Faithful Text-to-Image Generation by Selection

Anonymous ICCV Workshop submission

Paper ID 36

Abstract

Diffusion-based text-to-image (T2I) models can lack faithfulness to the text prompt, where generated images may not contain all the mentioned objects, attributes or relations. To alleviate these issues, recent post-hoc methods improve faithfulness without costly retraining, by modifying how the model processes the input prompt. In this work, we take a step back and show that large T2I models are more faithful than usually assumed, showing that faithfulness can be simply treated as a candidate selection problem instead. We introduce a straightforward pipeline that generates candidate images for a text prompt and picks the best one according to an automatic scoring system, leveraging existing T2I evaluation metrics. Quantitative comparisons alongside user studies on diverse benchmarks show consistently improved faithfulness over post-hoc enhancement methods, with comparable or lower computational cost.

1. Introduction

Text-to-Image (T2I) Generation has seen drastic progress with the advent of modern generative models, such as StableDiffusion [8]. However, even these large models appear to exhibit shortcomings when it comes to faithfully generating the input prompt, failing to correctly reflect attributes, counts, semantic object relations or even entire objects [7, 3, 2]. Thus, recent works [7, 3, 9] improve faithfulness by modifying the inference procedure. While resulting in a more expensive generation process qualitative results show superior faithfulness compared to the baselines, but often limited to specific prompt types.

Here we take a step back and investigate how unfaithful these diffusion models really are. Upon closer inspection, we observe that the faithfulness of Stable Diffusion is affected heavily by the random seed that determines the initial latent noise, suggesting that within the explorable latent space, faithful image generations are possible (c.f. for example image candidates in Fig. 1). We thus propose to improve the faithfulness in diffusion models by querying the model multiple times and automatically selecting the most suitable output. We denote this simple pipeline as `ImageSelect`. We utilize metrics from recently proposed text-to-image faithfulness benchmarks, TIFA [4] and ImageReward [10], to evaluate the faithfulness of our image generation. TIFA simplifies the text-to-image matching process into a set of Visual Question Answering tasks, solvable by existing pretrained models. ImageReward directly trains a a matching model trained on human preferences.

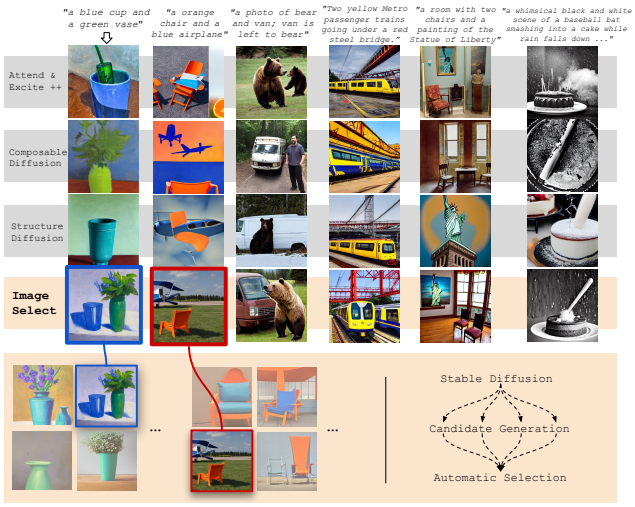

Figure 1. Our `ImageSelect` introduces automatic candidate selection to increase the faithfulness of a T2I generative model. We show that existing models are more faithful than assumed, and by simply querying them multiple times and selecting the most suitable image, we achieve significant improvements in T2I faithfulness, without requiring to explicitly adapt the generative process.

Experiments and user studies on existing benchmarks and a set of 1000 prompts aggregated from multiple datasets [1, 4, 6, 3]. show significant performance improvements against reference methods, with gains in faithfulness through automatic candidate selection consistently higher than that even achieved by changed model version generations (going for example from Stable Diffusion 1.4 to 2.1). The results showcase a simple, but large step forward for text-to-image faithfulness, and highlight our insights as a *crucial sanity check* for future work tackling the task of post-hoc enhancement of text-to-image generation.

