A Appendix

A.1 Additional Attention Statistics

In this section, we provide a set of examples that illustrates the relationship between the distribution of the values of the attention weights and the semantic accuracy of the images generated by the model. The experiment setup is identical to that in Section 3 except for the selection of diffusion step and cross-attention layer. In this setup, we chose to plot the maximum value of each attention head in the first up-sampling layer at the final diffusion step. From Figure 1 and Figure 2, we observe a positive relationship between the manifestation of a token in the generated image and the values of its attention weight. To illustrate, consider the first row of Figure 1. The attention values of the token "camera" is suppressed in the first four images (from the left), which corresponds to the absence of the camera in the generated image. However, in the final image in the first row, the attention values of the token "camera" is significantly higher, which corresponds to the presence of a camera in the generated image.

A.2 Additional Evaluation Metrics on Different Diffusion Models

In this section, we present additional evaluation results on all 6 approaches on 4 different Stable Diffusion models, namely Stable Diffusion 1.4, 1.5, 2 and 2.1. We evaluated the approaches on two datasets, namely the COCO dataset and the A&E dataset. We present results on two additional metrics, namely the detection success rate, see Table 1, and the LPIPS score, see Table 2.

Table 1: Detection success rate (Det.Rate) evaluation of different methods across diffusion models.

Methods	COCO Dataset				A&E Dataset			
	SD1.4	SD1.5	SD2	SD2.1	SD1.4	SD1.5	SD2	SD2.1
Original	51.6%	52.1%	55.9%	59.2%	34.8%	34.6%	46.7%	47.1%
ComposableDiffusion	36.8%	36.2%	27.0%	29.2%	24.9%	23.8%	19.4%	20.4%
SyntaxGeneration	61.7%	62.0%	63.8%	64.3%	57.2%	56.7%	61.4%	61.2%
DenseDiffusion	61.9%	63.5%	64.0%	65.2%	–	–	–	–
AttendAndExcite	55.8%	57.2%	64.5%	69.1%	46.0%	48.1%	60.8%	62.6%
Ours	**68.0%**	**68.7%**	**72.5%**	**73.2%**	**58.7%**	**59.3%**	**66.2%**	**66.8%**

We further run our method on Stable Diffusion 2 on 2 specific Text-to-Image benchmarks, T2I-CompBench [3] and TIFA [2]. Table 3 shows BLIP-VQA score from T2I-CompBench applied to our generation result. Attention Regulation similarly achieves significant improvement over existing methods. Table 4 shows the TIFA score of our generations against the baseline methods on a selected subset of the TIFA dataset. Our method performs marginally better on TIFA

Table 2: LPIPS score evaluation of different methods across diffusion models.

Methods	COCO Dataset				A&E Dataset			
	SD1.4	SD1.5	SD2	SD2.1	SD1.4	SD1.5	SD2	SD2.1
Original	-	-	-	-	-	-	-	-
ComposableDiffusion	0.580	0.579	0.606	0.613	0.584	0.584	0.598	0.579
SyntaxGeneration	0.468	0.469	0.452	0.468	0.432	0.427	0.423	0.436
DenseDiffusion	0.768	0.772	0.729	0.744	–	–	–	–
AttendAndExcite	0.462	0.468	0.393	0.410	0.642	0.639	0.508	0.506
Ours	0.495	0.496	0.508	0.546	0.543	0.544	0.666	0.538

Table 3: T2I-Benchmark on SD2.

	COCO Dataset	A&E Dataset
Original	0.564	0.571
ComposableDiffusion	0.311	0.410
SyntaxGeneration	0.613	0.739
DenseDiffusion	0.611	-
AttendAndExcite	0.608	0.689
Ours	**0.647**	**0.763**

score because questions in TIFA include many questions beyond the claim of this work, such as counting, orientation, and activities.

