Adversarial Attacks on Parts of Speech: An Empirical Study in Text-to-Image Generation

G M Shahariar, Jia Chen, Jiachen Li, Yue Dong
University of California, Riverside
{gshah010,jiac,jiachen.li,yue.dong}@ucr.edu

Abstract

Recent studies show that text-to-image (T2I) models are vulnerable to adversarial attacks, especially with noun perturbations in text prompts. In this study, we investigate the impact of adversarial attacks on different POS tags within text prompts on the images generated by T2I models. We create a high-quality dataset for realistic POS tag token swapping and perform gradient-based attacks to find adversarial suffixes that mislead T2I models into generating images with altered tokens. Our empirical results show that the attack success rate (ASR) varies significantly among different POS tag categories, with nouns, proper nouns, and adjectives being the easiest to attack. We explore the mechanism behind the steering effect of adversarial suffixes, finding that the number of critical tokens and content fusion vary among POS tags, while features like suffix transferability are consistent across categories. We have made our implementation publicly available at - https://github.com/shahariar-shibli/Adversarial-Attack-on-POS-Tags.

1 Introduction

Text-to-Image (T2I) generation models such as Stable Diffusion (Rombach et al., 2022; Podell et al., 2023), DALL-E2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), ediff-i (Balaji et al., 2022) have made steady progress in the field of image generation by bridging the semantic gap between textual descriptions and visual representations. Unlike traditional methods reliant solely on pixel manipulation, these models leverage multi-model alignments in latent spaces to interpret and synthesize complex visual content from textual prompts. Recent studies, such as Tang et al. (2023), have interpreted how cross-alignment from texts to images is transformed through text-image attribution analysis, demonstrating that different POS tags are well captured by cross-modal attention during synthesis.

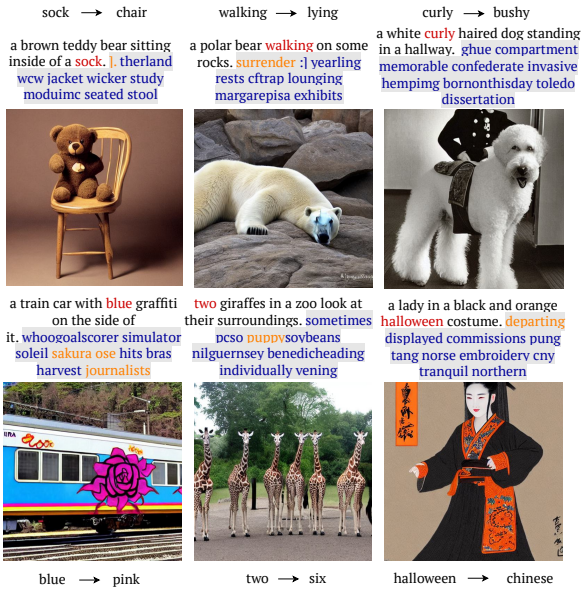

Figure 1: Examples of successful adversarial attacks on Stable Diffusion covering different POS tags drawn from our dataset. The POS tokens targeted by adversarial suffixes are highlighted in red. In addition, we observe that the attack success rate (ASR) varies significantly across POS tag categories, with features like the number of critical tokens (defined in §5, non-critical tokens are highlighted in orange) being highly associated with ASR.

On the other hand, recent research shows that T2I models are vulnerable to adversarial perturbations in text prompts, such as inserting nonsensical words (Millière, 2022), phrases (Maus et al., 2023), or irrelevant characters (Zhuang et al., 2023), which can significantly bias the generated images (Chefer et al., 2023; Salman et al., 2023). However, current adversarial attacks on T2I generation models, either manual heuristic-based methods (Zhuang et al., 2023; Gao et al., 2023; Maus et al., 2023) or automatic gradient-based approaches (Zhuang et al., 2023; Liang et al., 2023; Liu et al., 2023; Shahgir et al., 2023; Yang et al., 2024a,b; Du et al., 2024; Zhai et al., 2024), are specifically targeting entities

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 12874–12890
November 12-16, 2024

Research Paper	POS Tags to Attack	Data Source
Zhuang et al. (2023)	Noun	ChatGPT
Liu et al. (2023)	Noun	ImageNet-1K
Shahgir et al. (2023)	Noun	Manual MS-COCO
Yang et al. (2024a)	Noun	MS-COCO
Yang et al. (2024b)	Noun	ChatGPT
Du et al. (2024)	Noun	ImageNet-1K
This work	Noun, Proper Noun, Adjective, Verb, Numeral, Adverb	MS-COCO

Table 1: Comparison of T2I adversarial attacks based on targeting parts of speech.

or objects (i.e., *nouns*) in text prompts, neglecting other parts of speech. In this paper, we aim to answer the following two research questions:

- **Q1**: Do adversarial attacks, particularly gradient-based attacks on T2I models, behave similarly when targeting different POS tag categories?
- **Q2**: Are there common or distinct features relevant to attack success rates (ASR) when targeting different POS tag categories under adversarial attacks?

To bridge the gap in analyzing attack mechanisms across different POS tag categories beyond nouns, we first created a dataset with realistic scenarios for swapping different POS tag categories with adversarial attacks. Figure 1 provides a few examples drawn from our dataset covering six POS tags from Tang et al. (2023): *noun, adjective, verb, adverb, numeral*, and *proper noun*, with adversarial suffixes that successfully mislead T2I models into generating images related to the targeted attribute. Creating such a dataset is non-trivial, as Shahgir et al. (2023) noted that ASR to T2I models might be affected by internal bias rather than the attack itself; we tried to minimize such biases when creating the dataset. To the best of our knowledge, there is currently no dataset available for analyzing adversarial attacks on POS tags other than nouns (refer to Table 1).

We conduct targeted adversarial attacks over POS tag categories with a gradient-based token searching algorithm specifically designed for T2I models to effectively navigate the larger vocabulary size of the T2I text encoder (Shahgir et al., 2023). The attack objective is to create an adversarial prompt that causes a target POS token to appear in the generated image while ensuring the original POS token from the input prompt does not. We observe that the ASR differs significantly across different POS tag categories. Nouns, proper nouns, and adjectives are the easiest to attack, with increasing difficulties, while the other three categories, with the same gradient-based attack, offer almost no success, whether in restricted (preventing the target token's POS tag attribute from appearing in the adversarial suffix) or unrestricted settings.

This observation led us to further investigate whether there are features associated with these differences in ASR across POS categories. Through extensive experiments, we discovered a correlation between the number of critical tokens in adversarial suffixes and the attack success rate across different POS categories. Critical tokens are those whose removal from the adversarial suffix renders the attack unsuccessful.

Additionally, the results from our ablation study reveal that adversarial suffixes, while steering the generation of target attributes, often fail to completely remove the original attribute. For example, when attempting to change a purple grape to a green one using adversarial suffixes, the resulting image often shows a mixed color. In contrast, with noun attacks, both objects can be generated, whereas with verbs, it is difficult to mix or generate both original and swapped tokens. This varying ease of content fusion across POS categories may also contribute to the differences in ASR.

Furthermore, we identified a general feature shared across different POS categories: the transferability of the attack suffix. Similar to nouns, the adversarial suffixes found are universally transferable to other input prompts with the same attributes. This means an adversarial suffix can transform multiple input prompts with different attributes into the same target attributes in the generated images. For instance, we found that the same adversarial suffix targeting a 'blue' cup can steer the model to generate images of a blue cup across multiple input prompts with different original colors (e.g., red, yellow, orange).