2. ImageSelect: Faithfulness through Selection

Examples on Fig. 1 on various prompts over multiple seeds using vanilla Stable Diffusion indicate that faithful images *can be* generated, but are hidden behind a suitable selection of the starting latent noise. Thus, we introduce a simple, efficient and effective model that provides more faithful outputs for a given prompt by automatically selecting the most suitable candidates from multiple seeds.

Measuring Text-to-Image Alignment. For automatic selection, we can leverage existing T2I evaluation metrics. As *proof-of-concept*, we select TIFA and ImageReward.

TIFA Scores [4] evaluate T2I alignment via Visual-Question Answering (VQA). Specifically, given a text prompt y, a language model generates question-answer pairs $\mathcal{Q}(y) := \{(Q_i, A_i)\}_i$. The faithfulness score $\mathcal{F}$ of the generated image I is the ratio of questions that an off-the-shelf VQA model Ψ_{VQA}, *e.g.* [5], answered correctly,

$$\mathcal{F}_{\text{TIFA}}(I, y) = \frac{1}{|\mathcal{Q}(y)|} \sum_{(Q_i, A_i) \sim \mathcal{Q}(y)} \mathbb{I}\left[\Psi_{\text{VQA}}(I, Q_i) = A_i\right].$$

where $\mathbb{I}\left[\Psi_{\text{VQA}}(I, Q_i) = A_i\right]$ is 1 for correct answers. This strategy is interpretable and avoids any manual annotations.

ImageReward Scores [10] follow a completely different direction, performing end-to-end training on suitable data. In particular, [10] simply train a Multi-Layer Perception on top of BLIP [5] image and text features to regress 137k expert human preference scores on image-text pairs, with higher scores denoting higher levels of faithfulness. The resulting rating model $\mathcal{F}_{\text{ImageReward}}$ well-correlates with human ratings even on samples outside the training dataset.

Faithfulness through Selection. Both TIFA and ImageReward are only utilized as a benchmarking mechanism to evaluate current and future T2I methods on faithfulness. Instead, we showcase that these metrics can be easily used to supercharge the faithfulness of existing models without any additional retraining, by simply re-using them in a contrastive framework as a candidate selection metric. In particular, given a budget of N initialization starting points and a text prompt y, our associated generated output image I is thus simply given as

$$I_{\texttt{ImageSelect}}(y) = \arg\max_{n \in N} \mathcal{F}_{\texttt{ImageSelect}}\left(\mathcal{D}(\epsilon_\theta(\epsilon_n, T, y)), y\right)$$

where ϵ_θ denotes the text-conditioned denoising diffusion model in the latent space of the encoder-decoder model with decoder $\mathcal{D}$, total number of denoising iterations T, and initial latent noise $\epsilon_n \sim \mathcal{N}(0, 1)$ sampled anew for each n. We note that we use `ImageSelect` to refer to the use of any faithfulness measure s.a. $\mathcal{F}_{\text{TIFA}}$, $\mathcal{F}_{\text{ImageReward}}$, and highlight that this can be extended to any other scoring mechanism or combinations thereof. For a given selection method, we denote the respective `ImageSelect` operation as `TIFASelect` or `RewardSelect`.

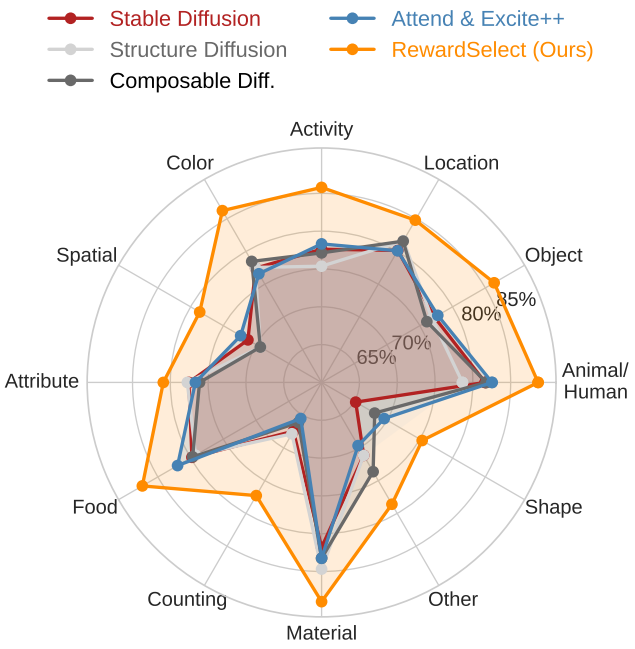

Figure 2. `RewardSelect` offers improved faithfulness across faithfulness categories as used in [4].