A.3 Experiment on Guidance Scale

Conditional Diffusion models generally apply the guidance in the form of Classifier-Free Guidance [1] to improve the semantic fidelity of generated images. We generate images on the COCO Dataset using Stable Diffusion 2 and evaluate the generated images on CLIP Score. Figure 3 shows generated images with different guidance scales. Our visual results show that increasing the guidance scale has limited influence in generating under-represented objects. The results in Table 5 show that while Guidance Scales can be used to improve the semantic fidelity of Conditional Diffusion models, they alone are insufficient to improve performance significantly. For reference, our attention regulation method with Stable Diffusion 2 scores 0.337 on the CLIP Score evaluation on the same dataset.

A.4 Additional Visual Comparisons of Images

In this section, we provide additional visual comparisons of generated images between the original Stable Diffusion, Attend-And-Excite, Composable Diffusion, Syntax Generation, Dense Diffusion and our attention regulation method. Figure 4, Figure 5 showcase an uncurated set of 3 images per prompt across multiple prompts covering multiple settings and objects. All images are generated with seed 42.

Table 4: TIFA score on SD2.

Original	CompDiff	SynGen	A&E	Ours
0.824	0.643	0.816	0.839	**0.846**

Guidance Scale	5	7.5	10	12.5	15	17.5
CLIP Score	0.325	0.328	0.329	0.329	0.330	0.329

Table 5: CLIP Score of images generated by Stable Diffusion 2 across a range of Guidance Scales.

A.5 Additional Ablation Results

In this section, we provide additional visual comparisons of generated images between the original Stable Diffusion and our attention regulation method with different hyperparameters. We showcase the generated images on three main hyperparameters, namely the cross attention layers, the diffusion steps and the β regularaisation term. All images are generated with seed 42. Figure 6 showcases the variation in the generated image as the cross attention layers on which attention regulation is performed varies. Figure 7 showcases the variation in the generated image as the diffusion steps on which attention regulation is performed varies. Figure 8 showcases the variation in the generated image as the value of the β regularisation term varies.

Moreover, in our experiment setup, with reference to the notation used in Section 3.3, we set $\sigma = \frac{w}{16}$, $\eta = 100$, $\kappa = 0.25$, $\lambda = 0.99$. Ablation results of these parameters are not included as their impact is less prominent as compared to the main hyperparameters that have been discussed.

References

1. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
2. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897 (2023)
3. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. arXiv preprint arXiv:2307.06350 (2023)