2 Related Work

Text-to-Image Diffusion Models. Nichol et al. (2021) formalized the initial text-to-image (T2I) diffusion model (GLIDE) that substituted class labels with text in class-conditioned diffusion models (i.e. Ablated Diffusion Models (Dhariwal and Nichol, 2021)). The authors explored two types of text conditioning methods: classifier guidance

and classifier-free guidance (CFG). Saharia et al. (2022) proposed Imagen by following the classifier-free guidance (CFG) of GLIDE for T2I generation. They utilized pre-trained large language models (LLMs) as the text encoder and found that scaling up language models is more efficient in improving sample fidelity and aligning images with text. Ramesh et al. (2022) created DALL-E2, a T2I generation model capable of sequentially generating images using text embeddings to guide the process. They achieved this by training a generative diffusion decoder to reverse the image encoding process of CLIP (Radford et al., 2021). Rombach et al. (2022) developed the Latent Diffusion Model (LDM) by incorporating denoising methods within the latent space of pre-trained autoencoders and improving the U-Net architecture with the cross-attention mechanism. Stability AI has utilized the LDM framework to create and introduce a variety of text-to-image diffusion models called the Stable Diffusion series.

Adversarial Attacks on T2I Models. Existing research on adversarial attacks on T2I models primarily falls into two categories: query or heuristic-based and gradient-based. Within the first category, recent studies have explored the excessive sensitivity of T2I diffusion models to minor changes in text prompts. Maus et al. (2023) introduced a query-based attack that discovers prepended prompts capable of causing T2I diffusion models to generate specific image categories. Zhuang et al. (2023) targeted the text encoder of diffusion models by appending extra nonsensical characters to the input prompt using a genetic algorithm. Gao et al. (2023) first identified keywords based on their impact on the generation distribution and then applied character-level substitutions, such as typos, glyphs, and phonetic variations. In the second category, there has been a recent increase in gradient-based adversarial attacks targeting the text encoder of T2I models. Liu et al. (2023) introduced a gradient-guided optimization process to refine a continuous token embedding, using gradients to navigate the prompt space and identify failure cases. Yang et al. (2024a) explored a focused targeted attack that adds target objects while removing original ones, and developed MMP-Attack, which incorporates multi-modal features. Du et al. (2024) proposed Auto-attack on Text-to-image Models (ATM), which automatically generates attack prompts that resemble clean prompts by replacing or adding words. Shahgir et al. (2023) applied gradient-based token perturbation methods to replace entities in the prompt with adversarial suffix tokens. We adopt the gradient attack proposed by (Shahgir et al., 2023) because it aligns with our attack objectives and demonstrates strong performance in targeting nouns.

3 Dataset Creation

In this section, we outline the procedure for constructing our dataset. We first specify the dataset source and then describe the steps involved in its construction.

Data Collection. The first obstacle we encountered in evaluating adversarial attacks across different POS categories beyond nouns was that there was no existing dataset for fair comparison. Table 1 compares existing adversarial attack datasets by size, parts of speech covered, and data sources. To construct our dataset, we chose MS-COCO (Lin et al., 2014) as the data source for its diverse and complex captions making it suitable for testing the robustness of SD. In the train split of MS-COCO, each image has five captions. We collected only the first caption among the five resulting in 118,287 rows.

Input Prompts Selection. We identified the POS tags in each caption from the initially collected data using the NLTK library (Bird, 2006) and a pre-trained POS tagging model (Sajjad et al., 2022). We only focused on six parts of speech tags: noun, verb, adverb, adjective, numeral, and proper noun. For each POS tag, we then randomly selected 20 unique captions, each containing at least one word from the corresponding POS tag, to be used as input prompts.

Target Prompts Generation. For each input prompt of every POS tag, we generated five target prompts, resulting in 100 prompt pairs per POS tag. Each input and target prompt differed by only one word, with the target words chosen from a pool of candidate words. The process of generating target prompts starts by extracting the POS-tagged word from the input prompt using the same NLTK library and pre-trained POS tagging model employed during the input prompt selection. Then, we compile a set of candidate words by gathering other words of the same POS category, identifying antonyms to introduce variety, `[MASK]` prediction to acquire the top-5 words, and exploring the CLIP token embedding space to find the top-k distant neighbors of the word. To extract antonyms, we use the

NLTK library and the WordNet database (Fellbaum, 2010). For [MASK] prediction, we employ BERT (Devlin et al., 2019) as a masked language model. To identify the farthest neighbor tokens in the vocabulary space, we calculate the cosine similarity between the extracted input word embedding and the embeddings of other tokens in the vocabulary, selecting the top 100 tokens with the lowest cosine similarity scores. These candidate words are then filtered to ensure they retain the same POS while removing synonyms, subwords, and substrings to maintain relevance and avoid redundancy. Using these filtered candidate words, we then generate ten candidate prompts ranked highest through [MASK] prediction probabilities. These prompts are subsequently ranked based on their perplexity scores, which measure the fluency and coherence of the prompts. The perplexity score is calculated using the GPT-2 model (Radford et al., 2019). Finally, the five prompts with the lowest perplexity scores, indicating the highest quality, are selected as the final target prompts. We repeat this process for each of the six POS tags, resulting in a total of 600 prompt pairs.

Annotator Recruitment. Our study involves two annotation tasks: dataset annotation and attack success evaluation. For these tasks, we chose two annotators with expertise and research experience in vision and language-related tasks. We chose them from a group of five candidates based on their trustworthiness scores (Price et al., 2020), which were determined through an assessment. We presented them with 30 image-text pairs and asked whether the image accurately reflected the text description (Yes/No). From our dataset, we randomly chose 20 text prompts and generated one image per prompt using SD. In addition, we created 10 text prompts with the help of ChatGPT using the prompt "Generate 10 simple scenes for text-to-image generative model" and then using SD generated one image per prompt. These 10 image-text pairs served as control samples, which were unknown to the participants in advance. Upon completion of the task, we assessed the number of correctly labeled control samples for each candidate. Candidates who achieved a trustworthiness score exceeding 90% were selected as annotators.

Dataset Annotation. We assigned one annotator the task of assessing the meaningfulness of the generated target prompts. The annotator was provided with 600 prompt pairs. For each prompt, we also presented the annotator with 10 candidate target words generated using ChatGPT. We used the prompt "Replace [MASK] with the most probable 10 words in the following text: " to generate candidate target words by ChatGPT. If the target prompt generated from our pipeline appeared meaningful and the annotator considered it visually representable, we instruct the annotator to retain it; otherwise, we ask to replace the corresponding word with an alternative from the pool of ChatGPT-generated words. Out of the 600 prompt pairs, the annotator opted to replace 97 target prompts.

4 Experiment

In this section, we outline the gradient-based adversarial attack method, describe the experimental setup, and report the results to assess the effectiveness of the attack.

4.1 Attack Method

Gradient-based attacks on Stable Diffusion (Zhuang et al., 2023; Shahgir et al., 2023; Yang et al., 2024a,b; Du et al., 2024) utilize the gradient information to perturb the input prompt in a way that maximizes the divergence from the intended output, effectively manipulating the image synthesis process. While previous studies have predominantly focused on nouns, our analysis extends this approach to other parts of speech by applying the gradient-based attack framework proposed by (Shahgir et al., 2023). The process of such an attack on T2I models generally starts with an initial prompt, which is modified iteratively to create an adversarial prompt that maximizes a predefined score function. This involves embedding the target prompt and the adversarial prompt using a token embedder and processing them through a text encoder. The core mechanism focuses on creating multiple candidate prompts by replacing tokens and computing the top-k token candidates. The best candidate prompt, which maximizes the score function, is selected, and the gradient of the loss function concerning the adversarial prompt is used to iteratively refine the prompt. This iterative optimization adjusts the adversarial prompt to gradually increase the discrepancy between the model's output for the target prompt and the adversarial prompt, effectively fooling the T2I generation model. Further details of the attack are provided in Appendix B. We conducted the targeted attack under two distinct settings: with and without restrictions. In the unrestricted setting, we

allow the adversarial prompt to include the target token or its sub-tokens as suffix tokens. However, in the restricted attack scenario, we confine the appearance of the target token within the adversarial prompt by constraining all possible substrings of the target token.

