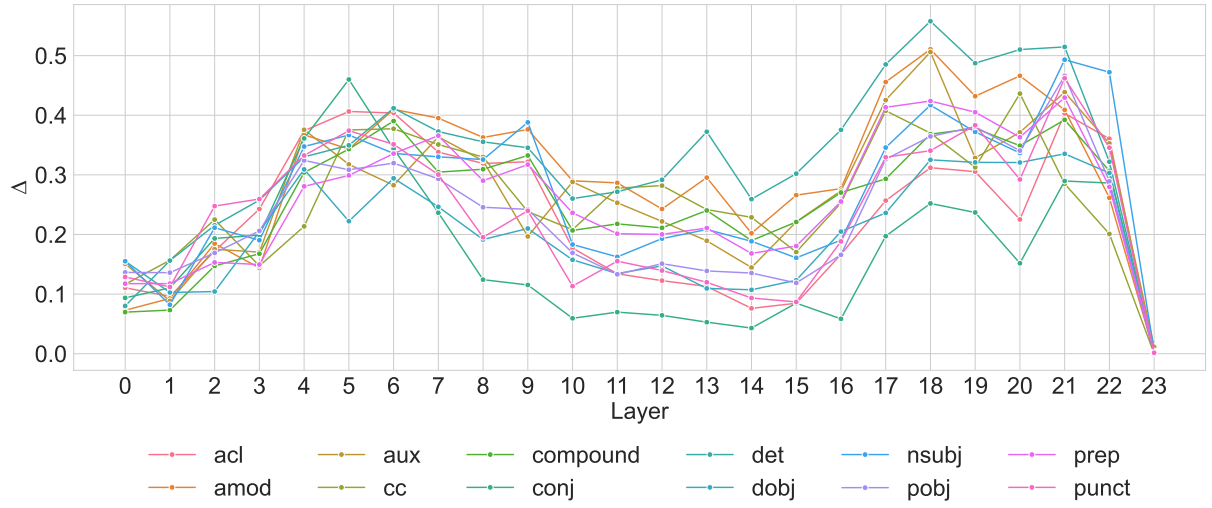


Figure 19 The difference between mIoD and mean mIoH for all tags across layers.

5.3 PPA Analysis

5.3.1 Experimental Setup

This experiment examined how the model visually interprets syntactic ambiguities arising from prepositional phrases, specifically investigating whether prepositional phrases attach to nouns or verbs, referred to as Prepositional Phrase Attachment (PPA).

An initial dataset was created by extracting captions containing prepositional phrases from the LAION (Schuhmann et al. 2021) and COCO (Lin et al. 2014) datasets using `spaCy`. The resulting dataset includes a total of 5,716 entries. Each entry follows a quadruple pattern $[v, n_1, p, n_2]$ and is annotated with an attachment label: *Verb Phrase (VP)* or *Noun Phrase (NP)*. For example, *['pouring', 'tea', 'into', 'cup']* and the attachment label *Verb Phrase (VP)* was extracted from the caption *'Someone pouring tea into a glass cup.'*.

From this dataset, 100 entries were manually selected after verifying each case, removing ambiguous examples, correcting attachment labels, and balancing the dataset with 50 noun attachments and 50 verb attachments. In total, eight prepositions were used: *at*, *for*, *from*, *in*, *into*, *of*, *on*, and *with*. Each caption was used ten times with different random seeds to augment the data, resulting in a total of 1,000 samples.

Following previous experiments, an image was generated for each entry. Attention scores were aggregated, interpolated to match the image size, binarized and Intersection-over-Union (IoU) scores were computed for n_1–n_2 and v–n_2 pairs. The predicted attachment label is determined by comparing the IoU scores: the noun n_2 is attached to the noun n_1 if the IoU score for n_1–n_2 is greater than the IoU score for v–n_2; otherwise, it is attached to the verb v (see Appendix A.3).

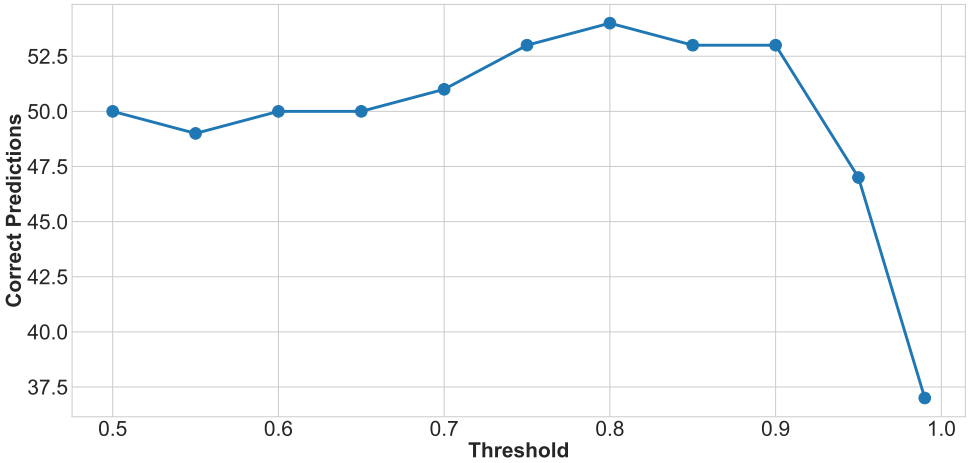


Figure 20 Number of correct attachment predictions (out of 80 samples) using quantile thresholding at different threshold values.

An additional experiment was conducted to determine the thresholding method and optimal threshold value for mask creation. A separate dataset was created by selecting ten distinct samples for each preposition, resulting in a total of 80 samples. As discussed earlier, the best result was previously reported by Tang et al. (2023) using max-based thresholding with a threshold value of 0.4 (DAAM-0.4). However, this method produced only 41 correct attachment predictions out of 80 entries. Improved results were observed using quantile-based thresholding, with the highest accuracy achieved at a threshold value of 0.8 (see Figure 20). This value was subsequently adopted for all PPA experiments.

Another dataset was created for layer-wise analysis by selecting 100 distinct captions with a fixed seed and extracting IoU scores from each layer, resulting in a total of 2,400 entries across all 24 layers.

5.3.2 Observations

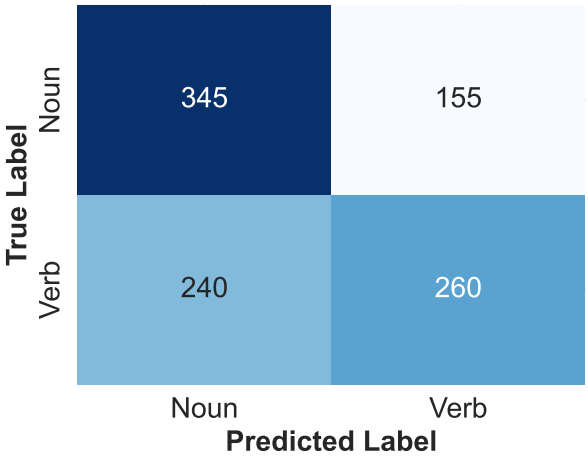


Figure 21 Confusion matrix for true and predicted attachments.

It was found that 605 of the 1000 attachments were correctly predicted. Although the mean pixel count for nouns was lower than for verbs in the POS analysis (see Section 5.1.2 and Figure 13), it was observed that a higher number of prepositional attachments were toward nouns than verbs. As illustrated in Figure 21, the true number of noun attachments was 500, but the predicted attachments increased to 585. In contrast, the total prediction for verb attachments decreased from 500 to 415.

After extending the quantitative analysis of the correct attachment predictions per preposition, it was found that the highest accuracy was achieved for the preposition *for* (87%), as shown in Figure 22. The preposition *of* is usually attached to nouns, but some entries were predicted as verb attachments. Conversely, *into* is usually attached to verbs; however, many noun attachments were observed.