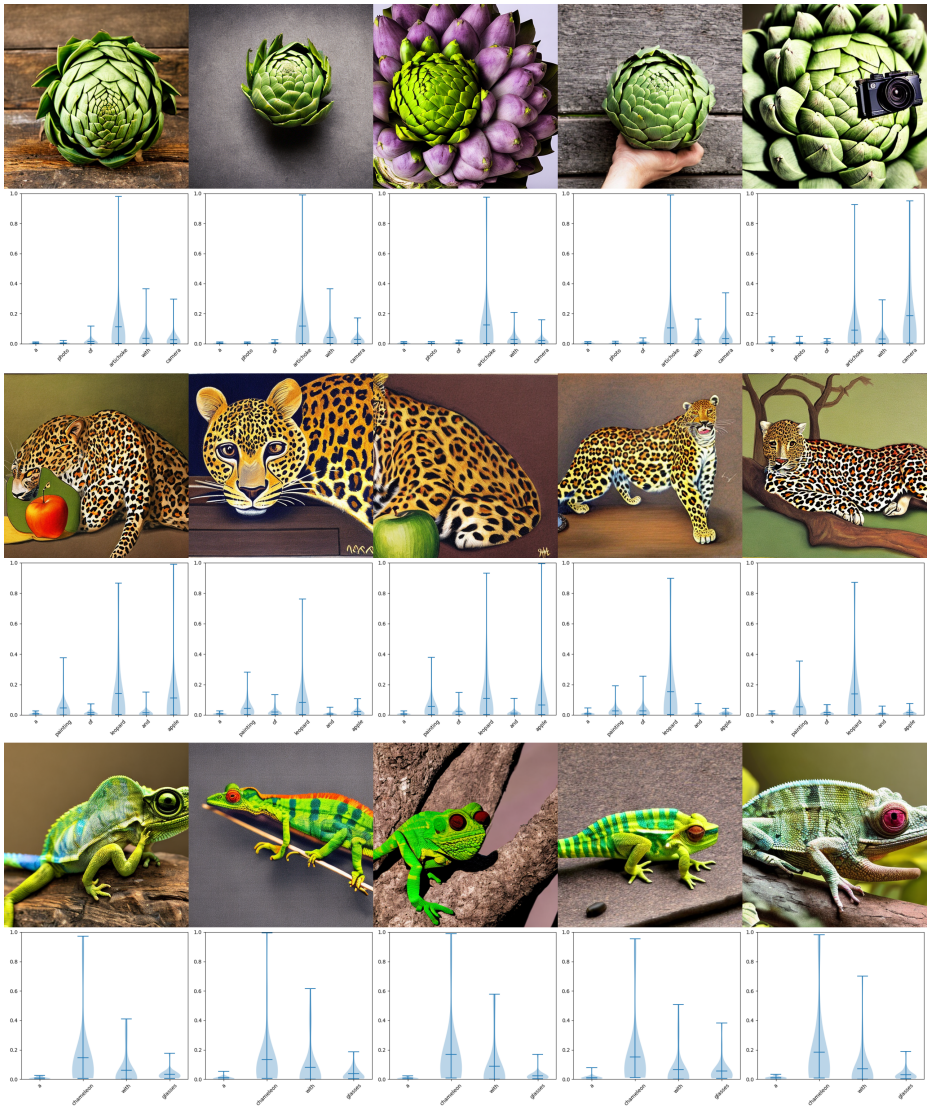

Fig. 1: Generated Images and their Corresponding Attention Plots. In the first row, the camera appears only when the attention values of the "camera" token match that of the "artichoke" token. In the second row, the apple appears only when the attention values of the "apple" token match that of the "leopard" token. In the third row, the attention values of the "glasses" token is significantly lower than the "chameleon" token, which corresponds to the absence of glasses in the generated images.