4.2 Experimental Setup

We followed the setup of Shahgir et al. (2023) and conducted the attack five times for each pair, with 100 steps per run, employing 10 adversarial tokens. For each step, we selected the top 256 tokens as candidate tokens and generated 512 new prompts by randomly substituting tokens using these candidates. Subsequently, we generated seven images per attack, resulting in the evaluation of a total of 21,000 generated images (600 pairs, 5 runs, and 7 images per run). During image generation, we set the resolution to 512 $\times$ 512, the number of inference steps to 50, and the scale of classifier-free guidance to 7.5. As the victim model, we utilized Stable Diffusion v1.5 (SD v15) for both image generation and performance assessment, leveraging a pre-trained CLIP model trained on a dataset comprising text-image pairs. All experiments (attack execution, evaluation, and image generation) were conducted using a single Nvidia RTX 3090 GPU, totaling approximately 600 GPU hours. The execution time to attack a single input-target prompt pair is approximately 8 minutes.

4.3 Evaluation Metrics

Attack Success Rate. We consider an attack successful if the image generated by the adversarial prompt matches the target text; otherwise, we consider it unsuccessful. Since we generate 7 images per adversarial prompt, to measure the attack success rate (ASR), we consider the attack as successful if at least 4 images have a higher matching score than a threshold. Following (Shahgir et al., 2023), we set this threshold value at 3.41. We determine the matching score by calculating the difference between the CLIP score of the input prompt and the generated image, and the CLIP score of the target text and the generated image. CLIP score measures the cosine similarity between the visual CLIP embedding of an image and the textual CLIP embedding of a text. For each input-target prompt pair, we run the attack five times, generating five adversarial prompts, and consider the attack successful if at least one of them succeeds.

Semantic Shift Rate. For a quantitative measure to evaluate the efficacy of adversarial suffix tokens, we utilized SemSR (Semantic Shift Rate) (Zhai et al., 2024) which measures the semantics between a generated image and a text prompt. SemSR utilizes CLIP's multi-modal embedding space and computes the similarity in semantics between a generated image and a prompt using cosine similarity. This metric quantifies the displacement in the vector space of the generated image after appending adversarial suffix tokens compared to the image generated using the input prompt. Since the amount of deviation necessary to attain diverse target semantics differs, it is adjusted by the maximum deviation. The SemSR equation is provided below:

$$SemSR = \frac{CS(E_{I_a}, E_{P_a}) - CS(E_{I_i}, E_{P_i})}{CS(E_{I_t}, E_{P_t}) - CS(E_{I_i}, E_{P_i})} \quad (1)$$

where CS denotes $CLIP\_Score$, I_a represents the generated image from the adversarial prompt P_a, I_i denotes the generated image from the input prompt P_i, and I_t denotes the generated image from the target prompt P_t. For a single input-target prompt pair, we measure the average of SemSR scores over five runs.

4.4 Results

In Figure 11, we showcase a few examples of images generated through both the unrestricted and restricted attack methods. Table 2 displays the average attack success rate (ASR) and average semantic shift rate (SemSR) over all the prompt pairs for each POS tag under both attack conditions. Below, we present both quantitative analysis and human evaluation of our experiments.

POS Tag	Unrestricted Attack		Restricted Attack	
	ASR	SemSR	ASR	SemSR
Noun	**0.65**	1.4394	**0.51**	1.3884
Proper Noun	0.40	0.8955	0.31	0.8606
Adjective	0.29	**2.0929**	0.24	1.1181
Verb	0.15	1.5963	0.12	**1.9121**
Numeral	0.13	1.9246	0.11	1.5943
Adverb	0.03	0.9313	0.01	1.0077

Table 2: Average Attack Success Rate (ASR) and average Semantic Shift Rate (SemSR) of both unrestricted and restricted attack on each POS Tag. The higher the values, the better. The highest values are **bold** marked.

Quantitative Evaluation. Table 2 presents the ASR and SemSR metrics, which are the average values across 100 data points for each POS tag.

POS Tag	Unrestricted Attack		Restricted Attack	
	Input	Target	Input	Target
Noun	0.13	0.87	0.27	0.73
Proper Noun	0.47	0.53	0.53	0.47
Adjective	0.67	0.33	0.53	0.47
Verb	0.73	0.27	0.67	0.20
Numeral	0.20	0.13	0.20	0.13
Adverb	0.93	0.07	0.87	0

Table 3: Human evaluation results on the matching of input and target text with the generated images for each POS tag in both unrestricted and restricted settings.

Higher ASR and SemSR values indicate better performance. From the table, we observe that in the case of the unrestricted attack, both ASR and SemSR surpass those of the restricted attack except for verb and adverb POS tags. This suggests that allowing the target token to be part of the concatenated adversarial suffix tokens leads to greater success in adversarial attacks. Clearly, the unrestricted attack excels in producing images containing the target POS token instead of the input POS token. We also observe that in both the restricted and unrestricted attacks, nouns demonstrate higher ASR values compared to other POS tags, implying their greater vulnerability to adversarial attacks. Proper nouns and adjectives show moderate success rates, while verbs and numerals exhibit lower success rates. Adverbs, on the other hand, have the lowest success rates in both types of attacks, indicating their higher resistance to adversarial manipulation. SemSR values quantify the semantic disparity between a text and its corresponding generated image caused by an adversarial attack. Higher SemSR values signify substantial semantic shifts. By analyzing SemSR values, we find that attacking nouns and adjectives is comparatively simpler, whereas adverbs present greater difficulty. This suggests that nouns and adjectives undergo more significant semantic alterations, while adverbs experience the least. Moreover, SemSR values remain relatively stable across various POS tags for both unrestricted and restricted attacks. However, with unrestricted attack, adjectives exhibit the greatest semantic shifts, which is not the case with restricted attack. Numerals consistently show the second-highest semantic changes across both attack types.

Human Evaluation. We evaluate the attack's effectiveness with the assistance of two annotators. We randomly choose 15 prompt pairs for each POS tag, amounting to a total of 90 prompt pairs for both unrestricted and restricted attack settings. Each prompt pair is presented with 7 images to the annotators (as 7 images were generated per run in our experiments), who then assess whether at least 4 images closely align with either the target prompt or the input prompt (Yes/No). We collect evaluations from the annotators using a Google Form (Appendix J), which includes the generated image and two checkboxes for the input text and target text. We determine the score of an annotator by the number of prompt pairs they classify as a match. Since there are two evaluators, we calculate the average of their scores and present the results in Table 3. The table indicates that annotators agree that verbs, adverbs, and numerals are more resistant to adversarial attacks. In the case of numerals, the annotators reported that the majority of the post-attack generated images do not align with either the target or the input prompts. We observe that unrestricted attack tends to generate images that more closely match the target prompt than the restricted attack. We used Cohen's Kappa (κ) metrics (Cohen, 1960) to measure annotator agreement on target text-image matching, obtaining scores of 0.796 for unrestricted and 0.745 for restricted settings, which indicate a high degree of agreement.

From Table 2 and 3, we observe that the average ASR shows a strong positive correlation with human evaluation of target text-image matching in both the unrestricted setting (Pearson = 0.988 and Spearman = 1.00) and the restricted setting (Pearson = 0.980 and Spearman = 0.986). On the other hand, the average SemSR exhibits a very weak negative correlation with human evaluation in both unrestricted attack scenario (Pearson = -0.126 and Spearman = -0.143), and restricted attack scenario (Pearson = -0.176 and Spearman = -0.087). Given that the average ASR has higher correlations with human judgment in both settings, it is more reliable than average SemSR for evaluating the success of attacking POS tags. Therefore, we use ASR to evaluate attack success in all subsequent sections.