3. Experiments

Implementation Details. We take off-the-shelf Stable Diffusion 1.4 and 2.1 and evaluate them on the TIFAv1.0 [4] benchmark. To ensure that our results are as representative as possible and do not overfit to a particular type of prompt generation mechanism introduced in a benchmark, we aggregate prompts from HRS [1], TIFA (containing also captions from MS-COCO), and prompts utilized in [3]. We called this dataset `diverse-1k`, reporting more details in the supplementary. We consider the Structure Diffusion (StrD) [3] & Composable Diffusion (CD) [7] (both available only with Stable Diffusion 1.4) and the Attend-and-Excite (A&E) [2][1] methods as our baselines.

3.1. Benchmark comparison

Faithfulness on `diverse-1k`. We begin by evaluating the faithfulness of baselines on top of Stable Diffusion Version 1.4 (SD1.4) and 2.1 (SD2.1) on `diverse-1k`, using both TIFA and ImageReward scores. We use `RewardSelect` for TIFA scores, and vice versa `TIFASelect` for the ImageReward score evaluation, over a pool of 10 randomly generated images per prompt. Results in Fig. 3 show a **clear** increase in faithfulness of `ImageSelect` over all baselines and metrics. Across diverse prompts only A&E improves faithfulness of the base-

[1] We extend Attend and Excite for automatic use, details in the suppmat.

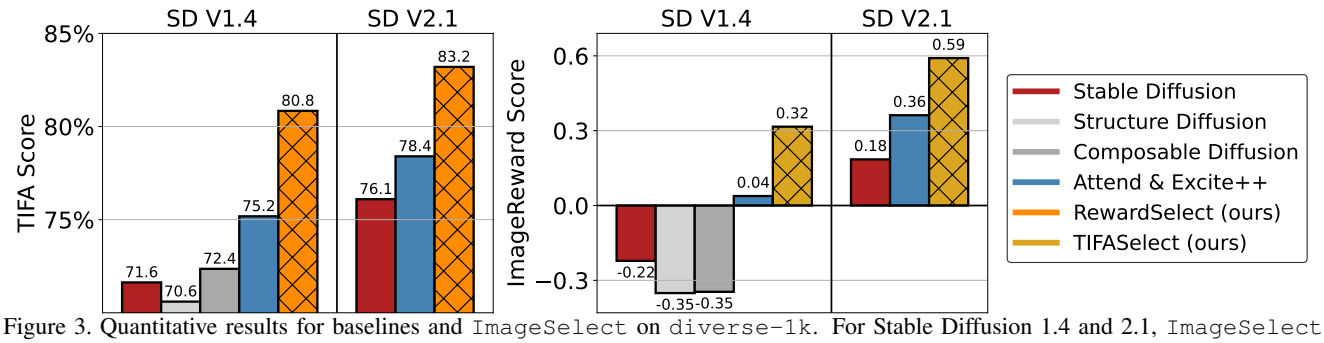

Figure 3. Quantitative results for baselines and `ImageSelect` on `diverse-1k`. For Stable Diffusion 1.4 and 2.1, `ImageSelect` outperforms all, irrespective of the selection and evaluation metric.

