Hallucination Detection and Mitigation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Modern AI systems are highly capable of generating content across various domains. However, a persistent challenge faced by these systems is **hallucination**—the generation of outputs that are plausible but incorrect. In this report, we present a detailed analysis of hallucination, their types and causes. The report specifically focuses on **Text-to-Image** (T2I) models. We also present a method to detect hallucinations in the image generated by a T2I model and two possible strategies to mitigate hallucinations.

1 Introduction

We live in a world that is filled with AI systems everywhere, from Self-Driving cars till Generative Models. Generative Models are set to be an essential part of lives of human beings soon with the power that they possess in terms of generating content in the right domain that the user needs it at the very moment.

However, many generative models face a very common issue **Hallucination**. As per [1], Hallucination can be defined as "*the output of the Neural Machine Translation (NMT) system is often quite fluent but entirely unrelated to the input*". Although Hallucination is very common in most of the Generative models, we tend to focus here more on Text to Image Generative Models (T2I). Hallucination for a T2I model as per [2] is: "*Text-to-image generation has shown remarkable progress with the emergence of diffusion models. However, these models often generate factually inconsistent images, failing to accurately reflect the factual information and common sense conveyed by the input text prompts*".

The current T2I models heavily depend on the stochastic sampling process, that acts as a big source of hallucinations to trip into the pipeline. As per studies [3], hallucinations of T2I model could be reduced by 95% just by having variance based metrics during sampling process. Thus, we propose a possible strategy to ensure that the noise added to T2I models (Diffusion Models) is not very random, but something that actually is useful enough for both diversification and hallucination reduction. We also propose another method to use a modified loss function to fine tune T2I Diffusion Models that can reduce hallucinations.

The complete implementation is available on this **GitHub**.

2 Literature Review

Recent research on hallucination detection in text-to-image (T2I) models has developed structured evaluation frameworks. The Fine-Grained Hallucination (FiG) Metrics 6 employ a three-stage hierarchical approach to compare generated images with prompts, using object intersection scores, dependency parsing for attribute pairs, and spurious object subtraction. While effective for granular analysis, it struggles with compositional prompts. The I-HallA v1.0 Pipeline 7 leverages GPT-4o to generate QA pairs from textbook-based prompts, achieving 72.3% detection accuracy via vision-language models but faces limitations in LLM dependency. Scene Graph-Based QA Agents 8 parse prompts into (subject-relation-object) triples, validate against image-derived knowledge graphs, and classify hallucinations through LLM-generated questions, though graph construction adds computational complexity. Dependency Parsing 9 converts prompts into grammatical triples for systematic omission/addition detection. Current methods prioritize structured evaluation but lack scalability for complex prompts. Our work addresses these gaps through lightweight attention regularization and noise conditioning, building on insights from cross-attention analysis.

3 Problem Statement

Text-to-Image (T2I) generative models often produce visually compelling images but struggle with hallucination, where generated outputs fail to accurately reflect the input text. We wanted to suggest a novel approach that can reduce hallucination (sticking to reducing spurious additions) without actual need of human intervention (means no question answering etc). Most importantly we want to detect the hallucinating region in an image generated from TTI model and then also mitigating it by quantifying hallucination and coming up with a new loss function.

4 Categorizing Hallucinations

Hallucinations by a T2I model are generally of following categories [4]:

- Attribute Hallucination: Occurs when an unspecified attribute of an object becomes part of the image. For example, prompting **red apple** generated **green apple** instead.
- Relation Hallucination: Occurs when the relation between objects specified in the prompt is assumed by the model, or distorts/ignores the object relations specified in the prompt. It mainly consists of spatial relations. For example, prompting cat on the table generated cat under the table.
- Object Hallucination: Occurs when objects that were not part of the prompt appear/ objects mentioned in the prompt don't appear in the generated image. For example, prompting Issac Newton generated Issac Netwon under an apple tree.

From hallucination detection point of view, here are the categories that we assume:

- **Prompt Omission**: The T2I model ignores important non-redundant tokens from the prompt during generation.
- **Spurious Addition**: The model introduces objects or elements that were not specified in the prompt.

5 Causes of Hallucinations

Hallucinations in text-to-image (T2I) models are often a result of the interplay between model architecture and the training process. The causes can be traced back to how the models generate images based on input prompts, and how they deal with noise and ambiguity in the input data. In this section, we categorize T2I models into two main types: Generative Adversarial Networks (GANs) and Diffusion Models (DMs), with a particular focus on their inherent susceptibility to hallucinations.