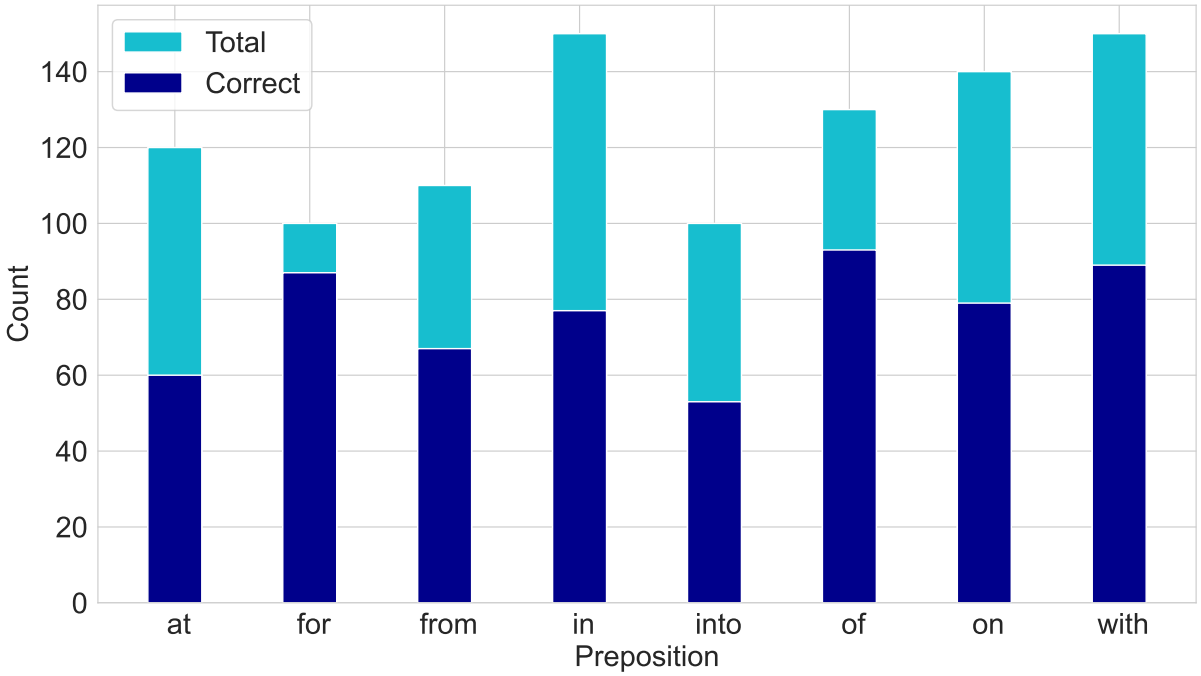


Figure 22 Total and correct predictions per preposition

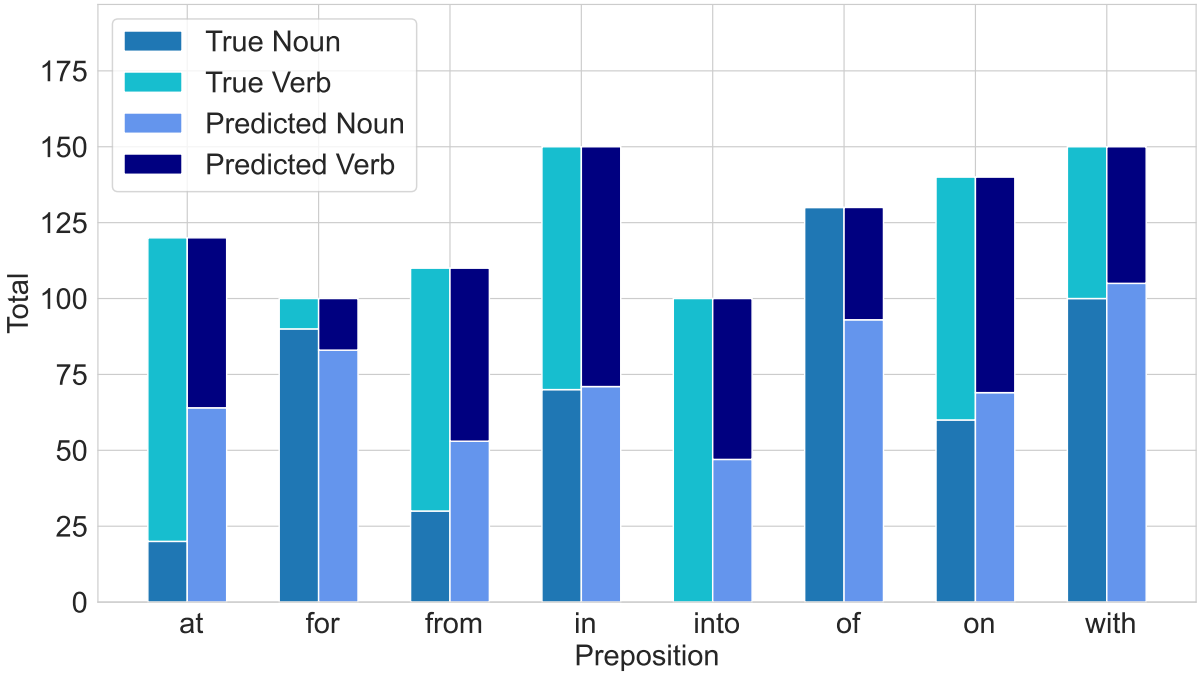


Figure 23 True and predicted number of noun and verb attachments per preposition

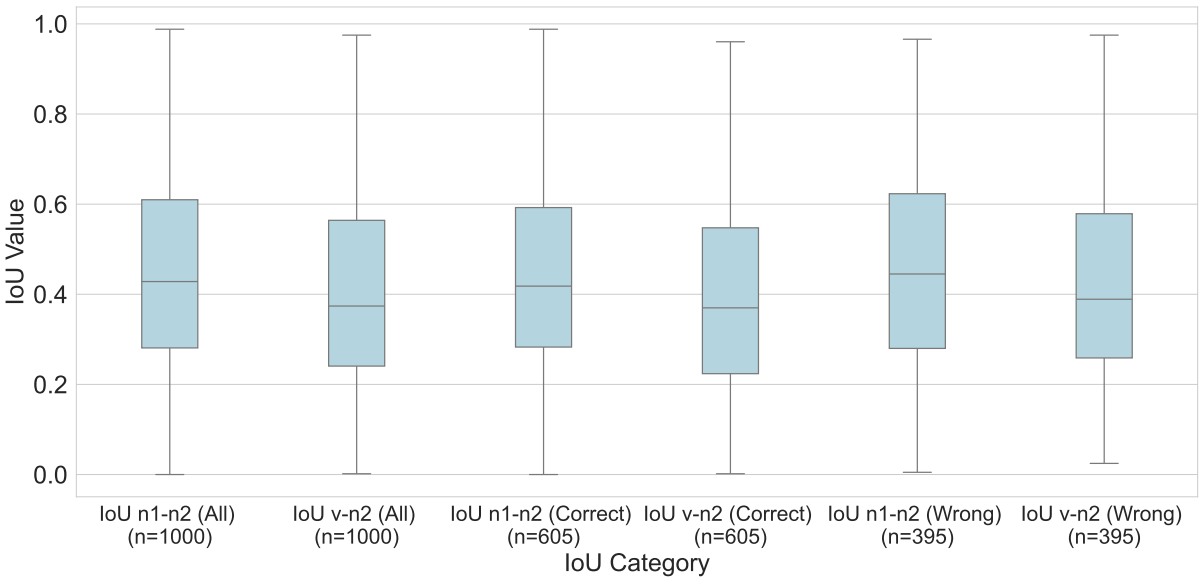


Figure 24 Distribution of IoU scores for all, correct, and incorrect attachments.

In general, the predictions for the noun attachments increased for all prepositions except for a slight decrease for the preposition *for* and *of* (see Figure 23). It can also be seen from the distribution of the IoU scores in Figure 24 that the noun attachments have slightly higher IoU scores, regardless of the predicted attachments.

Cohen's *d* (Cohen 1988, p. 19–27) was used to measure the statistical difference between the two IoU score distributions, and a small effect size of 0.2 was found. This indicates a slight difference between the IoU scores of n_1–n_2 and v–n_2, but not a substantial one. However, a large difference was observed for the prepositions *for* and *of*, despite them having the highest number of correct predictions (see Figure 25).

In terms of layer-wise analysis, the IoU scores for verb attachments are lower than those for noun attachments across all layers (see Figure 26). The scores are higher in the early and later layers; however, they are lower in the middle layers. Nevertheless, the scores show a consistent pattern with each other throughout all layers.