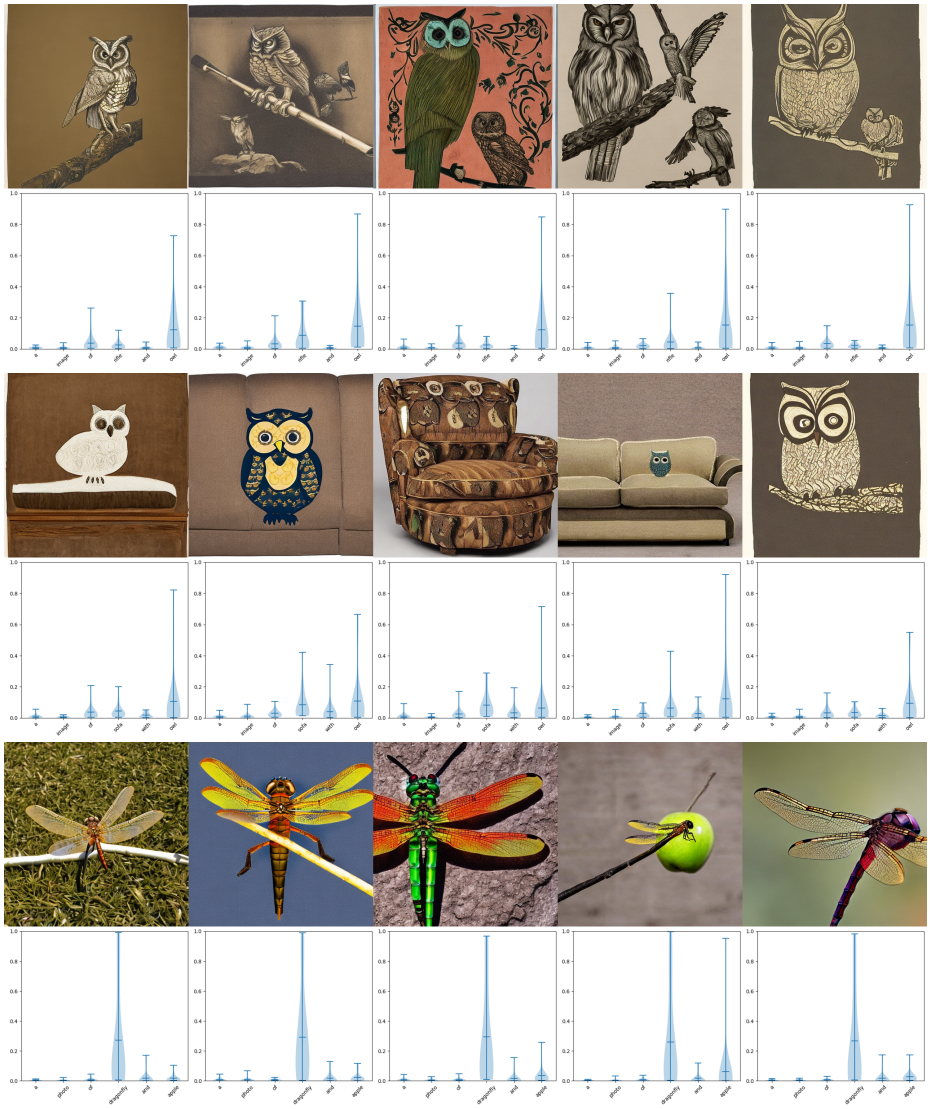

Fig. 2: Generated Images and their Corresponding Attention Plots. In the first row, the attention values of the "rifle" token is significantly lower than the "owl" token, which corresponds to the absence of rifle in the generated images. In the second row, the sofa appears when the attention values of the "sofa" token follows a similar distribution as that of the "owl" token. In the third row, the apple appears only when the attention values of the "apple" token match that of the "dragonfly" token.