5.1 GAN Models

Generative Adversarial Networks (GANs) are inspired by the working of the human brain, with two primary components: a generator that creates samples, and a discriminator that evaluates them. GANs are trained through a game-theoretic approach where the generator attempts to produce realistic images while the discriminator strives to differentiate real images from the generated ones.

Despite their promise, GANs can be prone to hallucinations. These hallucinations arise due to several factors in their design and training process.

5.1.1 Human Brain Analogy

The analogy to the human brain is relevant here. Just as the brain has a sophisticated mechanism for distinguishing real sensory input from imagined or hallucinated stimuli, GANs aim to simulate this with their generator and discriminator structure. However, the human brain has an advantage — it is trained on a vast amount of data starting from infancy, allowing it to learn detailed, nuanced representations of the world. In contrast, GANs are often trained on much smaller datasets and in highly controlled environments, leading to several challenges:

5.1.2 Challenges in GAN Training

- **Limited Data Coverage:** GANs, when trained on insufficient datasets, often lack samples from all parts of the true data distribution. This means that the generator may not be exposed to certain modes of data, such as specific variations in color, texture, or object forms. The discriminator, in turn, cannot effectively distinguish between real and fake data in these less-represented modes, causing the generator to produce unrealistic or hallucinated results when attempting to sample from these unrepresented modes.
- **Distribution Troughs:** The data distribution, especially in complex domains like images, often contains "troughs" — areas where data is sparse or missing. If the generator is tasked with creating an image from a region of the distribution where data is insufficient or underrepresented, the gradients that would typically guide the learning process become very small. This can lead to poorly trained weights, resulting in the generator producing incomplete or incoherent images that are considered hallucinations. This issue becomes more pronounced when the data does not fully capture all possible variations.

In conclusion, GANs hallucinate due to data limitations and the difficulty of representing all possible data modes with a finite dataset. This is particularly evident in real-world scenarios where the GAN's training data is often incomplete or biased.

5.2 Diffusion Models

Diffusion models (DMs) represent a different approach to generative modeling. These models work by simulating a forward process that gradually adds random noise to clean images across multiple time steps, resulting in highly corrupted representations. In the reverse process, the model tries to reconstruct the original image by progressively removing noise, guided solely by a given text prompt.

While diffusion models show great potential, they are not immune to hallucinations. Hallucinations in diffusion models often occur due to issues related to the noise process and the dependence on the text prompt for reconstruction.

5.2.1 The Role of Random Noise in Diffusion Models

The core strength of diffusion models is their ability to generate images from noise, and this is where their susceptibility to hallucinations lies. At inference time, diffusion models begin from pure random noise. The text prompt is then used to guide the reconstruction process, essentially acting as a blueprint for how the model should remove the noise and revert to a coherent image. However, the challenge arises when the input prompt is weak, underspecified, or ambiguous.

- **Weak or Ambiguous Prompts:** If the prompt provided is not clear or detailed enough, the model may rely on prior learned biases from the training data to fill in the gaps. These biases are often not directly aligned with the intended image, leading to hallucinations. For instance, in cases where the prompt is vague or does not specify enough detail about a particular region of the image, the model might generate random, ungrounded details in those regions, resulting in incoherent or irrelevant portions of the image.
- **Random Noise Sampling Issues:** The randomness inherent in the noise process is a double-edged sword. While noise serves as a tool for generating diverse outputs, poor noise sampling or bias in the noise generation process can exacerbate hallucinations. For example, if the noise process is not properly calibrated, it may lead to overly random or corrupted image generation, making it difficult for the model to recover a meaningful image from the

noise, especially when guided by weak prompts. The quality of the Gaussian noise used during training and inference directly impacts the fidelity of the generated images.

- **Dependency on Inference Process:** The reverse process in diffusion models, which is responsible for denoising the corrupted images, heavily relies on the noise added during the forward process. If the noise distribution is biased or poorly designed, it becomes harder for the model to reliably recover a coherent image. The model's denoising mechanism might end up relying on flawed assumptions or prior learned patterns that are not consistent with the actual image, again leading to hallucinations.

Thus, hallucinations in diffusion models are often a result of weak prompts and the challenge of controlling the randomness of the noise process. Ensuring that the noise process is of high quality and that the model is provided with clear, detailed prompts is essential to reduce hallucination risk.