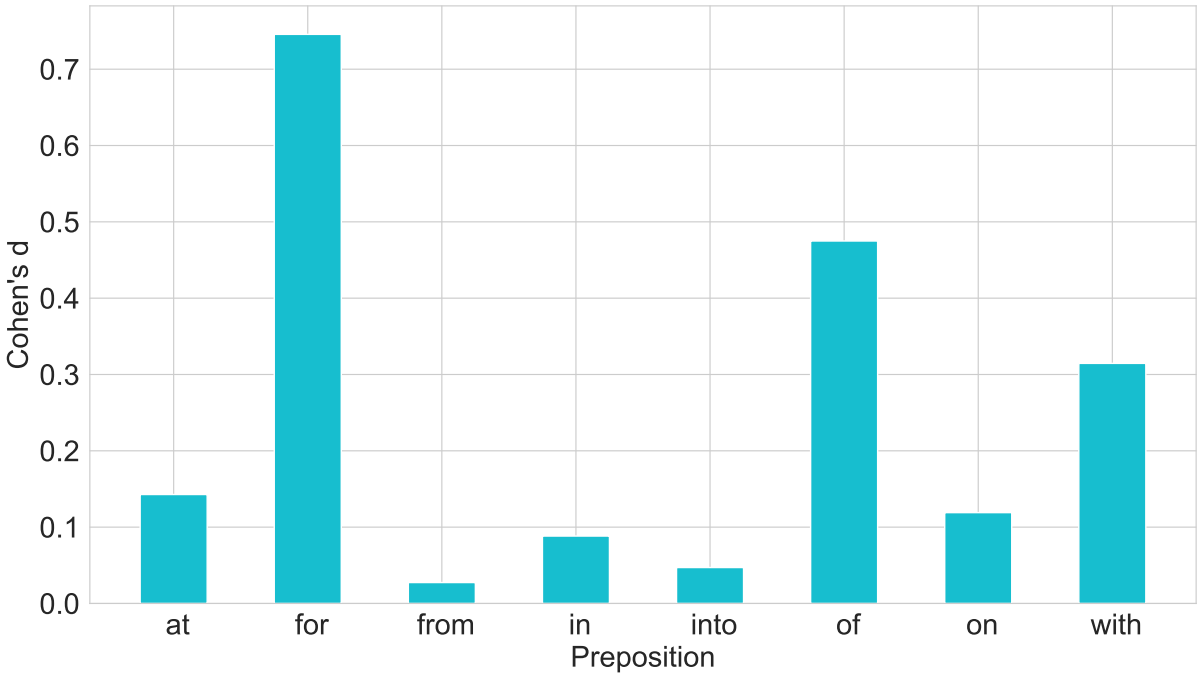


Figure 25 Cohen's d between IoU scores of n_1–n_2 and v–n_2 per Preposition.

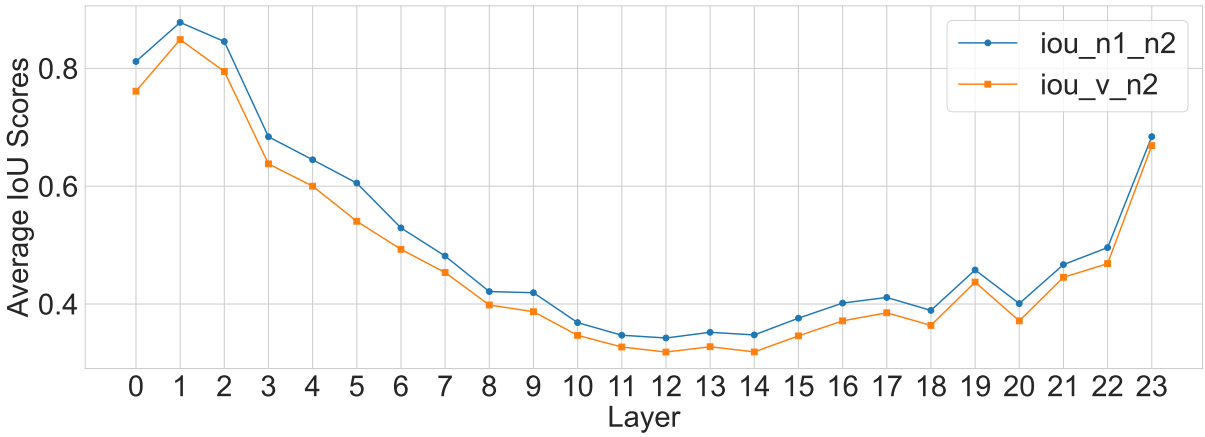


Figure 26 Average IoU per Layer.

6 Discussion

It is very difficult to determine whether the SD2 or SD3 architecture captures the linguistic structures more effectively. We are unable to find comparable work on POS and PPA analysis for the SD2 architecture (see Section 2.4). This could be a valuable direction for future research.

However, Tang et al. (2023) mentioned that they are the first to interpret large diffusion models from a visuolinguistic perspective. Since our DEP analysis is similar to the visuosyntactic analysis of Tang et al. (2023), we can compare the results to gain an approximate understanding of which architecture captures the linguistic structures better.

From dependency analysis (DEP), the mIoU scores with SD3 (see Table 1) are reduced by approximately half for almost all tags compared to the mIoU scores with SD2 reported by Tang et al. (2023) in the visuosyntactic analysis section. To better understand the dependency relationships, we can focus on mIoD, mIoH, and the differences between them instead of the mIoU scores alone. A larger difference indicates a stronger dominance relationship. A broad attention pattern is observed for the dependent word (*'is', 'the', 'a', ','*) of particular tags: *det*, *punct*, and *aux* (although *aux* is not included in the visuosyntactic analysis in the DAAM paper) in both analyses (Ours and DAAM). However, Tang et al. (2023) reports that the differences between mIoD and mIoH scores are very small for these particular tags (*punct*: 1.8 and *det*: 2.6). We found in our analysis that the differences are larger (*punct*: 29.9, *det*: 42.6, and *aux*: 30.3), which is obvious because of the spread out attention of the dependent words and concentrated attention of the head words of these tags. Additionally, it can be observed from the POS analysis in Figure 13 that the pixel counts for the POS tags (*punct*, *det*, *aux*) of the words corresponding to the dependent words of these DEP tags are considerably higher than the pixel counts for the POS tags (usually *noun* and *verb*) of words corresponding to the head words of those DEP tags. Therefore, more abstract tags usually show broad attention, while content tags show a concentrated pattern, and the differences between them, along with their dominance relationship, are expected to be larger. From this perspective, we can draw an approximate conclusion that the SD3 architecture captures linguistics better than SD2.

In general, the SD3 architecture not only joins sequences, but also applies modulation at different points to inject textual information. This leads to better alignment between text and image, generating more concentrated similarity scores in the attention matrix (see Section 3.4). Therefore, the SD3 architecture can capture much more detailed linguistic information than cross-attention-based SD2.

7 Conclusion

In conclusion, we study recent Stable Diffusion models based on rectified flow and transformer architectures from a syntactic perspective. We follow the DAAM method to generate attribution maps; however, we aggregate attention scores from the relevant region of the attention matrix.

We apply our method to examine how individual parts of speech tags are captured in the generated images and find that several tags, such as *DET*, *AUX*, and *PUNC*, attend too broadly, while others, such as *NOUN* and *VERB*, show a more concentrated pattern. We extend our analysis layer-wise and identify that layer 14 plays a critical role, exhibiting the highest mean attention values and the lowest pixel counts across nearly all tags.

We further investigate dependency relationships between the head and the dependent words and compare our findings with those reported in the DAAM paper. We find that dependents dominate heads in most cases, except for *NSUBJ* and *CONJ*, and that layers around 18 play a critical role in capturing these dominance relationships.

Finally, We analyze how the model resolves prepositional phrase ambiguities and show a slight preference for noun attachment that is consistent across layers.

Appendix

A Examples

A.1 POS Experiments



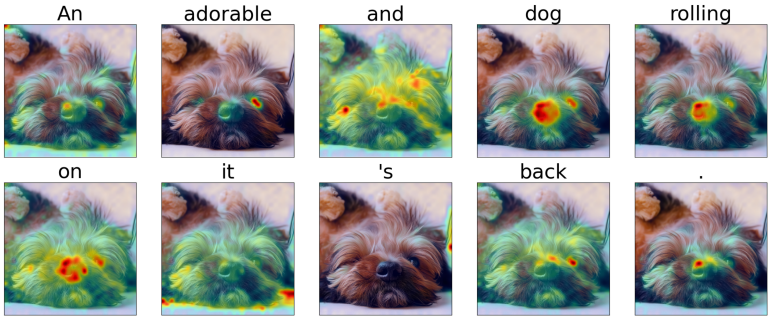
Pos Tag	Word	Pixel Count	Mean
DET	A	73193	0.038
ADJ	black	23666	0.062
NOUN	cat	11893	0.084
AUX	is	56944	0.047
VERB	sitting	62422	0.047
ADP	on	26398	0.063
CCONJ	and	126654	0.038
ADV	back	92671	0.04
PRON	its	56359	0.046
PUNCT	.	107130	0.036

Caption: A black cat is sitting on the bench and turning back its head.