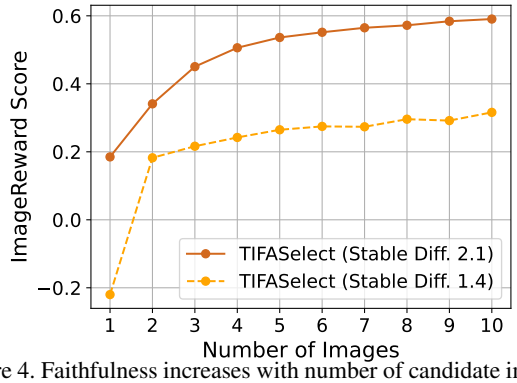

Figure 4. Faithfulness increases with number of candidate images.

line (e.g. +3.6% TIFA for SD1.4) while CD and StrD may have a detrimental effect, e.g. SD1.4 scoring +1% on TIFA and +0.12 on ImageReward w.r.t StrD. These results are surpassed by `ImageSelect`, which e.g. on SD1.4 achieves an impressive 80.4% - over 4*pp higher than the change from SD1.4 to SD2.1* gives in terms of text-to-image faithfulness. This fact is only exacerbated on the ImageReward score (−0.22 SD1.4, 0.18 SD2.1 and 0.32 for `TIFASelect`). These results indicate that suitable candidate selection can have a much higher impact on faithfulness than current explicit changes to the generative process.

Breakdown by Categories. We repeat our previous experiments on the original TIFAv1.0 benchmark [4], as it offers easy category-level grouping such as "counting", "shape" etc. For all methods and `RewardSelect` on SD1.4, we show results in Fig. 2. When breaking down the overall improvement in faithfulness into respective categories, the benefits of `ImageSelect` become even clearer, improving over every baseline across every category, with significant changes in categories such as "counting" (over 10*pp*) - a well-known shortcoming of T2I models, "spatial (relations)" (7*pp*) or "object (inclusion)" (8*pp*). Note that these improvements are not a result of potential overfitting to the evaluation metric, as the scoring approaches are entirely different (VQA vs human preferences).

Faithfulness vs Number of Candidate Images. We visualize the relation between T2I faithfulness and the number of candidate images taken into consideration in Fig. 4, as measured by the ImageReward score on `diverse-1k`. Our experiments show a drastic improvement with already two candidates, raising the faithfulness of SD1.4 to that of SD2.1. Going further, we find monotonic improvements, but with diminishing returns for larger candidate counts. Notably, a small number of candidate images (e.g. 4) is already sufficient to surpass all baselines. This is not caused by any single seed being more effective, as we find all seeds to behave similarly (77.9% to 78.5% for 10 seeds on TIFAv1.0), *but rather the per-prompt candidate selection.*

3.2. User Study

Since image generation is subjective, we expand our quantitative studies with extensive human evaluations. For every `diverse-1k` prompt, we generate images using all baselines (CD [7], StrD [3] and A&E++) as well as `RewardSelect` and `TIFASelect` on SD1.4, and on SD2.1 for all `ImageSelect` variants and A&E++. Using the generated images, we set up a comparative study following the layout shown in supplementary. Voluntary users interact with the study through a webpage, and are tasked to select the most faithful generation between the output of either a baseline method or an `ImageSelect` variant. In total, we collect 5093 human preference selections, distributed over 68 unique users and each comparative study.Results are shown in Fig. 5, where we also compare `RewardSelect` and `TIFASelect` directly.

We find a clear preference in faithfulness for images generated by `ImageSelect`, particularly `RewardSelect`, with `ImageSelect` chose in parts twice (e.g. +126.3% for `TIFASelect` vs CD on SD1.4) or even three times more often (e.g. +207.9% on `RewardSelect` vs. StrD on SD1.4). Even against A&E++ on the improved SD2.1, `RewardSelect` has a 84.4% higher chance to be chosen. In general, `RewardSelect` is better aligned with human insights on faithfulness, and better suited as a candidate selector w.r.t. `TIFASelect` (Fig. 5i-j). This indicates that a model trained to mimic human preferences might work better as a selection metric than one that looks for faithfulness as a numerical metric, weighing every semantic aspect