5.2.2 Conclusion: Balancing Randomness and Control

Both GANs and diffusion models face unique challenges related to their intrinsic reliance on noise and randomness during the generative process. For GANs, limited data and insufficient training in certain regions of the distribution lead to hallucinations, while for diffusion models, weak prompts and noise sampling issues cause the model to generate incoherent or irrelevant content. Understanding these factors is critical for mitigating hallucinations and improving the reliability and accuracy of T2I models.

6 Methods to Detect Hallucinations

In text-to-image diffusion models, the cross-attention mechanism is critical for establishing the connection between the input text prompt and the spatial regions in the generated image. Our proposed method leverages this mechanism to identify and flag hallucinated regions—those parts of the image that are not sufficiently grounded in the given prompt. In the following subsections, we describe the overall approach, its mathematical formulation, and practical integration into the diffusion pipeline.

6.1 Conceptual Overview

During the diffusion process, each attention layer in the generative model computes a set of weights that quantify how much each token in the text influences various spatial locations of the latent representation. These weights can be viewed as a function

$$\mathcal{A}^{(l)} : (x, y, t) \to \mathbb{R},$$

where l indexes the layer, (x, y) are the spatial coordinates in the latent feature map, and t denotes a textual token. Intuitively, a high weight indicates strong grounding of that particular image region to a specific part of the prompt.

6.2 Aggregation Across Layers and Tokens

To build a comprehensive indicator of grounding, the attention maps are aggregated across the different layers of the model. Let L be the number of cross-attention layers. For each layer l, denote by

$$\mathcal{A}^{(l)}(x, y, t) \in \mathbb{R}^{H \times W \times T}$$

the attention map produced during that diffusion step, where H and W represent the spatial resolution of the latent space and T is the number of tokens in the prompt.

We first compute the average across layers:

$$\bar{\mathcal{A}}(x, y, t) = \frac{1}{L} \sum_{l=1}^{L} \mathcal{A}^{(l)}(x, y, t).$$

Subsequently, the model aggregates the token influence by averaging over the token dimension:

$$\mathcal{H}(x, y) = \frac{1}{T} \sum_{t=1}^{T} \bar{\mathcal{A}}(x, y, t).$$

The resulting map $\mathcal{H}(x, y)$ reflects the overall influence of the input text across the spatial domain of the latent representation.

6.3 Normalization and Visual Representation

For interpretability, the spatial attention map is normalized to yield a heatmap that directly correlates with the level of textual grounding. The normalization is performed as follows:

$$\tilde{\mathcal{H}}(x, y) = \frac{\mathcal{H}(x, y) - \min_{(x,y)} \mathcal{H}(x, y)}{\max_{(x,y)} \mathcal{H}(x, y) - \min_{(x,y)} \mathcal{H}(x, y) + \epsilon},$$

where ϵ is a small constant to prevent division by zero. This normalized map is then scaled to an 8-bit intensity range and converted into a three-channel RGB image to form a visual heatmap.

To overlay the heatmap onto the generated image, a linear blending operation is performed:

$$I_{\text{blend}}(x, y) = \alpha \, I_{\text{image}}(x, y) + (1 - \alpha) \, I_{\text{heatmap}}(x, y),$$

where $I_{\text{image}}(x, y)$ is the original generated image, $I_{\text{heatmap}}(x, y)$ is the heatmap, and α controls the blending ratio. This composite image allows for a clear visual inspection of the regions that are well grounded versus those that are less influenced by the prompt.

6.4 Detecting Hallucinations

The core assumption behind our detection mechanism is that well-grounded regions will exhibit high normalized attention values, while regions with low values indicate insufficient grounding and potential hallucination. Specifically:

- Regions where $\tilde{\mathcal{H}}(x, y)$ is close to 1 indicate that the area receives significant influence from the text tokens.
- Conversely, regions with $\tilde{\mathcal{H}}(x, y) \ll 1$ suggest that those spatial locations were weakly attended to by the prompt and are therefore potential sites of hallucination.

Thus, by analyzing the heatmap overlay, one can quantitatively and qualitatively assess the fidelity of the generated image with respect to the input prompt.

6.5 Integration with the Diffusion Process

This diagnostic framework is integrated into the diffusion pipeline as a non-invasive, post-generation analysis tool. The procedure can be summarized in three main steps:

1. **Interception:** During image generation, the cross-attention weights from multiple layers are stored.
2. **Aggregation:** After the diffusion process completes, the collected attention maps are averaged first over layers and then over tokens to form a spatial grounding map $\mathcal{H}(x, y)$.
3. **Visualization:** The map is normalized into a heatmap $\tilde{\mathcal{H}}(x, y)$, which is then blended with the generated image to provide a visual representation of text-to-image alignment.