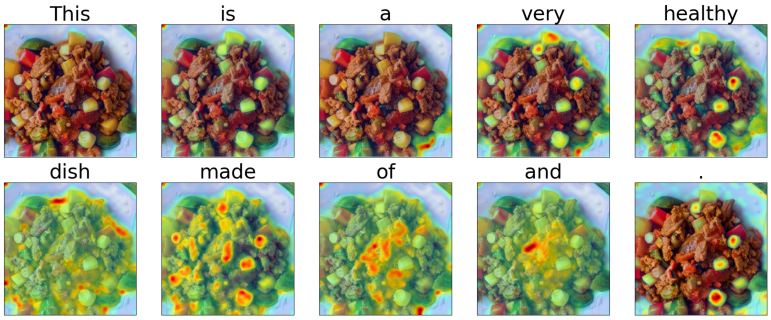
Pos Tag	Word	Pixel Count	Mean
DET	The	180111	0.031
NOUN	bus	1826	0.077
AUX	is	136653	0.034
ADV	mainly	76867	0.043
ADJ	white	86000	0.041
CCONJ	but	42690	0.048
ADP	on	54843	0.046
VERB	has	69861	0.046
PRON	it	32139	0.053
PUNCT	.	8135	0.059

Caption: The bus is mainly white but has stripes of color on it as well.



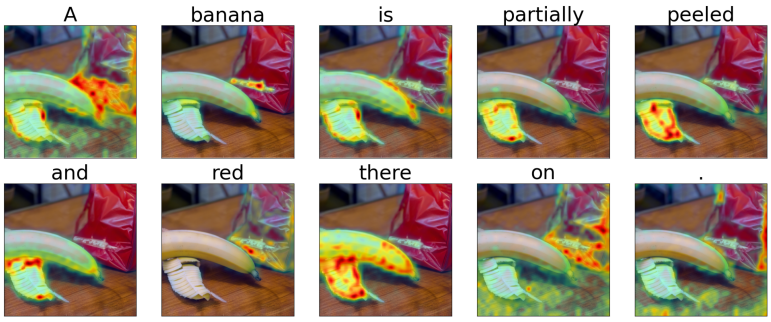
Pos Tag	Word	Pixel Count	Mean
DET	An	215850	0.029
ADJ	adorable	7222	0.066
CCONJ	and	228873	0.031
NOUN	dog	48838	0.046
VERB	rolling	66149	0.038
ADP	on	197409	0.032
PRON	it	171219	0.034
AUX	's	5797	0.062
ADV	back	178486	0.032
PUNCT	.	77791	0.037

Caption: An adorable brown and black dog rolling on it's back.



Pos Tag	Word	Pixel Count	Mean
PRON	This	67737	0.032
AUX	is	118708	0.033
DET	a	42216	0.043
ADV	very	127321	0.035
ADJ	healthy	174954	0.032
NOUN	dish	244261	0.03
VERB	made	242674	0.031
ADP	of	249240	0.03
CCONJ	and	233212	0.031
PUNCT	.	116153	0.035

Caption: This is a very healthy dish made up of meat and vegetables.



Pos Tag	Word	Pixel Count	Mean
DET	A	253490	0.028
NOUN	banana	45818	0.044
VERB	is	137861	0.034
ADV	partially	57948	0.044
VERB	peeled	68705	0.043
CCONJ	and	58430	0.048
ADJ	red	59442	0.045
PRON	there	86254	0.045
ADP	on	205229	0.031
PUNCT	.	232379	0.026

Caption: A banana is partially peeled and there is a red bag on a wooden table.



Pos Tag	Word	Pixel Count	Mean
DET	A	231211	0.03
NOUN	man	2658	0.22
CCONJ	and	17242	0.068
PRON	his	7014	0.1
AUX	are	44488	0.047
VERB	standing	4508	0.114
ADV	next	56231	0.057
ADP	to	65116	0.049
ADJ	open	30785	0.047
PUNCT	.	180813	0.032

Caption: A man and his son are standing next to an open fire hydrant.

Figure 27 Examples of attribution maps generated for all 10 POS tags, using captions randomly selected from the COCO dataset. Each caption includes at least one instance of every POS tag.

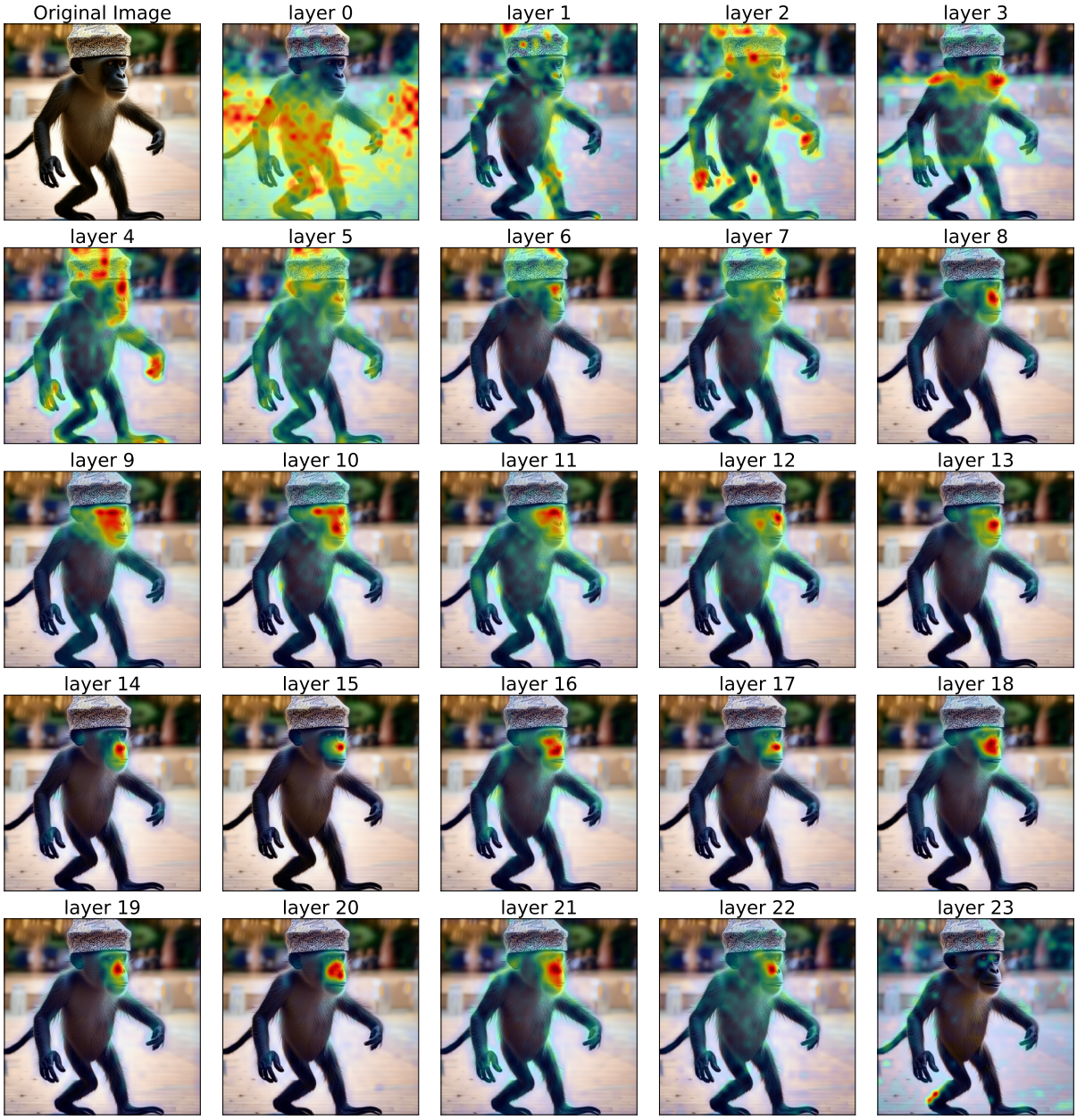


Figure 28 Image generated with the prompt *'A monkey with a hat walking.'* and the corresponding attribution maps across individual layers for the word *'monkey'*.