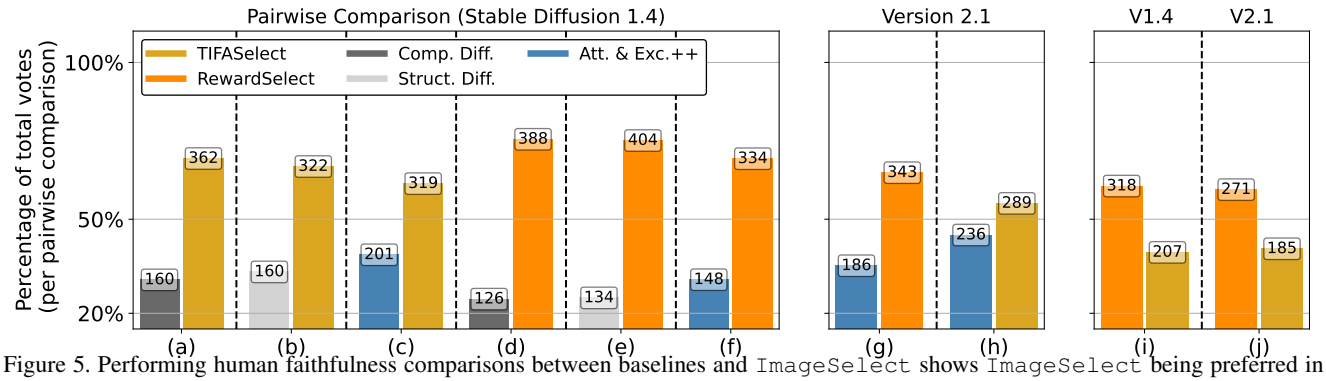

Figure 5. Performing human faithfulness comparisons between baselines and `ImageSelect` shows `ImageSelect` being preferred in the majority of cases for prompts from `diverse-1k`.

equally. Regardless of the variations in `ImageSelect`, our user study provides compelling evidence that automatic candidate selection is a highly promising approach for post-hoc faithfulness in large-scale pretrained T2I diffusion models, especially when compared to existing approaches explicitly adapting the generative process.

3.3. Qualitative Examples

We also show additional qualitative examples to illustrate the successes of `ImageSelect` in Fig.X of the supplementary, which captures both simple and complex prompts well, particularly compared to other methods that struggle with the issues of catastrophic neglect [2], attribute binding [3], and incorrect spatial arrangement. For instance, `ImageSelect` is able to capture the objects and spatial relations in prompts like *"three small yellow boxes on a large blue box"* or *"Two men in yellow jackets near water and a black plane."*, while also faithfully rendering creative prompts like *"an oil painting of a cat playing checkers."*. Other methods perform worse in comparison, often missing objects entirely or generating objects with an incorrect spatial arrangement or false association of attributes (c.f. *"A green chair and a red horse"*).

4. Conclusion

In this work, we both highlight and leverage the dependence of faithfulness on initial latent noises in diffusion-based T2I to introduce `ImageSelect`. By viewing the problem of post-hoc faithfulness improvements as a candidate selection problem, we propose a simple pipeline, in which an automatic scoring system selects the most suitable candidate out of multiple model queries. In doing so, we are able to significantly improve faithfulness, particularly when compared to recent approaches adapting the diffusion process directly. We validate the success of `ImageSelect` with quantitative experiments and user studies on diverse test benchmarks, showcasing significant gains in faithfulness. Overall, we hope that our work serves as a useful practical tool and a valuable sanity check for future work on post-hoc enhancement of text-to-image generation.

References

[1] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. *arXiv:2304.05390*, 2023. 1, 2

[2] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *SIGGRAPH*, 2023. 1, 2, 4, 5

[3] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*, 2023. 1, 2, 3, 4, 5

[4] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv:2303.11897*, 2023. 1, 2, 3, 5

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 5

[7] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 1, 2, 3, 5

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[9] Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis, 2023. 1, 5

[10] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv:2304.05977*, 2023. 1, 2

A. Implementation Details

We take off-the-shelf Stable Diffusion 1.4 and 2.1 and evaluate them on the TIFAv1.0 [4] benchmark - consisting of prompts from MS-COCO and other sources that benchmark T2I generation for more creative tasks - and our `diverse-1k` prompts list. We consider the Structure Diffusion (StrD) [3] & Composable Diffusion (CD) [7] (both available only with Stable Diffusion 1.4) and the Attend-and-Excite (A&E) [2] methods as our baselines. While StrD can be applied directly, CD requires us to split the prompts and join them together using the "AND" operator.