5 Attack Success Mechanism

In this section, we explore the mechanism behind the steering effect of adversarial suffixes. We identify (a) features that vary across POS categories and explain differences in attack success rates (ASR), such as the number of critical tokens and content fusion, and (b) features that are consistent across

POS Tag	Unrestricted			Restricted		
	Number of Successful Attack	Avg no. of critical tokens	Avg ASR by removing critical tokens	Number of Successful Attack	Avg no. of critical tokens	Avg ASR by removing critical tokens
Noun	65	7.800	0.195	51	8.902	0.136
Proper Noun	40	8.175	0.175	31	8.935	0.115
Adjective	29	7.862	0.173	24	8.960	0.111
Verb	15	8.200	0.166	12	9.000	0.076
Numeral	13	8.615	0.150	11	9.180	0.034
Adverb	3	9.000	0.078	1	10.000	0

Table 4: Comparison of the average attack success rates (ASR) by removing critical tokens across different POS tags, under all unrestricted and restricted successful attack examples.

different POS categories and do not explain variations in ASR rates, but provide general insights such as suffix transferability.

Correlation between the number of critical tokens in adversarial suffixes and ASR. A successful attack demonstrates that appending an adversarial suffix to an input prompt effectively shifts the text embedding toward the target prompt, highlighting the significant role of the suffix tokens. To investigate, we tokenized several adversarial suffixes, generated an image for each token to isolate their contributions, and found that some tokens generate images associated with the target POS token. This observation led us to identify the most contributing tokens within adversarial suffixes. We define "critical tokens" as those whose removal causes the attack to fail. To determine critical tokens in a suffix, we generated all possible combinations of replacing suffix tokens with <|endoftext|> token. For each combination, we generated an image and queried the pre-trained vision-language model BLIP[1] (Li et al., 2022) to check if the generated image matched the target prompt. We identified the combination with the fewest tokens replaced by <|endoftext|> and considered those tokens as critical since their absence leads to an unsuccessful attack. Tokens not replaced by <|endoftext|> were considered non-critical.

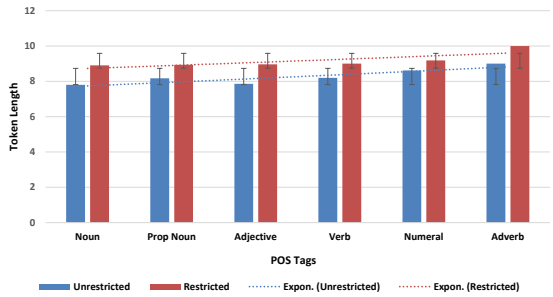

Figure 2: Average length of critical tokens across different POS tags in unrestricted and restricted settings. Exponential trend lines are included for both settings to highlight the general pattern.

We present the average number of critical tokens across all POS tags in Table 4 and compare the lengths of critical tokens in Figure 2. The number of critical tokens is generally higher across POS categories in both attack settings. However, the restricted setting shows significantly higher numbers of critical tokens, as the absence of the target word necessitates other tokens to compensate and maintain the attack's effectiveness. We find that adverbs, numerals, and verbs are more resistant to adversarial attacks due to their dependency on the high number of critical tokens in the suffixes. This prompted us to explore whether every critical token within a suffix contributes equally to the attack's success. Therefore, removing some or all critical tokens from the suffixes should notably decrease the ASR. To test this in both settings, we removed critical tokens from the suffixes in all possible combinations while keeping the non-critical tokens unchanged. We then calculated the ASR for each combination by querying BLIP and took the average. We find that the ASR significantly decreases across POS categories. Table 4 shows that adverbs, numerals, and verbs are the hardest to attack due to their reliance on a higher number of critical tokens, resulting in a significantly lower ASR when these tokens are removed. However, nouns, proper nouns, and adjectives are relatively

[1] https://huggingface.co/Salesforce/blip-vqa-capfilt-large

easier to attack. Despite having a substantial number of critical tokens, the ASR for these categories remains moderately high when critical tokens are removed but still shows a significant drop. For instance, in the unrestricted setting, the attack success rate drops from 65 (total successful attacks) to around 13 (0.195 * 65) when critical tokens are removed. Thus, we conclude that the number of critical tokens in adversarial suffixes is highly associated with ASR. Some POS tags are harder to attack because the attack algorithm must find adversarial suffixes with a higher number of critical tokens.

Ease of Content Fusion. We observe that while adversarial suffixes steer the generation of target attributes, they often fail to completely remove the original token. This results in images generated by the Stable Diffusion containing both the input and target attributes, a phenomenon we refer to as content fusion. We found that content fusion across different POS categories decreases with decreasing ASR. For example, when attempting to change a noun like "car" to "motorcycle", the resulting image often contains both the car and the motorcycle. For a proper noun, changing a "Santa costume" to a "Halloween costume" might result in a Santa costume with Halloween-themed colors. With adjectives, trying to change a "white swan" to a "black swan" can lead to an image of a swan that is both black and white. We showcase examples of adjective fusion in Appendix I. In contrast, verbs are harder to mix or generate together; for instance, it is difficult to create an image where a person is both standing and lying down. Similarly, in the case of numerals, attempting to change "three apples" to "five apples" often fails to produce an image with both three and five apples. With adverbs, changing "running quickly" to "running slowly" does not result in an image that simultaneously depicts both quick and slow running. We posit that this varying ease of content fusion is due to the number of critical tokens associated with ASR. In categories like nouns, proper nouns, and adjectives, where the number of critical tokens is relatively lower, fusion is easier. However, in categories with a higher number of critical tokens, such as verbs, numerals, and adverbs, fusion is not possible.

Suffix Transferability. We discovered a common feature across different POS categories: the transferability of adversarial suffixes. We observed that the identified adversarial suffixes can universally transfer to other input prompts within the same POS tag. This indicates that a single adversarial suffix can convert various input prompts with distinct attributes into images with the same target attributes. For example, an adversarial suffix targeting the noun "motorcycle" can prompt the model to generate motorcycle images from diverse noun prompts like "plane", "car", and "bird". We present some examples in Appendix G (Figure 10). Additionally, to explain why such universal transferability works, we demonstrate by following the approach of (Du et al., 2024) that the adversarial suffix alone can dictate the output of the Stable Diffusion by steering the generated image toward the target prompt. At first, we divide a successful adversarial prompt into two segments: the input prompt and the adversarial suffix. Then we extract text embeddings of both the input prompt and the suffix separately. This step ensures that each segment is processed into its own embedding without influence from the other segment. The embeddings of the input prompt and the suffix are then concatenated. Concatenation ($\oplus$) means joining these two embeddings into a single combined embedding that the Stable Diffusion will use for image generation. We observe that the final image generated by Stable Diffusion using the concatenated text embedding matches the target prompt. We repeat this procedure for all successful attack examples and find consistent results across all POS tags. Further details can be found in Appendix G.

6 Conclusion

In this study, we evaluate a gradient-based adversarial attack aimed at six POS tags within text prompts in both unrestricted and restricted attack strategies. We assess the impact of these attacks on the Stable Diffusion, revealing valuable insights into the factors contributing to their success. Our findings reveal that nouns, proper nouns, and adjectives are particularly vulnerable to perturbation, resulting in adversarial image generation. However, we see that verbs, adverbs, and numerals exhibit a higher level of resilience against adversarial attack, exerting minimal influence on the visual output generated by the Stable Diffusion. We hypothesize that the number of critical tokens in an adversarial suffix and the ease of content fusion are primarily responsible for such resilience against attacks. We believe these findings will be valuable for enhancing the robustness of T2I generation systems.

7 Limitations

We utilized the Stable Diffusion model for the gradient-based attack. It is important to note that the attack approach might not generalize effectively to other closed-source T2I generation models like Imagen (Saharia et al., 2022) or DALL-E2 (Ramesh et al., 2022), owing to differences in architecture, text encoder, and training data. Moreover, the metrics utilized in this study to assess the attack may not fully capture the visual plausibility or semantic accuracy of images after the attack. We evaluated the attack only on six specific POS tags, which may not encompass all possible scenarios, such as *prepositions, conjunctions, interjections, articles,* and *determiners*. Furthermore, the approach relies on appending suffix tokens to the original prompt, which may not always be the most optimal method for manipulating the image generation process, considering the T2I model's sensitivity to the order of tokens. As the appended adversarial suffix tokens may lack meaning, the resulting adversarial prompt found by the attack methods exhibits reduced naturalness.