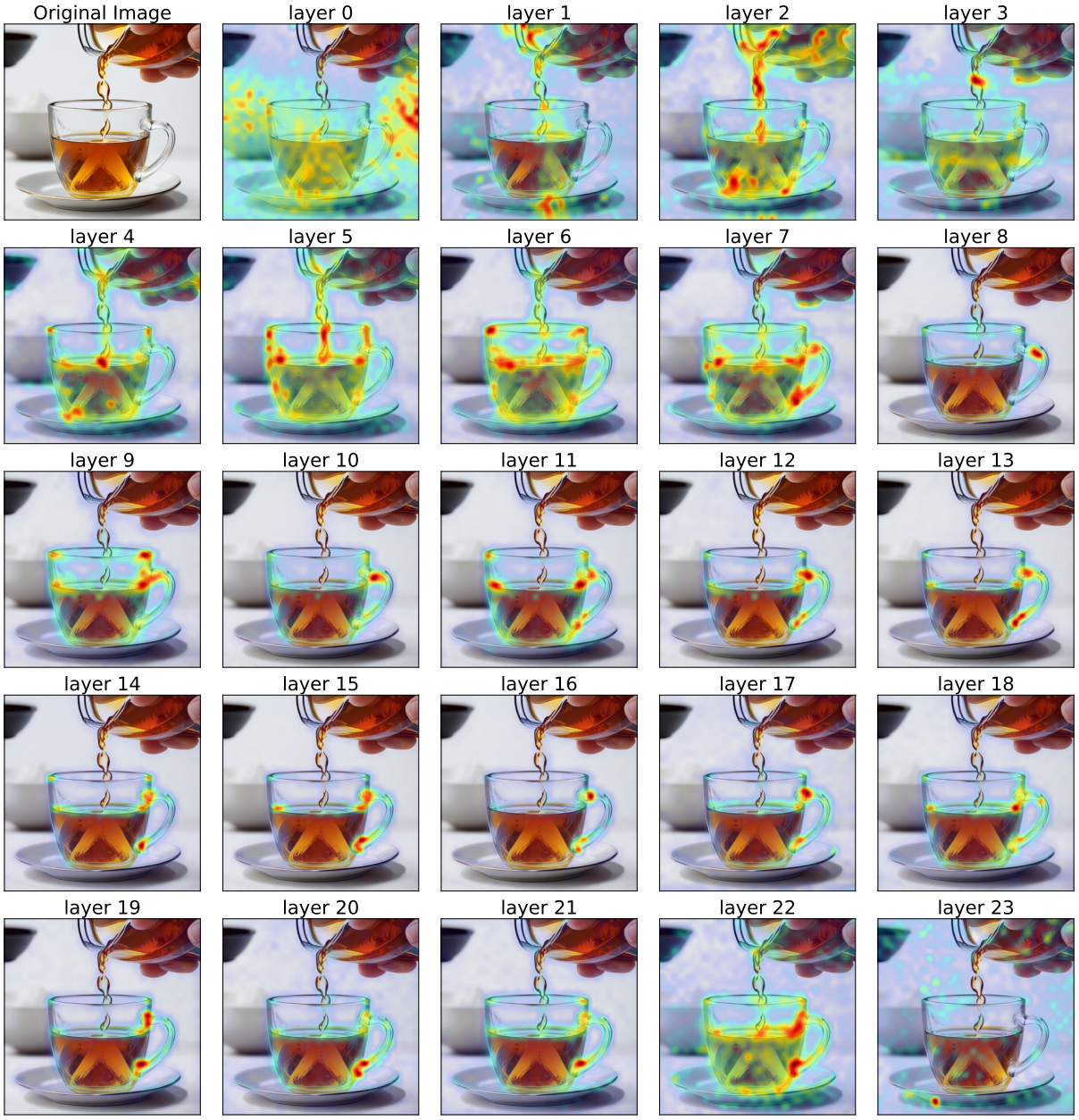
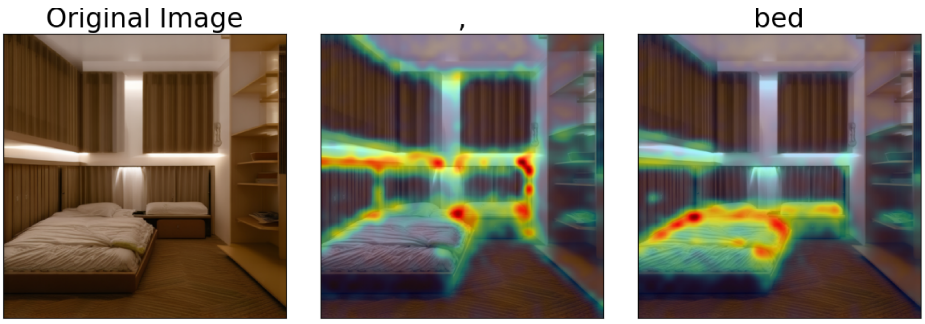


Figure 29 Image generated with the prompt *'Someone pouring tea into a glass cup.'* and the corresponding attribution maps across individual layers for the word *'cup'*.

A.2 DEP Experiments



Caption: a plane at the airport with some cars by it
Dependency tag: det, Dependent: a, Head: plane
IoU: 0.366, IoD: 0.366, IoH: 0.998, Δ: 0.631



Caption: A room with a bed, sheets and shelves.
Dependency tag: punct, Dependent: ,, Head: bed
IoU: 0.45, IoD: 0.497, IoH: 0.828, Δ: 0.331



Caption: An elephant is chewing the side of a tree.
Dependency tag: aux, Dependent: is, Head: chewing
IoU: 0.016, IoD: 0.016, IoH: 1.0, Δ: 0.984



Caption: The tablet is sitting beside the laptop computer on the desk.
Dependency tag: nsubj, Dependent: tablet, Head: sitting
IoU: 0.37, IoD: 0.805, IoH: 0.407, Δ: 0.398



Caption: a small boy with a green shirt is eating a sandwich
Dependency tag: amod, Dependent: green, Head: shirt
IoU: 0.157, IoD: 0.16, IoH: 0.906, Δ: 0.746



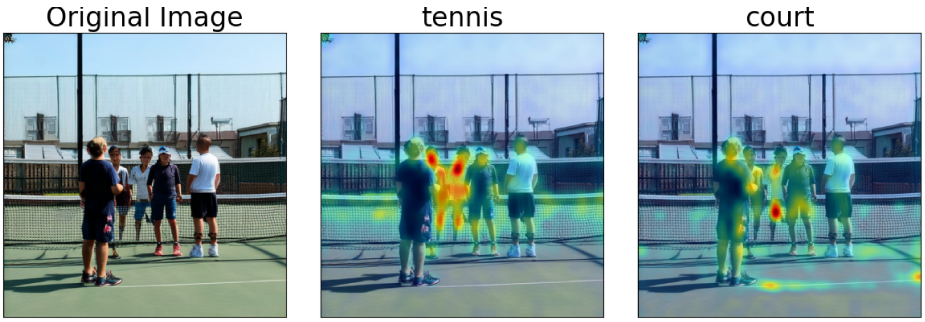
Caption: A large room filled with furniture and clothes.
Dependency tag: conj, Dependent: clothes, Head: furniture
IoU: 0.44, IoD: 0.999, IoH: 0.44, Δ: 0.559



Caption: A small cat sitting neatly in a ceramic bowl
Dependency tag: acl, Dependent: sitting, Head: cat
IoU: 0.369, IoD: 0.504, IoH: 0.578, Δ: 0.074



Caption: A small cat sitting neatly in a ceramic bowl
Dependency tag: prep, Dependent: in, Head: sitting
IoU: 0.456, IoD: 0.558, IoH: 0.713, Δ: 0.154



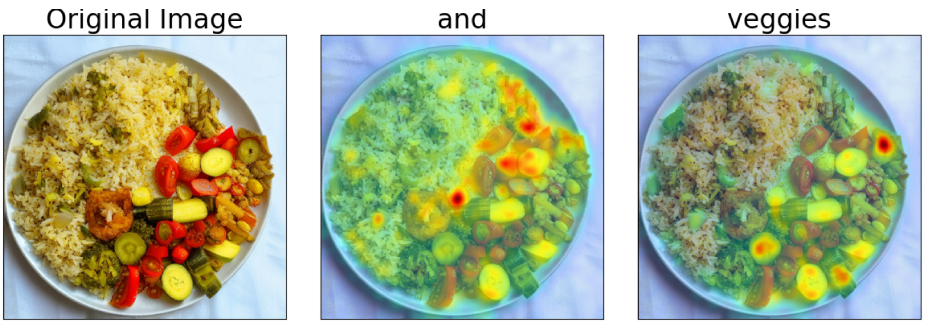
Caption: Some people standing around on a tennis court.
Dependency tag: compound, Dependent: tennis, Head: court
IoU: 0.409, IoD: 0.593, IoH: 0.569, Δ: 0.024



Caption: A three dimensional rendering of a sign on a beach.
Dependency tag: pobj, Dependent: sign, Head: of
IoU: 0.647, IoD: 0.794, IoH: 0.778, Δ: 0.016



Caption: A man in a hat riding a horse.
Dependency tag: dobj, Dependent: horse, Head: riding
IoU: 0.477, IoD: 0.671, IoH: 0.623, Δ: 0.048



Caption: A plate filled with assorted veggies and rice.
Dependency tag: cc, Dependent: and, Head: veggies
IoU: 0.603, IoD: 0.603, IoH: 1.0, Δ: 0.397

Figure 30 Example of attribution maps of head and dependent words for the top 12 tags, along with their overlap statistics.