This framework not only facilitates the detection of hallucinated regions but also aids in evaluating the effectiveness of other mitigation strategies.

7 Proposed Methods to mitigate Hallucinations

To address the issue of hallucination in diffusion models, we propose two complementary strategies aimed at improving grounding between the generated image and the input text prompt.

7.1 Hallucination Loss via Cross-Attention Maps

Diffusion models leverage cross-attention mechanisms to align textual tokens with corresponding image regions during denoising. However, hallucinations can arise when certain image regions are not adequately attended to by the prompt.

To mitigate this, we propose a hallucination loss that explicitly penalizes activations in poorly grounded regions. Specifically, during training, we extract cross-attention maps from intermediate transformer layers. These maps, after normalization and aggregation over tokens, provide a per-pixel grounding score indicating how strongly each region is linked to the input text. Regions with low attention scores are treated as hallucinated areas.

We then define a hallucination mask and compute an auxiliary loss by applying this mask to the model's predicted features at each timestep. The loss encourages the model to suppress ungrounded activations, thereby promoting better semantic fidelity to the prompt. The total training loss becomes a weighted combination of the standard denoising loss and the hallucination loss.

During denoising at timestep t, the model generates intermediate features $x_t \in \mathbb{R}^{C\times H\times W}$. Simultaneously, the cross-attention maps $A \in \mathbb{R}^{N\times H\times W}$, where N is the number of text tokens, encode the correspondence between tokens and spatial regions.

We first normalize each attention map over the token dimension using a softmax:

$$\tilde{A}_{n,h,w} = \frac{\exp(A_{n,h,w})}{\sum_{n'=1}^{N} \exp(A_{n',h,w})} \tag{1}$$

We then aggregate the normalized attention maps over all tokens to obtain an overall grounding map:

$$A_{\text{sum}}(h,w) = \sum_{n=1}^{N} \tilde{A}_{n,h,w} \tag{2}$$

The hallucination mask is defined as:

$$M_{\text{halluc}}(h,w) = 1 - \text{clip}(A_{\text{sum}}(h,w), 0, 1) \tag{3}$$

where clip$(\cdot)$ restricts values to the $[0,1]$ range.

Given M_{halluc}, we define the hallucination loss as:

$$\mathcal{L}_{\text{halluc}} = \|M_{\text{halluc}} \odot x_t\|_1 \tag{4}$$

where $\odot$ denotes element-wise multiplication and $\|\cdot\|_1$ denotes the ℓ_1 norm over all pixels and channels.

Finally, the overall training objective combines the standard denoising loss $\mathcal{L}_{\text{denoise}}$ and the hallucination loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{denoise}} + \lambda\mathcal{L}_{\text{halluc}} \tag{5}$$

where λ is a hyperparameter balancing the two terms.

7.2 Context-Aware Noise Conditioning

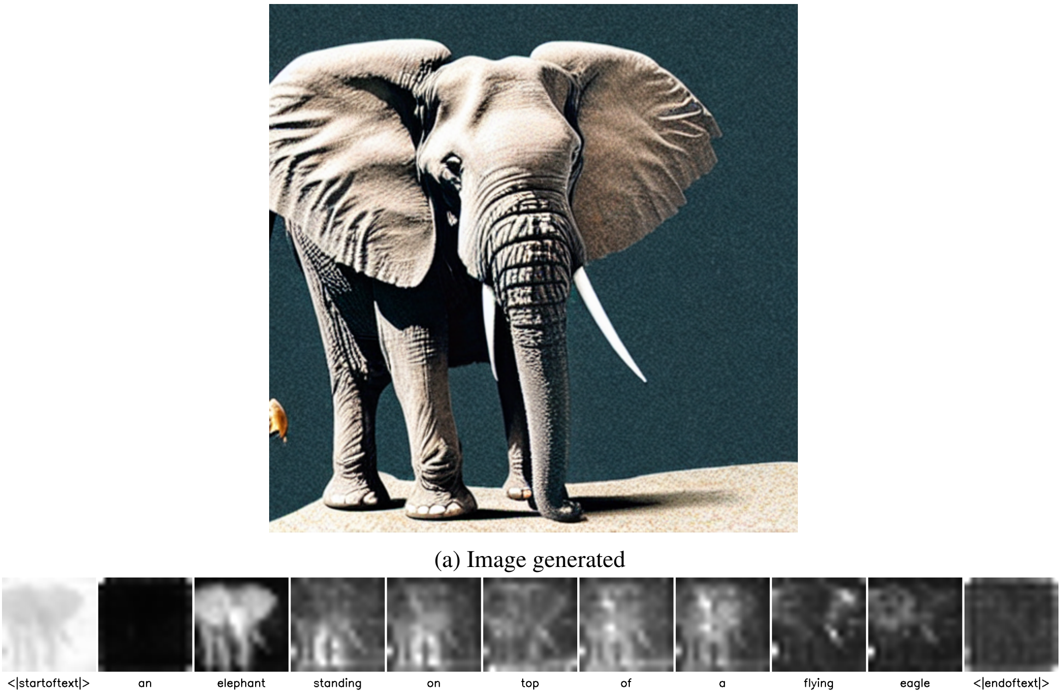
In standard diffusion models, purely random Gaussian noise is added to images during the forward process. While effective, this approach makes the denoising process heavily reliant on the prompt alone for reconstructing fine details, increasing the risk of hallucination when prompts are weak.