A&E requires a user to manually select tokens the model should attend to. We modify this to work automatically by selecting categories from MS-COCO, as well as utilizing NLTK to determine nouns which cannot be treated as either a verb or adjective. For any prompt for which the above protocol provides no target tokens, we continuously relax the constraints over the nouns. In limit cases where nothing suitable is selected, A&E defaults back to the original Stable Diffusion it extends. We denote A&E equipped with this formalism as *Attend-and-Excite++* (A&E++). We find that on normal prompts or those qualitatively studied in the original paper [2], our protocol comes very close to the generations reported in [2].

B. Additional Results

Comparison to Ground Truth Faithfulness. To provide a better reference for the quantitative change in performance, we also evaluate on the MS-COCO captions used in [4], for which ground truth images exist. Using `RewardSelect` and the TIFAScore for evaluation, we report results in Tab. 1. While clearly outperforming baseline methods, we also see `RewardSelect` matching ground truth TIFA faithfulness scores of true MS-COCO image-caption pairs (89.85% versus 89.09%). While attributable to increases in measurable faithfulness through `ImageSelect`, it is important to note both the noise in ground truth captions on MS-COCO [6] and a focus on a particular prompt-style (descriptive natural image captions - hence also our use of `diverse-1k` for most of this work). Still, these ground truth scores provide strong support for the benefits of candidate selection as a means to increase overall faithfulness.

Computational Efficiency. While Stable Diffusion takes 5 seconds to generate a single image (NVIDIA 2080Ti), Attend-and-Excite requires 30 with double the memory requirements. Other recent methods such as Space-Time-Attention [9] can require nearly five times the VRAM and over 10 minutes. Thus even from a computational perspective, there is a clear benefit of leveraging simple candidate selection through `ImageSelect`, and generating as many candidates as possible within a computational budget. Finally, the process of producing respective images for a prompt is parallelizable, and directly benefits from extended GPU counts even on a single-prompt level.

Table 1. Faithfulness comparison with our `RewardSelect` (`RS`) using the TIFA-score on the ground-truth MS-COCO image-caption pairs. Our `RS` closes the gap with GT=89.09% in faithfulness.

V1.4	SD	A&E++	RS
	82.69%	82.04%	88.69%
V2.1	SD	A&E++	RS
	85.28%	85.87%	**89.85%**

Figure 6. *Additional Examples* highlighting favorable faithfulness of `ImageSelect` (rightmost) compared to Attend-and-Excite++, Composable Diffusion [7] and Structure Diffusion [3].

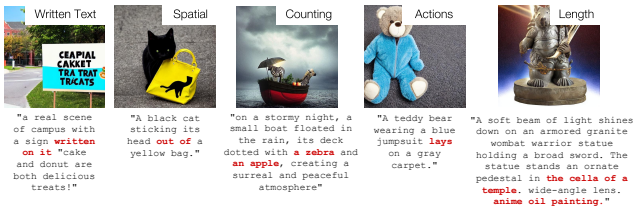

Figure 7. *Qualitative Failure Cases.* Despite significantly improving faithfulness, `ImageSelect` can not fully account for fundamental shortcomings. Details on faithfulness categories, see e.g. Fig. 2.

Qualitative Examples and Limitations. We also show additional qualitative examples to illustrate the successes of `ImageSelect` in Fig. 6, which captures both simple and complex prompts well, particularly compared to other methods that struggle with the issues of catastrophic neglect [2], attribute binding [3], and incorrect spatial arrangement. For instance, `ImageSelect` is able to capture the objects and spatial relations in prompts like ``three small yellow boxes on a large blue box'' or ``Two men in yellow jackets near water and a black plane.'', while also faithfully rendering creative prompts like ``an oil painting of a cat playing checkers.''. Other methods perform worse in comparison, often missing objects entirely or generating objects with an incorrect spatial arrangement or false association of attributes (c.f. ``A green chair and a red horse'').

Limitations. We illustrate failures in Fig. 7. While `ImageSelect` significantly improves faithfulness, it can still struggle with challenges inherent to the underlying model such as rendering text, exact spatial relations, counting or very long prompts. However, due to its applicability to any T2I model, these shortcomings can be addressed by jointly tackling fundamental issues in vision-language models and leveraging orthogonal extensions such for character generation.