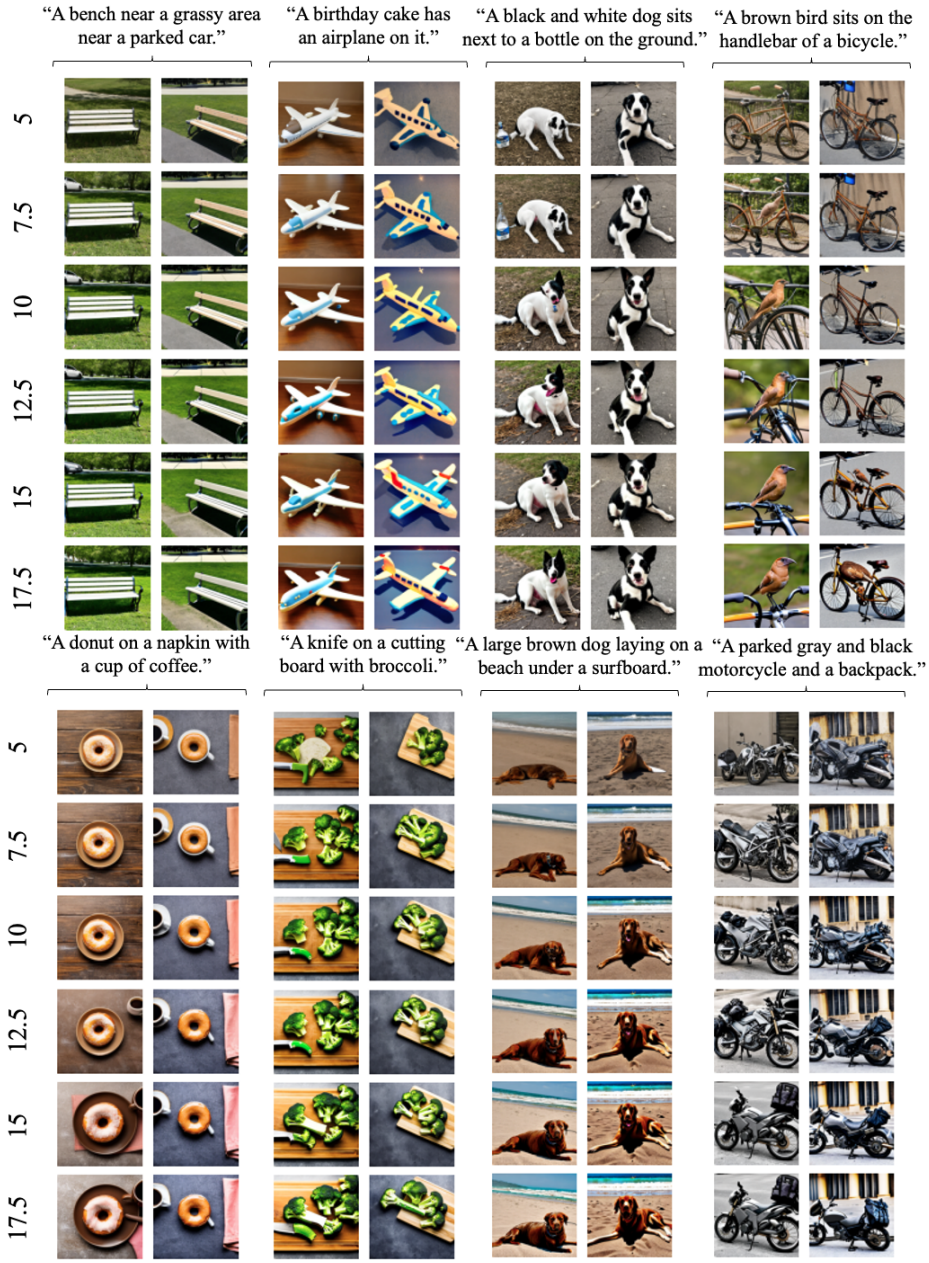

Fig. 3: Visual Comparison of Images across Different Guidance Scales. For each prompt, we show two generated images where we use the same set of seeds for all six guidance scale.