To alleviate this, we propose conditioning the added noise on the model's internal feature representations. Rather than injecting purely random noise, the noise at each timestep is generated as a stochastic function of the feature maps obtained during the forward pass. This ensures that the noise remains random but retains contextual relevance to the input image distribution.

By injecting context-aware noise, we aim to constrain the model's denoising trajectories within more meaningful regions of the data manifold. As a result, the model becomes less dependent solely on the prompt for reconstruction, leading to reduced hallucination and more faithful generations.

8 Experiments and Results

In this work, we perform targeted fine-tuning of the diffusion model `CompVis/stable-diffusion-v1-4` by applying Low-Rank Adaptation (LoRA) specifically to the cross-attention layers. LoRA is a parameter-efficient fine-tuning technique that injects



(a) Image generated

(b) Attention Maps

Figure 1: Prompt used: An elephant standing on top of a flying eagle

trainable low-rank matrices into existing weights, allowing for effective adaptation without modifying the entire model. This approach enables lightweight fine-tuning of the attention mechanism responsible for text-image alignment, which is critical for mitigating hallucinations.

We use approximately 20% of the original test dataset for this fine-tuning phase, selecting a diverse subset to ensure broad textual and visual coverage while maintaining computational feasibility. The fine-tuning process focuses on strengthening the alignment between text tokens and the spatial grounding in generated images, based on the hallucination loss described earlier.

Currently, the model is under active fine-tuning, and quantitative as well as qualitative evaluation metrics are being collected. As such, the final results, including visual samples, attention maps, and hallucination error rates, will be formally presented during the viva examination. We anticipate that the proposed loss functions and fine-tuning approach will lead to a noticeable reduction in hallucinated regions and improved fidelity of text-grounded generation.