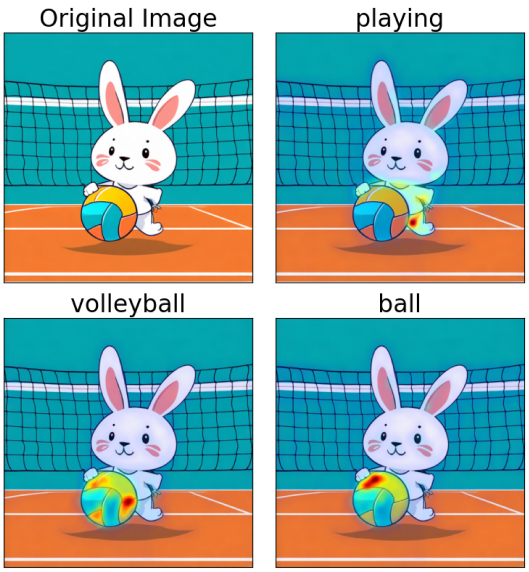
A.3 PPA Experiments



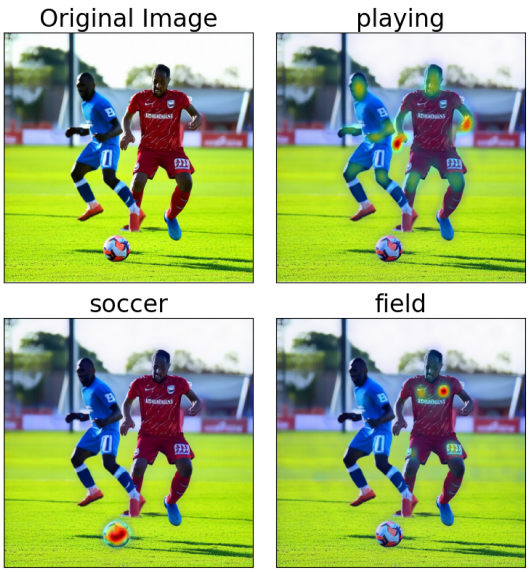
Caption: A woman wearing an orange shirt with black writing.
IoU (Noun): 0.356, IoU (Verb): 0.138, seed: 13579
True Attachment: Noun, Predicted Attachment: Noun
Preposition: with, Example Type: Good



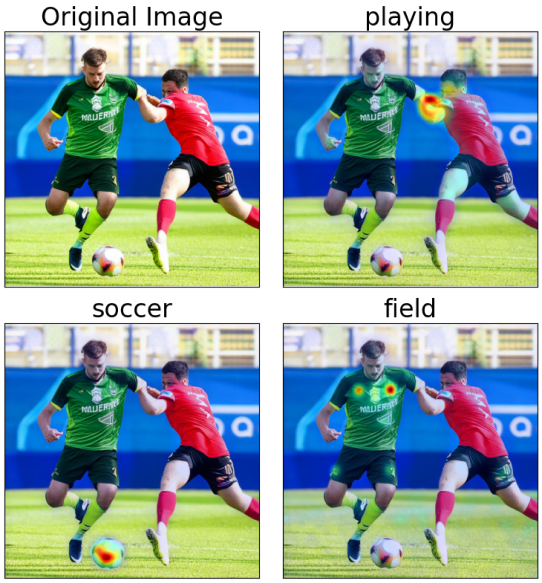
Caption: A woman celebrating her birthday with cake and confetti.
IoU (Noun): 0.442, IoU (Verb): 0.283, seed: 123456
True Attachment: Verb, Predicted Attachment: Noun
Preposition: with, Example Type: Bad



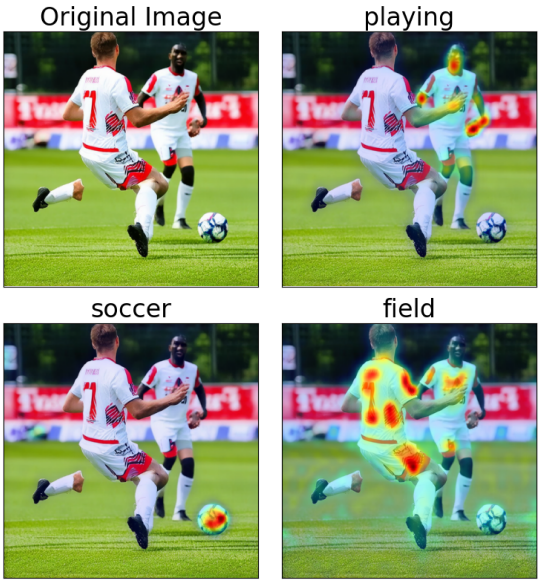
Caption: A cartoon bunny is playing volleyball with the ball.
IoU (Noun): 0.680, IoU (Verb): 0.680, seed: 13579
True Attachment: Verb, Predicted Attachment: Verb
Preposition: with, Example Type: Ambiguous



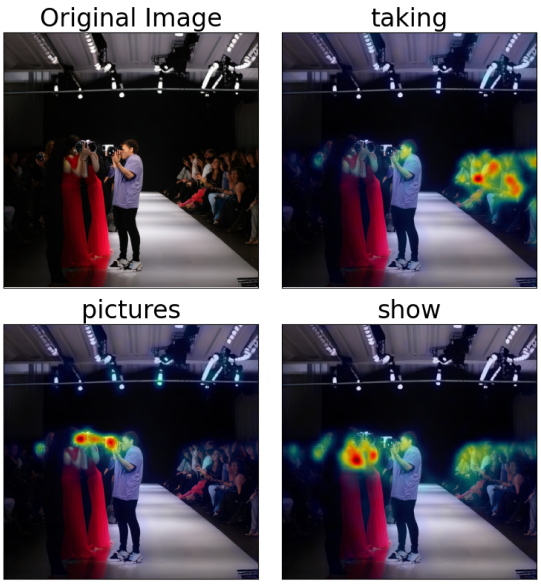
Caption: Two men are playing soccer on the field.
IoU (Noun): 0.400, IoU (Verb): 0.691, seed: 123456
True Attachment: Verb, Predicted Attachment: Verb
Preposition: on, Example Type: Good



Caption: Two men are playing soccer on the field.
IoU (Noun): 0.401, IoU (Verb): 0.297, seed: 82645
True Attachment: Verb, Predicted Attachment: Noun
Preposition: on, Example Type: Bad



Caption: Two men are playing soccer on the field.
IoU (Noun): 0.422, IoU (Verb): 0.421, seed: 654321
True Attachment: Verb, Predicted Attachment: Noun
Preposition: on, Example Type: Ambiguous



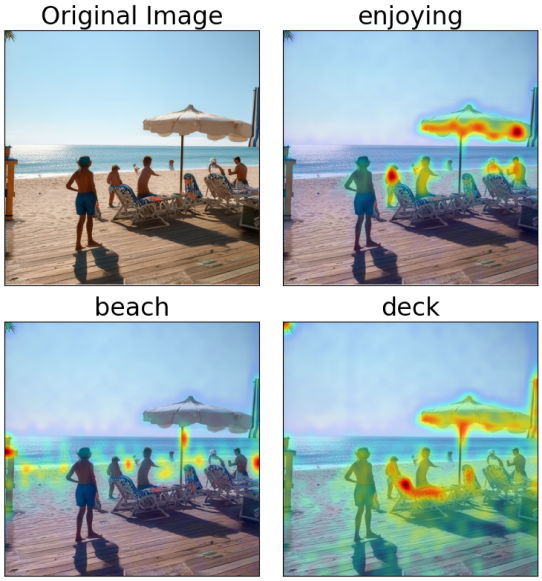
Caption: People taking pictures at the runway show.
IoU (Noun): 0.420, IoU (Verb): 0.537, seed: 192837
True Attachment: Verb, Predicted Attachment: Verb
Preposition: at, Example Type: Good



Caption: People taking pictures at the runway show.
IoU (Noun): 0.505, IoU (Verb): 0.273, seed: 123456
True Attachment: Verb, Predicted Attachment: Noun
Preposition: at, Example Type: Bad



Caption: An older man giving a speech at the UN.
IoU (Noun): 0.425, IoU (Verb): 0.424, seed: 2468
True Attachment: Verb, Predicted Attachment: Noun
Preposition: at, Example Type: Ambiguous



Caption: People are enjoying the beach from a deck.
IoU (Noun): 0.246, IoU (Verb): 0.579, seed: 123456
True Attachment: Verb, Predicted Attachment: Verb
Preposition: from, Example Type: Good



Caption: Two men sitting down drinking water from glasses.
IoU (Noun): 0.627, IoU (Verb): 0.364, seed: 97531
True Attachment: Verb, Predicted Attachment: Noun
Preposition: from, Example Type: Bad



Caption: Two men sitting down drinking water from glasses.
IoU (Noun): 0.557, IoU (Verb): 0.559, seed: 2468
True Attachment: Verb, Predicted Attachment: Verb
Preposition: from, Example Type: Ambiguous