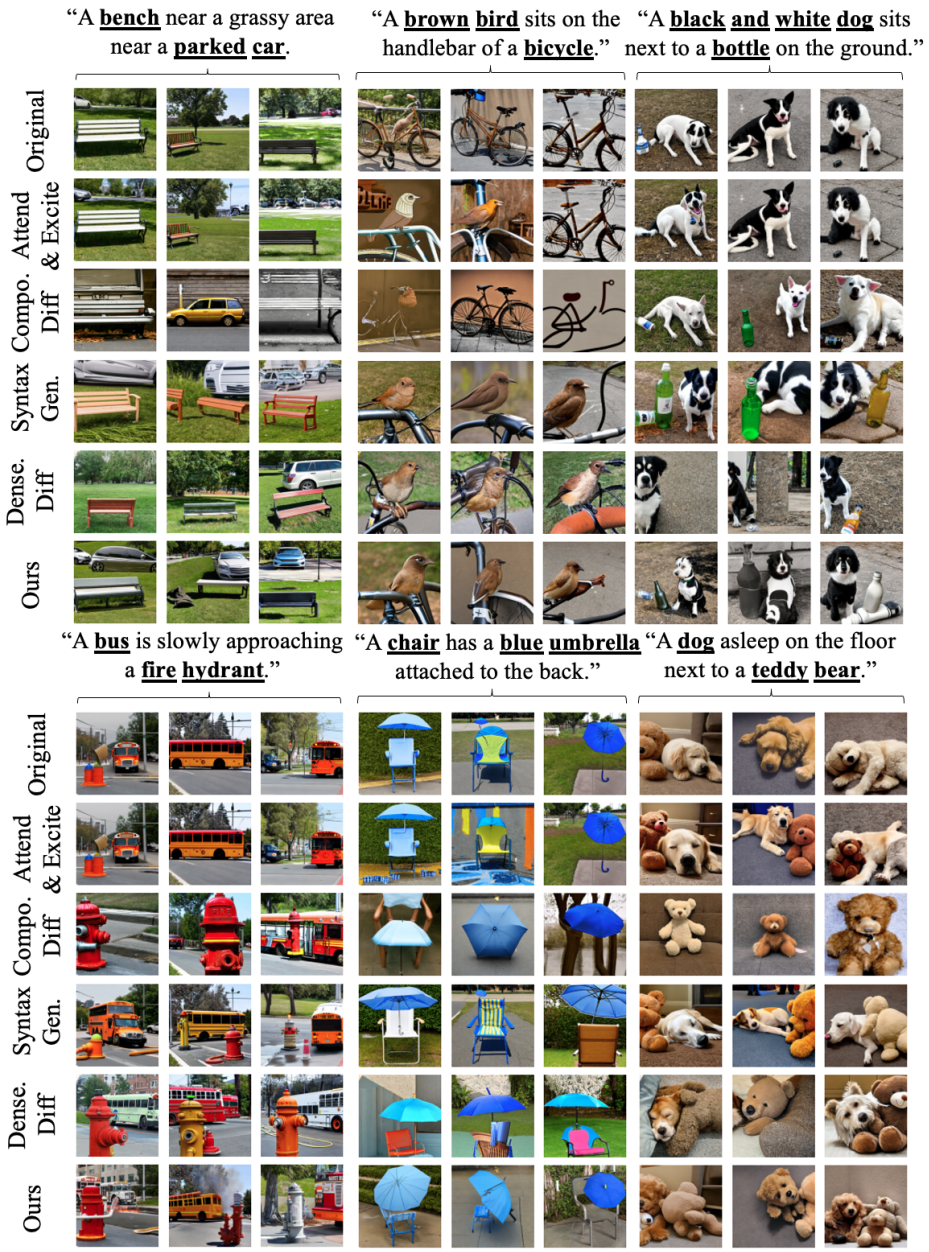

Fig. 4: Additional Visual Comparison of Images across Different Approaches. For each prompt, we show three generated images where we use the same set of seeds for all six approaches. The subject tokens optimised by our approach is in bold and underlined.