References

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Chengbin Du, Yanxi Li, Zhongwei Qiu, and Chang Xu. 2024. Stable diffusion is unstable. *Advances in Neural Information Processing Systems*, 36.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Hongcheng Gao, Hao Zhang, Yinpeng Dong, and Zhijie Deng. 2023. Evaluating the robustness of text-to-image diffusion models against real-world attacks. *arXiv preprint arXiv:2306.13103*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. 2023. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 20763–20786. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Qihao Liu, Adam Kortylewski, Yutong Bai, Song Bai, and Alan Yuille. 2023. Discovering failure modes of text-guided diffusion models via adversarial search. In *The Twelfth International Conference on Learning Representations*.

Natalie Maus, Patrick Chao, Eric Wong, and Jacob Gardner. 2023. Black box adversarial prompting for foundation models. *arXiv preprint arXiv:2302.04237*.

Raphaël Millière. 2022. Adversarial attacks on image generation with made-up words. *arXiv preprint arXiv:2208.04135*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Ilan Price, Jordan Gifford-Moore, Jory Fleming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversation. *arXiv preprint arXiv:2010.07410*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Khan, and Jia Xu. 2022. Analyzing encoded concepts in transformer language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3082–3101, Seattle, United States. Association for Computational Linguistics.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. 2023. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*.

Haz Sameen Shahgir, Xianghao Kong, Greg Ver Steeg, and Yue Dong. 2023. Asymmetric bias in text-to-image generation with adversarial attacks. *arXiv preprint arXiv:2312.14440*.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2023. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659, Toronto, Canada. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Dingcheng Yang, Yang Bai, Xiaojun Jia, Yang Liu, Xiaochun Cao, and Wenjian Yu. 2024a. Cheating suffix: Targeted attack to text-to-image diffusion models with multi-modal priors. *arXiv preprint arXiv:2402.01369*.

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. 2024b. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 123–123. IEEE Computer Society.

Shengfang Zhai, Weilong Wang, Jiajun Li, Yinpeng Dong, Hang Su, and Qingni Shen. 2024. Discovering universal semantic triggers for text-to-image synthesis. *arXiv preprint arXiv:2402.07562*.

Haomin Zhuang, Yihua Zhang, and Sijia Liu. 2023. A pilot study of query-free adversarial attack against stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2384–2391.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Appendix

A Preliminaries of Stable Diffusion

Stable diffusion is a latent diffusion model comprising three key components: a Variational Autoencoder (VAE), a UNet, and a CLIP text encoder (Radford et al., 2021) for conditioning. The VAE consists of an encoder E and a decoder D, where the encoder compresses an image y into a lower-dimensional latent space representation $E(y)$, while the decoder reconstructs the image from the latent space $\bar{y} = D(E(y))$. During the T2I generation process, at first CLIP tokenizer tokenizes a text prompt into a sequence of tokens $W = \{w_1, w_2, ...w_n\}$, ensuring uniform length by

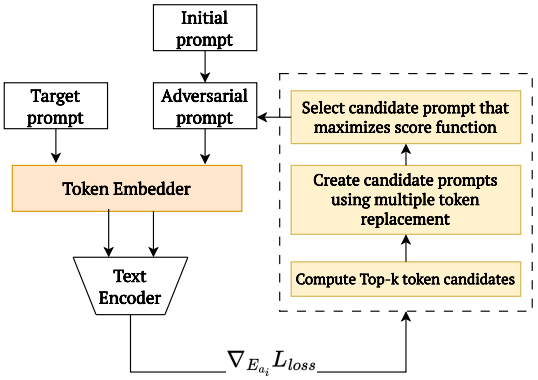

Figure 3: Schematic view of the POS-Attack pipeline. At first, hidden state representations from the CLIP text encoder using input and target token embeddings are extracted. Then, we compute loss, take gradients, and select the top-k candidate tokens for substitution. Next, we create several candidate prompts by randomly replacing multiple tokens from the pool. The candidate prompt maximizing a score function is chosen for the next optimization step.

Figure 4: Examples of vulnerabilities revealed by SD model with prompts containing adverbs and proper nouns.

padding or truncating sequences to 77 tokens for computational ease. Each token is then converted into a text representation W_{emb} using the text encoder of CLIP. CLIP comprises both an image encoder and a text encoder, each responsible for encoding an image and its corresponding text description into representations that closely align with one another. As a result, the text representation W_{emb} generated by CLIP's text encoder for a given text prompt is expected to contain relevant information about the images described in the prompt. Next, a random latent image representation I_0, drawn from a Gaussian distribution is created, and noise is gradually eliminated to get a noise-free representation $E(y) = I_{emb}$ through a reverse diffusion process. Guided by the latent text embedding W_{emb}, a UNet neural network $U(I_0, W_{emb}, t)$ employs a cross-attention mechanism to predict and eliminate noise from the latent space $E(y)$ at each time step t. The level of noise reduction is regulated by a scheduler, progressively refining image quality. Finally, the VAE decoder D upscales the latent image $E(y)$ back into pixel space, resulting in a high-resolution image $\bar{y}$.

B Details of Adversarial Attack

Adversarial Prompt Generation. We start the process by considering the input prompt as the adversarial prompt. Subsequently, we extract the embeddings for both the adversarial prompt tokens and the target prompt tokens. These embeddings are then fed into the CLIP text encoder to obtain the final hidden state representations. Following this, we compute the loss using a loss function and calculate gradients with respect to one hot token vector to determine the top-k candidate tokens for substitution. Then, we generate several candidate prompts by randomly replacing multiple tokens of the initial adversarial prompt from the pool of candidate tokens. The candidate prompt that maximizes the score function is chosen as the adversarial prompt for the subsequent optimization step. This iterative process continues for a set number of iterations until a final adversarial prompt is obtained. Notably, this attack method relies only on the text encoder and does not necessitate access to the image generation model. An illustration of the adversarial attack is depicted in Figure 3. From the adversary's viewpoint, the concatenated suffix tokens in the adversarial prompt should be nonsensical to humans yet encode specific semantics predetermined by the adversary.

Loss Function. The attack focuses on manipulating the CLIP embedding space to optimize a score function, which quantifies how much the adversarial token embeddings at an intermediate optimization stage deviate towards the target token embeddings using cosine similarity. The objective is to steer away from the embeddings of input tokens and progressively approach those of the target tokens by discovering more effective adversarial tokens. This process of maximizing the score function is similar to Shahgir et al. (2023). To compute the loss, we adopt the negated score function. Maximizing the score is equivalent to minimizing the loss.