Caption: The projection screen is displaying an award for excellence.
IoU (Noun): 0.701, IoU (Verb): 0.446, seed: 123456
True Attachment: Noun, Predicted Attachment: Noun
Preposition: for, Example Type: Good



Caption: The projection screen is displaying an award for excellence.
IoU (Noun): 0.369, IoU (Verb): 0.558, seed: 2468
True Attachment: Noun, Predicted Attachment: Verb
Preposition: for, Example Type: Bad



Caption: A baseball player is signing autographs for fans.
IoU (Noun): 0.262, IoU (Verb): 0.260, seed: 2468
True Attachment: Verb, Predicted Attachment: Noun
Preposition: for, Example Type: Ambiguous



Caption: A woman decorating a cake in a kitchen.
IoU (Noun): 0.069, IoU (Verb): 0.267, seed: 123456
True Attachment: Verb, Predicted Attachment: Verb
Preposition: in, Example Type: Good



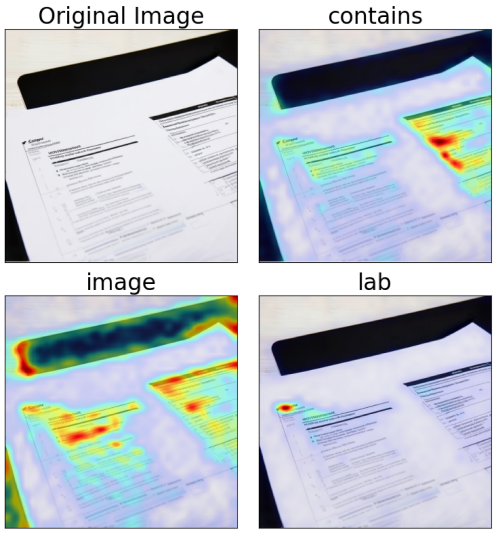
Caption: A woman crossing the street in New York City.
IoU (Noun): 0.230, IoU (Verb): 0.427, seed: 13579
True Attachment: Noun, Predicted Attachment: Verb
Preposition: in, Example Type: Bad



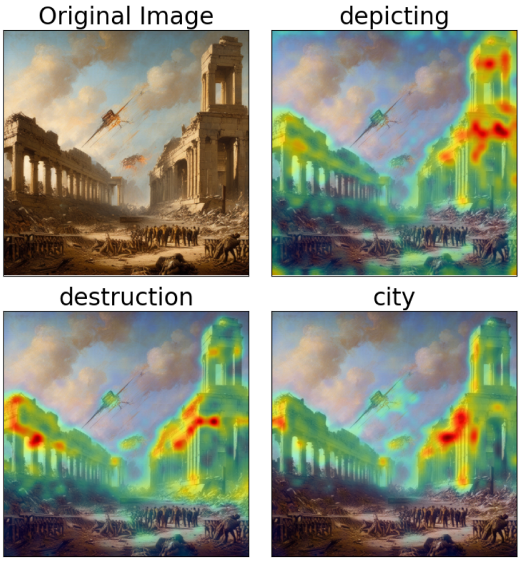
Caption: A man holding up a baby in the air.
IoU (Noun): 0.459, IoU (Verb): 0.456, seed: 82645
True Attachment: Verb, Predicted Attachment: Noun
Preposition: in, Example Type: Ambiguous



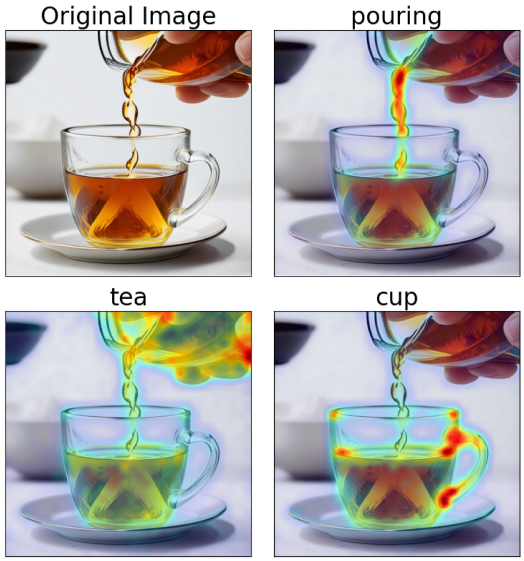
Caption: An old photo shows a group of people.
IoU (Noun): 0.647, IoU (Verb): 0.047, seed: 13579
True Attachment: Noun, Predicted Attachment: Noun
Preposition: of, Example Type: Good



Caption: The document is in full view and contains an image of a lab.
IoU (Noun): 0.510, IoU (Verb): 0.792, seed: 13579
True Attachment: Noun, Predicted Attachment: Verb
Preposition: of, Example Type: Bad



Caption: A painting depicting the destruction of an ancient city.
IoU (Noun): 0.585, IoU (Verb): 0.587, seed: 654321
True Attachment: Noun, Predicted Attachment: Verb
Preposition: of, Example Type: Ambiguous



Caption: Someone pouring tea into a glass cup.
IoU (Noun): 0.191, IoU (Verb): 0.434, seed: 2468
True Attachment: Verb, Predicted Attachment: Verb
Preposition: into, Example Type: Good

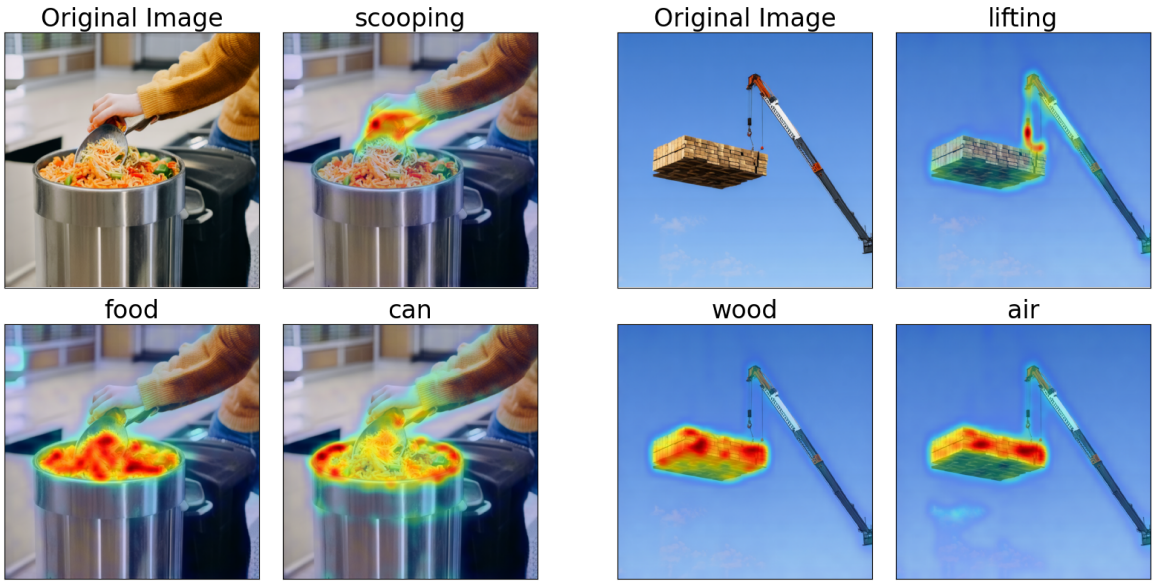


Figure 31 Good, bad, and ambiguous examples for each preposition, with corresponding overlap scores and prepositional attachment labels.

References

Abnar, S., and Zuidema, W. 2020. "Quantifying Attention Flow in Transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197.

Bahdanau, D., Cho, K., and Bengio, Y. 2015. "Neural Machine Translation by Jointly Learning to Align and Translate," in *3rd International Conference on Learning Representations (ICLR)*.

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. 2023. "All Are Worth Words: A Vit Backbone for Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22,669–22,679.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., and Ramesh, A. 2023. "Improving Image Generation with Better Captions," https://cdn.openai.com/papers/dall-e-3.pdf.

Beňová, I., Košecká, J., Gregor, M., Tamajka, M., Veselý, M., and Šimko, M. 2025. "Beyond Image-Text Matching: Verb Understanding in Multimodal Transformers Using Guided Masking," in *SOFSEM 2025: 50th International Conference on Current Trends in Theory and Practice of Computer Science*.

Chefer, H., Gur, S., and Wolf, L. 2021a. "Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 397–406.

Chefer, H., Gur, S., and Wolf, L. 2021b. "Transformer Interpretability Beyond Attention Visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791.