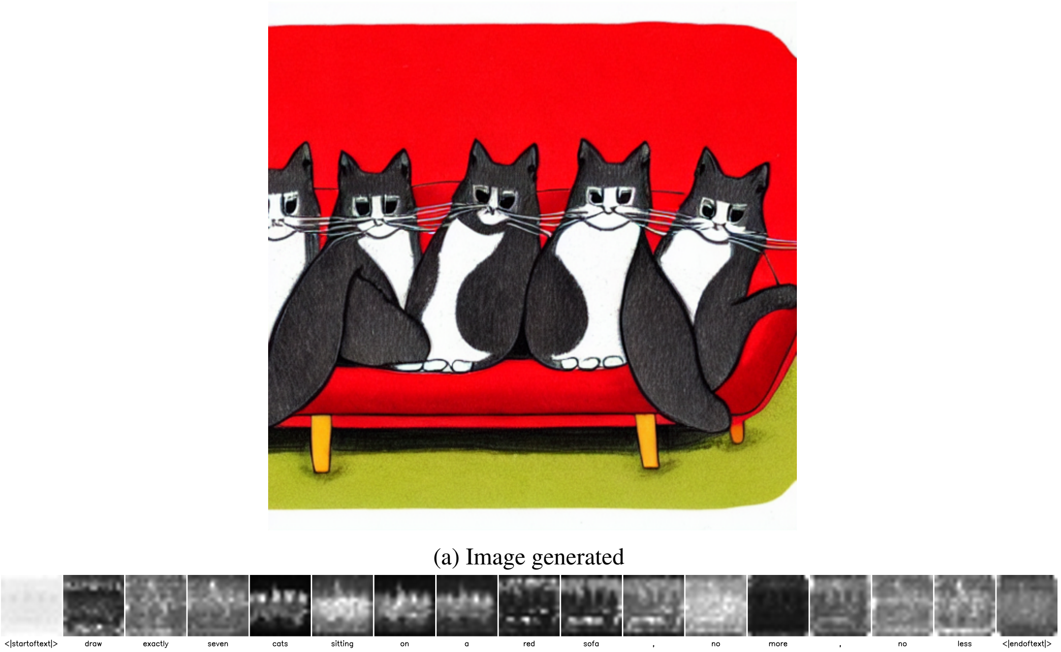
9 Conclusion

In this work, we took a closer look at why hallucinations occur in text-to-image diffusion models and proposed two ways to address them: regularizing cross-attention and making the noise injection process more structured. Our methods are based on a simple observation: when the prompt signal is weak, the model tends to "guess" during denoising, leading to hallucinations. To counter this, we designed a hallucination-aware loss and fine-tuned only the cross-attention layers using LoRA, keeping the overall model lightweight and efficient. Although the model is still under training, our preliminary findings suggest that tuning attention alone could make a noticeable difference in how faithfully the images reflect the prompts.

References

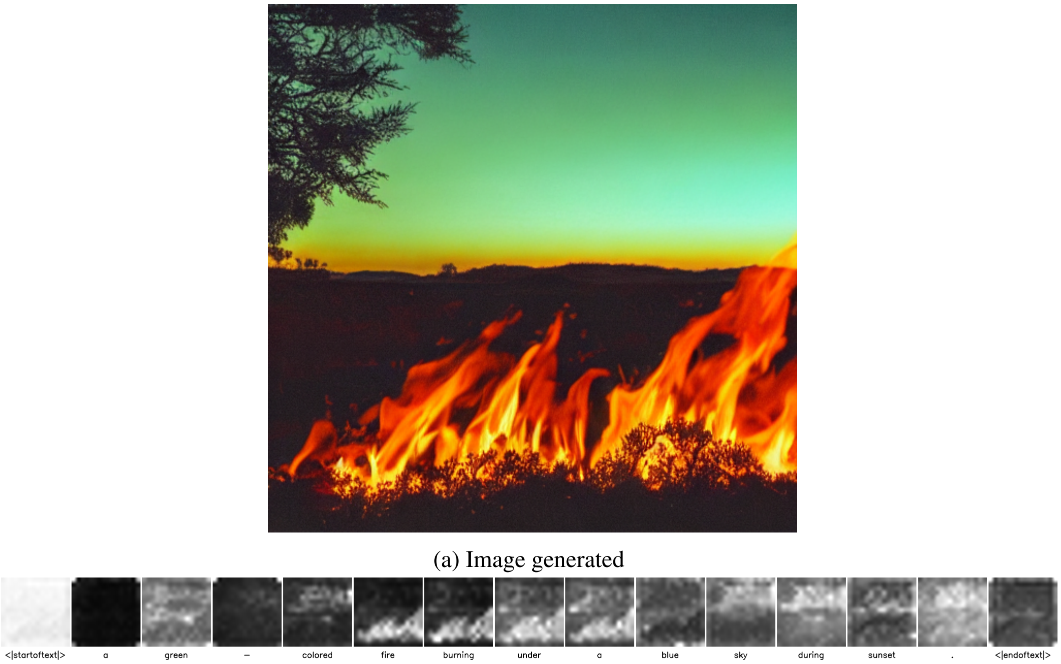
[1] Understanding Hallucinations in Text-to-Image Diffusion Models

[2] Mitigating Hallucination in Multimodal Diffusion Models


(a) Image generated

(b) Attention Maps

Figure 2: Prompt used: Draw exactly seven cats sitting on a red sofa, no more, no less


(a) Image generated

(b) Attention Maps

Figure 3: Prompt used: A green-colored fire burning under a blue sky during sunset.

[3] Investigating Cross-Attention Failures in Diffusion Models

[4] Evaluating Hallucination in Diffusion Models Using Scene-Graph-Based QA Agents

[5] A Survey on Hallucination in Generative Models

[6] Fine-Grained Hallucination (FiG) Metrics for T2I Models

[7] I-HallA v1.0 Pipeline for Multimodal Evaluation

[8] Scene Graph-Based QA Agents for Hallucination Detection

[9] Dependency Parsing in Text-to-Image Alignment