Gradient-based Search. The attack employs an effective greedy coordinated gradient-based search algorithm (Zou et al., 2023), utilizing the loss function discussed above. At each optimization step, the algorithm selects k tokens with the highest negative loss and computes gradients with respect to

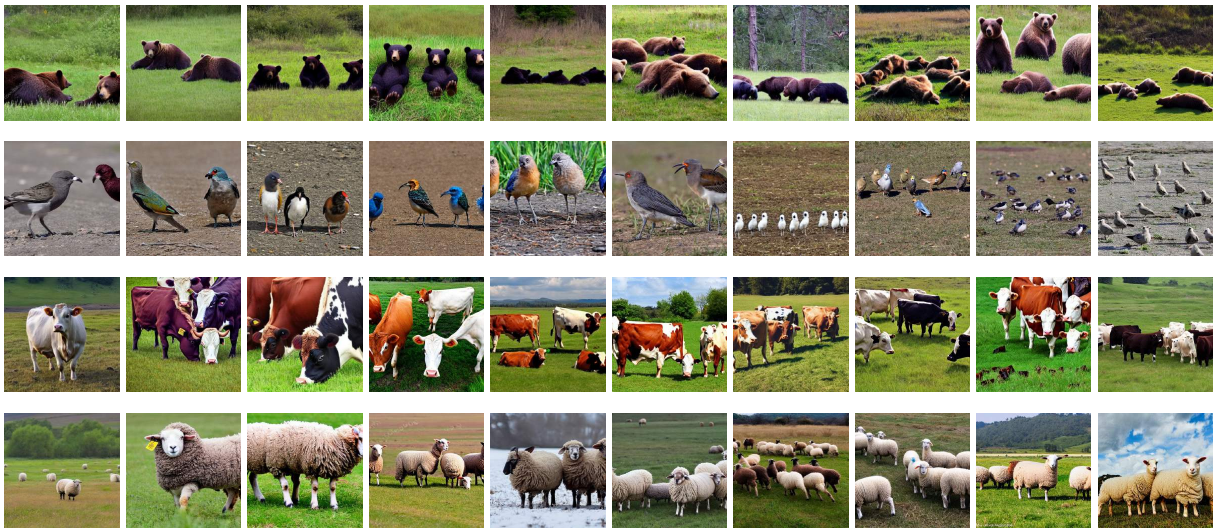

Figure 5: Examples of vulnerabilities revealed by SD model with numerals. Each row contains 10 images with numerals one to ten from left to right sequentially. The first row contains ten images of the prompt *"___ bears lying in the field"* where ___ is replaced by "one" to "ten" serially. Similarly, the second, third, and fourth rows contain prompts *"___ birds looking around while on the ground", "___ cows eating grass"* and *"___ sheep roaming in the field"*.

the one-hot token vectors to identify a promising set of candidates for replacing adversarial suffix token positions. New candidate prompts are generated by randomly replacing multiple token positions using the pool of token candidates, repeating this process T times. Following the approach of Shahgir et al. (2023), we initially replace all tokens and then gradually reduce the replacement rate to 20%.

C Vulnerabilities Observed across POS Tags

In this section, we present some vulnerabilities observed on Stable Diffusion across a few POS tags. We noticed that the SD model inherently faces difficulty generating images from prompts that include numerals. Specifically, the model struggles to produce images with a precise count of identical objects. For instance, if the prompt is to generate an image of five birds, the SD model will fail to create exactly five birds and instead produce images with a random number of birds, such as three, four, or more than five. Examples of this issue are shown in Figure 5. Images generated by the model using prompts where the adverb tokens have shared linguistic structures, close semantic representation in the feature space, and unrelated to emotions generally have minimal impact on visual output. We show such an example in Figure 4(a). In this example, substituting *"beautifully"* with *"partly"* in the prompt *"a bench that is beautifully shaded by a tree"* results in close perplexity[2] scores for the first (128.18) and second prompts (123.49). exhibit little difference. Furthermore, we observed that the SD model struggles to generate images involving logos, such as those of Microsoft, Disney, or Google. As shown in Figure 4(b), instead of producing accurate images, the model generates images with misspelled words as logos.

D Impact of Semantic Distance on Attack Success

In this section, we examine why certain POS tags are easier to attack by considering the impact of semantic distance. To explore this, we plotted text embeddings of all the data across six POS tags from the dataset in Figure 6 using t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008). Nouns, proper nouns, and adjectives show clear clustering with visible distances between markers, indicating considerable differences in their text embeddings. This distance allows the gradient algorithm to minimize the gap from input to target, making attacks on these POS tags easier. However, for numerals, verbs, and adverbs, the markers are very close or even overlapping, indicating that the input and target prompts have similar semantic representations. This prox-

[2]We utilized GPT-2 (Radford et al., 2019) for perplexity score calculation.

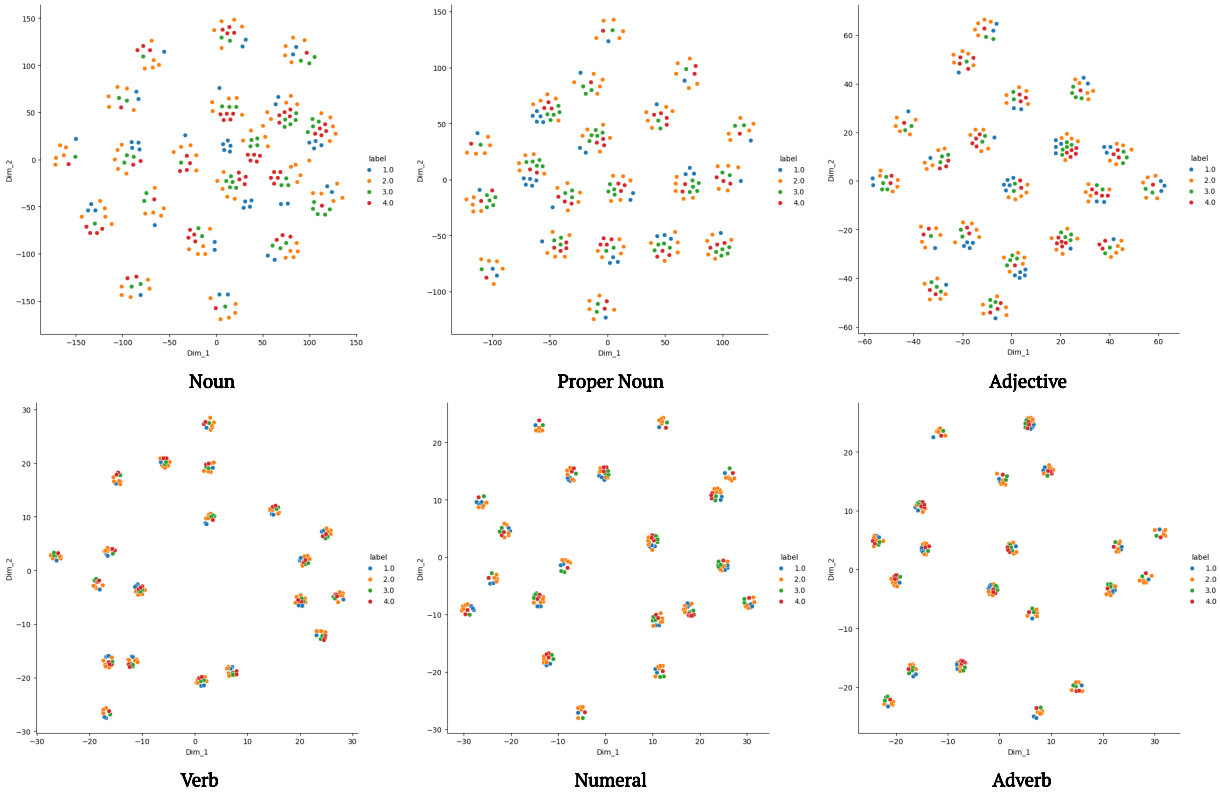

Figure 6: Visual representation of text embeddings of all input-target prompt pairs across six POS tags in 2D space. Blue marker is for input prompt, yellow indicates target prompt, green for inherent bias inducing prompt, and red for prompt where the target word is removed.

imity hampers the algorithm's ability to optimize the distance gap, leading to lower attack success rates.

Figure 7: Examples of adversarial attack on Adjective using ASCII/non-alphabetic characters.

E Attack using ASCII/non-alphabetic characters

We evaluated the adversarial attack by limiting it to adding only ASCII or non-alphabetic characters at the end of the input prompt. However, these attempts were unsuccessful. We conducted experiments by adding 5, 20, and 50 adversarial characters to a text prompt containing an adjective token, yet there was no alteration in the resulting image. In all three scenarios, we noticed that even with the addition of characters, the SD continued to generate images identical to those generated from the original input prompt. We present a few examples in Figure 7.