Chefer, H., Lang, O., Geva, M., Polosukhin, V., Shocher, A., Irani, M., Mosseri, I., and Wolf, L. 2024. "The Hidden Language of Diffusion Models," in *Proceedings of the Twelfth International Conference on Learning Representations*.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. 2018. "Neural Ordinary Differential Equations," in *Advances in Neural Information Processing Systems*, vol. 31.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. 2023. "Reproducible Scaling Laws for Contrastive Language-Image Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. 2019. “What Does BERT Look at? An Analysis of BERT’s Attention,” in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, p. 276–286.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, 2nd ed.

Dhariwal, P., and Nichol, A. 2021. “Diffusion Models Beat GANs on Image Synthesis,” in *Advances in Neural Information Processing Systems*, vol. 34.

Dinh, L., Krueger, D., and Bengio, Y. 2015. “NICE: Non-Linear Independent Components Estimation,” in *International Conference on Learning Representations*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. 2017. “Density Estimation using Real NVP,” in *International Conference on Learning Representations*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. 2021. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Elsharif, W., Alzubaidi, M., She, J., and Agus, M. 2025. “Visualizing Ambiguity: Analyzing Linguistic Ambiguity Resolution in Text-to-Image Models,” *Computers* (14:1). Article 19.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., and Rombach, R. 2024. “Scaling Rectified Flow Transformers for High-Resolution Image Synthesis,” in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, vol. 235, pp. 12,606–12,633.

Hao, Y., Dong, L., Wei, F., and Xu, K. 2021. “Self-Attention Attribution: Interpreting Information Interactions Inside Transformer,” in *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, pp. 12,963–12,971.

Helbling, A., Meral, T. H. S., Hoover, B., Yanardag, P., and Chau, D. H. P. 2025. “ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features,” in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Hendricks, L. A., and Nematzadeh, A. 2021. “Probing Image–Language Transformers for Verb Understanding,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3635–3644.

Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. 2019. "Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design," in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, vol. 97, pp. 2722–2730.

Ho, J., Jain, A., and Abbeel, P. 2020. "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems*, vol. 33.

Jain, S., and Wallace, B. C. 2019. "Attention is not Explanation," 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)*, pp. 3543–3556.

Karras, T., Aittala, M., Aila, T., and Laine, S. 2022. "Elucidating the Design Space of Diffusion-Based Generative Models," in *Advances in Neural Information Processing Systems*, vol. 35.

Kim, D., Thomas, X., and Ghadiyaram, D. 2025. "Revelio: Interpreting and Leveraging Semantic Information in Diffusion Models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, accepted for publication.

Kingma, D. P., and Dhariwal, P. 2018. "Glow: Generative Flow with Invertible 1×1 Convolutions," in *Advances in Neural Information Processing Systems*.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. 2014. "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. 2023. "Flow Matching for Generative Modeling," in *International Conference on Learning Representations*.

Liu, X., Gong, C., and Liu, Q. 2023. "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow," in *International Conference on Learning Representations*.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. 2022. "DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps," in *Advances in Neural Information Processing Systems*, vol. 35.

Mehrabi, N., Goyal, P., Verma, A., Dhamala, J., Kumar, V., Hu, Q., Chang, K.-W., Zemel, R., Galstyan, A., and Gupta, R. 2023. "Resolving Ambiguities in Text-to-Image Generative Models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14,367–14,388.

Nakashole, N., and Mitchell, T. M. 2015. "A Knowledge-Intensive Model for Prepositional Phrase Attachment," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 365–375.

Nichol, A., and Dhariwal, P. 2021. "Improved Denoising Diffusion Probabilistic Models," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, vol. 139, pp. 8162–8171.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., and Chen, M. 2022. "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, vol. 162 of *Proceedings of Machine Learning Research*, pp. 16,784–16,804.

Park, J.-H., Ju, Y.-J., and Lee, S.-W. 2024. "Explaining Generative Diffusion Models via Visual Analysis for Interpretable Decision-Making Process," *Expert Systems with Applications* (248), p. 123,231.

Peebles, W., and Xie, S. 2023. "Scalable Diffusion Models with Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. 2021. "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, vol. 139, pp. 8748–8763.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research* (21:140), pp. 1–67.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. 2022. "Hierarchical Text-Conditional Image Generation with CLIP Latents," https://arxiv.org/abs/2204.06125.

Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., and Chechik, G. 2023. "Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment," in *Advances in Neural Information Processing Systems*, vol. 36.

Rezende, D. J., and Mohamed, S. 2015. "Variational Inference with Normalizing Flows," in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, vol. 37, pp. 1530–1538.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 2022. "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10,684–10,695.

Ronneberger, O., Fischer, P., and Brox, T. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer, vol. 9351 of *Lecture Notes in Computer Science*, pp. 234–241.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Lopes, R. G., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. 2022. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," in *Advances in Neural Information Processing Systems*, vol. 35.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. 2021. "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs," .

Serrano, S., and Smith, N. A. 2019. "Is Attention Interpretable?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951.

Shahariar, G. M., Chen, J., Li, J., and Dong, Y. 2024. "Adversarial Attacks on Parts of Speech: An Empirical Study in Text-to-Image Generation," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12,874–12,890.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. 2015. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, pp. 2256–2265.

Song, J., Meng, C., and Ermon, S. 2021a. "Denoising Diffusion Implicit Models," in *International Conference on Learning Representations*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. 2021b. "Score-Based Generative Modeling through Stochastic Differential Equations," in *International Conference on Learning Representations*.

Stability AI 2022. "Stable Diffusion 2 Base," https://huggingface.co/stabilityai/stable-diffusion-2-base, accessed: 2025-09-10.

Stability AI 2024. "Stable Diffusion 3.5 Medium," https://huggingface.co/stabilityai/stable-diffusion-3.5-medium, accessed: 2025-08-10.

Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., and Ture, F. 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)*, pp. 5644–5659.

Tenney, I., Das, D., and Pavlick, E. 2019. "BERT Rediscovers the Classical NLP Pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601.

Toker, M., Orgad, H., Ventura, M., Arad, D., and Belinkov, Y. 2024. "Diffusion Lens: Interpreting Text Encoders in Text-to-Image Pipelines," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9713–9728.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. 2017. “Attention Is All You Need,” in *Advances in Neural Information Processing Systems*, vol. 30.

Vig, J. 2019. “A Multiscale Visualization of Attention in the Transformer Model,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 37–42.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., and Wolf, T. 2022. “Diffusers: State-of-the-art diffusion models,” https://github.com/huggingface/diffusers.

Wang, X., Kang, Z., Zhai, W., Lou, X., Lai, Y., Wang, Z., Wang, Y., Huang, K., Wang, Y., Li, P., and Liu, Y. 2025. “MUCAR: Benchmarking Multilingual Cross-Modal Ambiguity Resolution for Multimodal Large Language Models,” https://arxiv.org/abs/2506.17046.

Wiegreffe, S., and Pinter, Y. 2019. “Attention is not not Explanation,” 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)*, pp. 11–20.

Yeh, C., Chen, Y., Wu, A., Chen, C., Viégas, F., and Wattenberg, M. 2024. “AttentionViz: A Global View of Transformer Attention,” *IEEE Transactions on Visualization and Computer Graphics* (30:1), pp. 262–272.

Zhang, B., and Sennrich, R. 2019. “Root Mean Square Layer Normalization,” in *Advances in Neural Information Processing Systems*.

Zhang, Q., and Chen, Y. 2023. “Fast Sampling of Diffusion Models with Exponential Integrator,” in *International Conference on Learning Representations*.

Declaration of Originality

I, Syed Ikhtiar Ahmed, student registration number: 23039053, hereby confirm that I completed the submitted work independently and without the unauthorized assistance of third parties and without the use of undisclosed and, in particular, unauthorized aids. This work has not been previously submitted in its current form or in a similar form to any other examination authorities and has not been accepted as part of an examination by any other examination authority.

Where the wording has been taken from other people's work or ideas, this has been properly acknowledged and referenced. This also applies to drawings, sketches, diagrams and sources from the Internet.

In particular, I am aware that the use of artificial intelligence is forbidden unless its use an aid has been expressly permitted by the examiner. This applies in particular to such programs that could process the task of the examination or parts thereof instead of me, for example chatbots (especially ChatGPT) and other methods suitable for this purpose.

Furthermore, I am aware that working with others in one room or by means of social or other media represents the unauthorized assistance of third parties within the above meaning, if group work is not expressly permitted. Each exchange of information with others during the examination, with the exception of examiners, persons authorized by them and invigilators, about the structure or contents of the examination or any other information such as sources is not permitted. The same applies to attempts to do so. Proofreading of the work by third parties with the aim of correcting linguistic deficiencies is permitted.

Any infringements of the above rules constitute fraud or attempted fraud and shall lead to the examination being graded "fail" ("nicht bestanden"). In the event of multiple or particularly drastic breaches of the rules, further sanctions may be imposed by the Examination Board (Prüfungsausschuss).

Place, Date

Signature