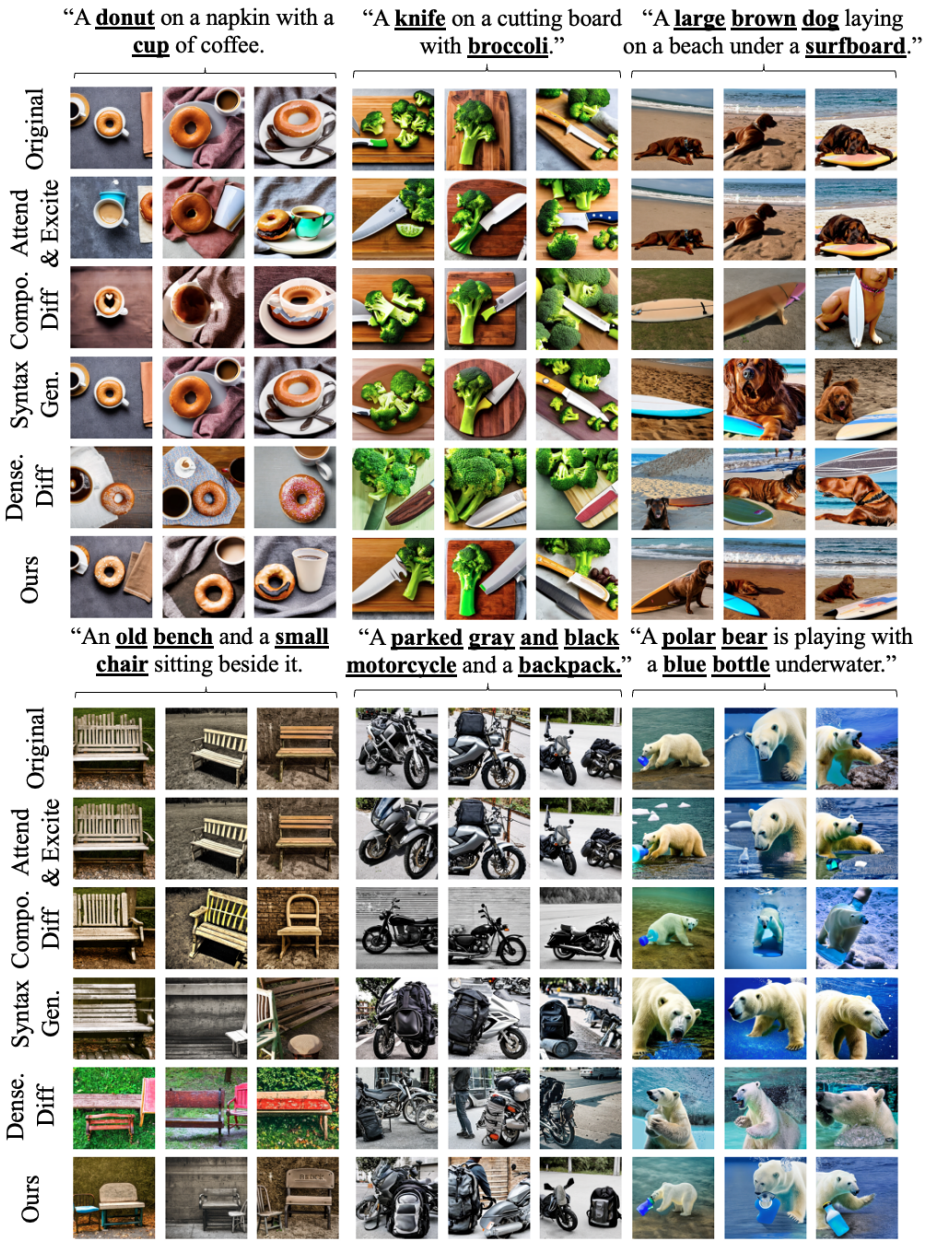

Fig. 5: Additional Visual Comparison of Images across Different Approaches. For each prompt, we show three generated images where we use the same set of seeds for all six approaches. The subject tokens optimised by our approach is in bold and underlined.

Fig. 6: Additional Visual Comparison of Images for Layer Ablation. Across all three prompts, the results show that editing 2 layers is sufficient to resolve catastrophic neglect.

Fig. 7: Additional Visual Comparison of Images for Diffusion Steps Ablation. Across all three prompts, the results show that editing up to 25 steps is sufficient to resolve catastrophic neglect.

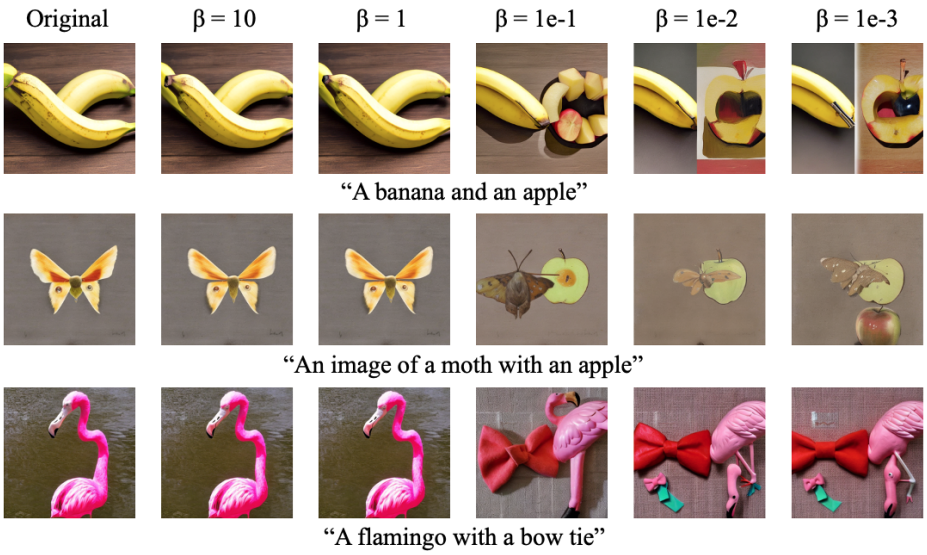

Fig. 8: Additional Visual Comparison of Images for β Ablation. Across all three prompts, the results show that the value of 0.1 strikes a reasonable balance between over-editing and under-editing.