F Attack Transferability

We employed Stable Diffusion v1-5, utilizing CLIP ViT-L/14 (Radford et al., 2021) as the pre-trained text encoder, for both unrestricted and restricted attack methods. Our investigation reveals that adversarial suffixes generated with this version of SD are ineffective when applied to Stable Diffusion v2-1, which uses a different pre-trained text encoder, OpenCLIP-ViT/H (Cherti et al., 2023). However, we note that the attacks successfully transfer to Stable Diffusion v1-4. Examples generated using SD v1-4 are presented in Figure 8.

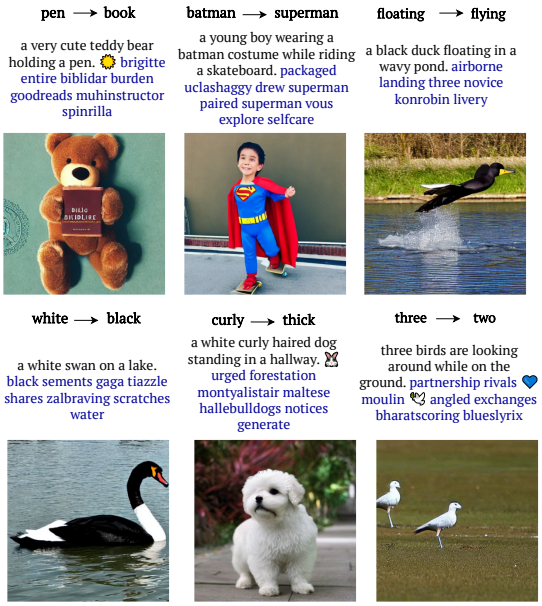

Figure 8: Some examples of successful attack on Stable Diffusion v1-4.

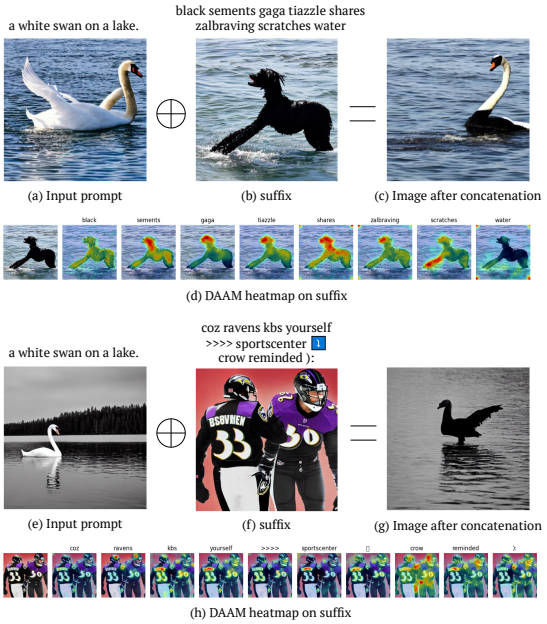

Figure 9: Examples demonstrating how the adversarial suffix independently dictates the SD model's output.

G Examples of Suffix Transferability

We provide a successful attack example in Figure 9 where the input and target prompts are *"a white swan on a lake."* and *"a black swan on a lake."* respectively. Figure 9(a - d) represents the unrestricted setting. Figure 9(a) and 9(b) represent the images generated by the SD model using the text embedding of the input prompt and the suffix respectively while 9(c) is the final image generated using the concatenated text embedding. As the suffix tokens are largely nonsensical, we employed DAAM (Tang et al., 2023) to generate word attribution heatmap in order to delineate which suffix words correspond to which portion of the suffix image, as shown in Figure 9(d). Similarly, we repeat the procedure for the restricted setting in Figure 9(e - g).

Figure 10: Examples of adversarial suffix transferability. The top two rows are the examples of *noun* and *proper noun* POS tags in unrestricted settings where the target words are "motorcycle" and "Halloween" respectively. The last two rows correspond to the examples of *verb* and *adjective* POS tags in restricted settings where the target words are "watching" and "blue" respectively.

H More Examples of Adversarial Attack

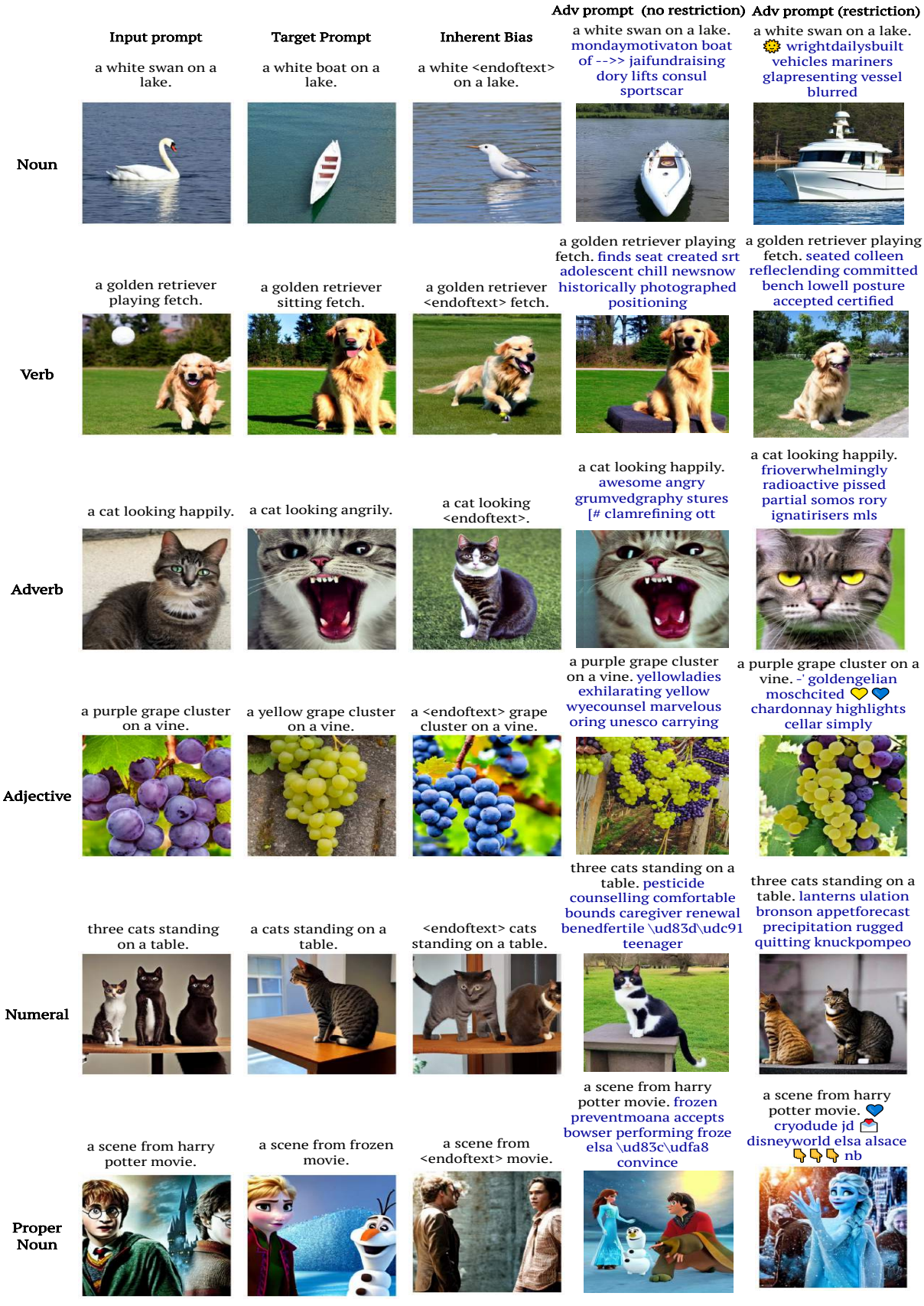

Figure 11: Few examples of successful attacks on out-of-dataset instances for each POS tag.

I Some Examples of Adjective Color Fusion

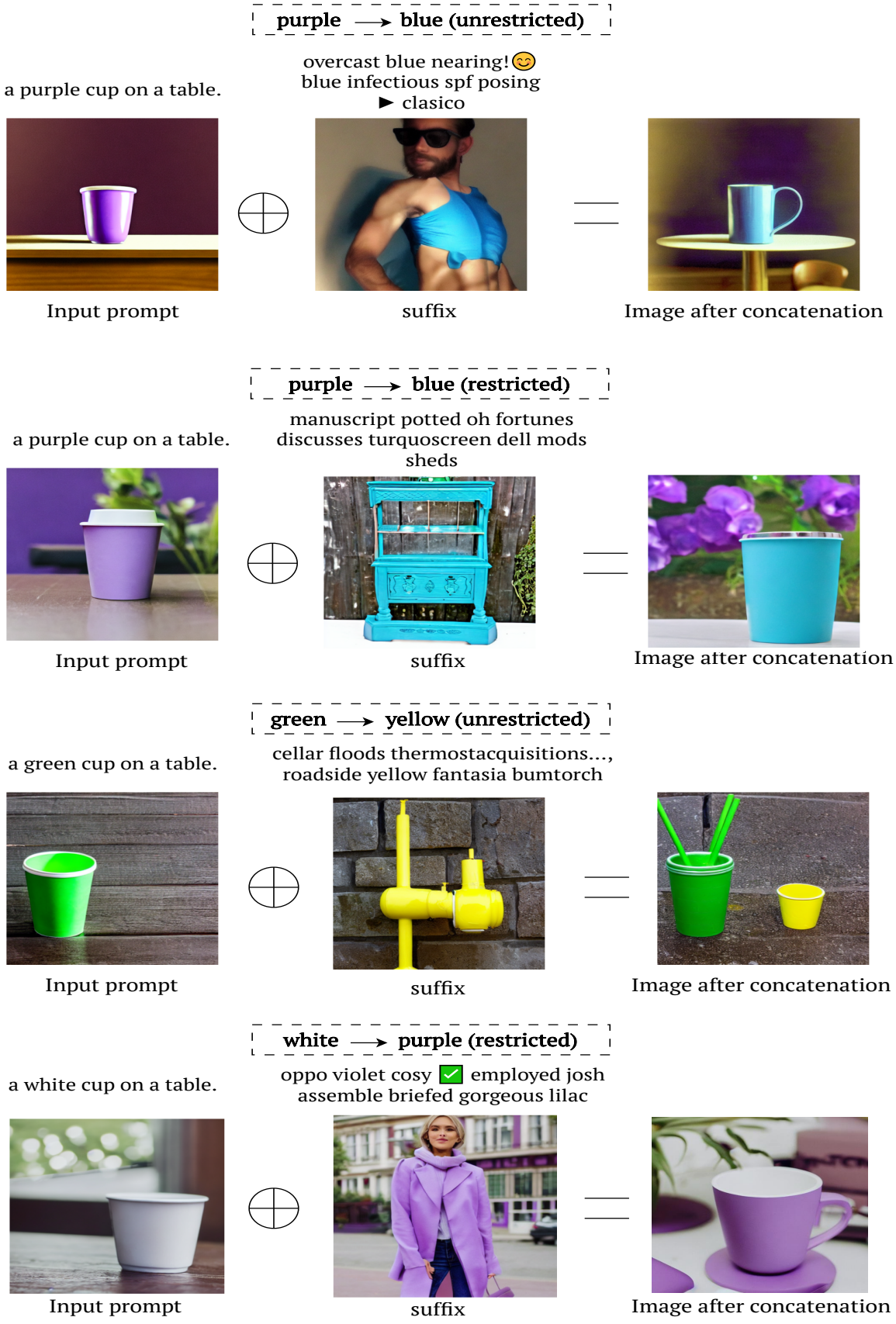

Figure 12: Examples of adversarial suffix found by the adversarial attack responsible for fusion of color adjectives.

J Human Evaluation Template

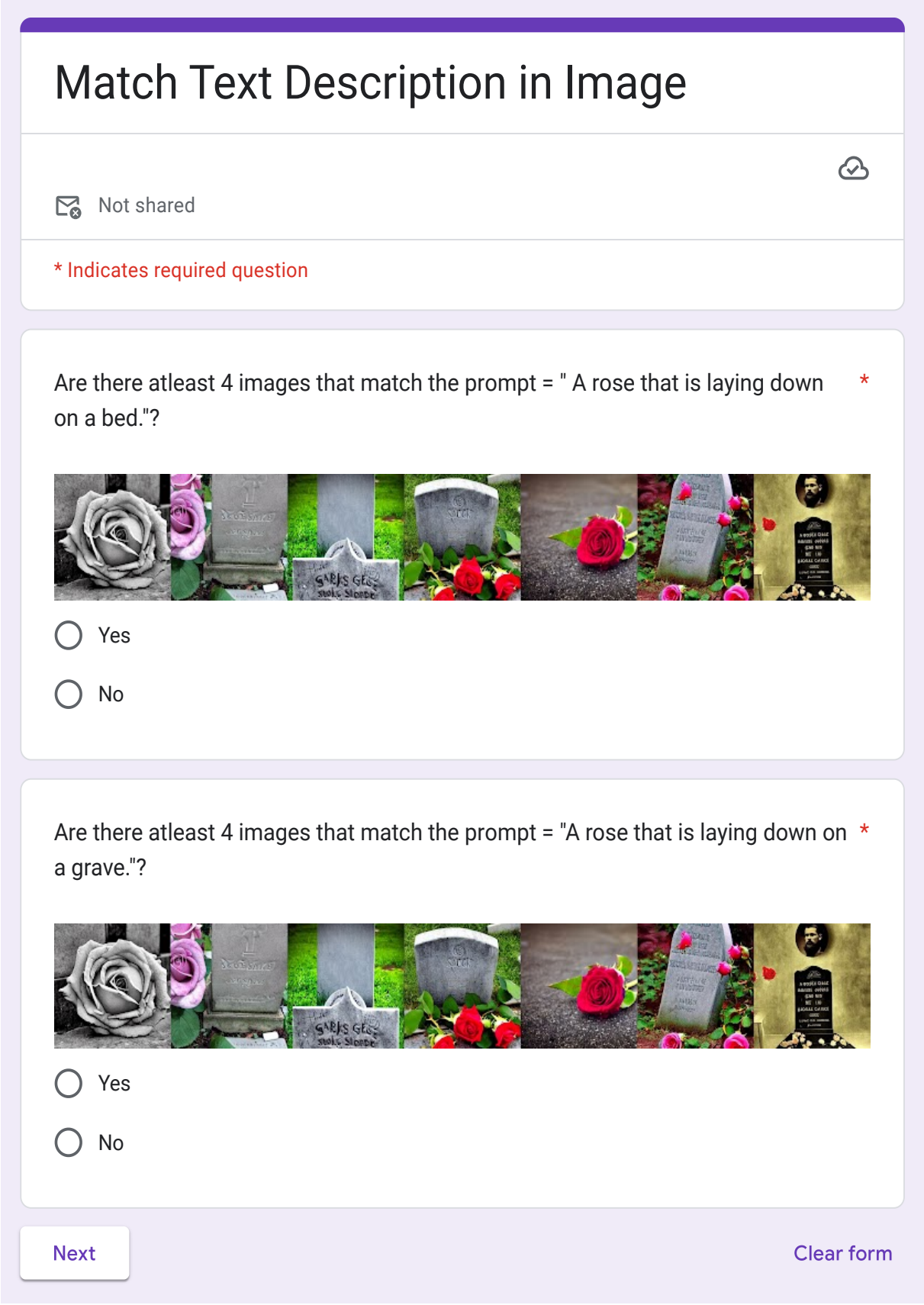

Figure 13: An overview of the human evaluation template.