TokenCompose: Text-to-Image Diffusion with Token-level Supervision

Supplementary Material

1. Pipeline Setup

We provide detailed data generation, training and inference settings for the TokenCompose pipeline setup in Table 1. Unless otherwise specified, experiments are run based on this set of settings. For settings that are not reported in Table 1, we follow the default values provided by their respective codebases.

Setting	Value
Data Generation	
CLIP Model	`ViT-L/14@336px` [3, 15]
POS Tagger	`flair/pos-english` [1]
Target POS Tags	NN / NNS / NNP / NNPS
G. DINO Model	`groundingdino_swint_ogc` [2, 11, 12]
G. DINO Box Thres.	0.25
G. DINO Text Thres.	0.25
SAM Model	`sam_hq_vit_h` [3, 9, 10]
Training	
Resolution SD 1.4	512×512
Resolution SD 2.1	768×768
Image Processing	Center Crop + Resize
Batch Size	1
Grad. Accum. Steps	4
Grad. Ckpting.	`True`
Train Steps SD 1.4	24000
Train Steps SD 2.1	32000
Learning Rate	5e-6
LR Scheduler	Constant
LR Warmup	`False`
λ for $\mathcal{L}_{token}$	1e-3
γ for $\mathcal{L}_{pixel}$	5e-5
$\mathcal{L}_{token}$ and $\mathcal{L}_{pixel}$	all layers in U_{Mid} and U_{D}
Clf-Free Guidance [7]	`False`
Inference	
Resolution SD 1.4	512×512
Resolution SD 2.1	768×768
Timestep Scheduler	`PNDMScheduler`
# Inference Steps	50
Clf-Free Guidance [7]	`True`
Guidance Scale	7.5

Table 1. **Model choices and settings for data generation, training, and inference.** We provide a comprehensive list of pipeline details for data generation (*e.g.,* caption selection, noun extraction and segmentation map generation), training (*e.g.,* finetuning the Stable Diffusion model), and inference (*e.g.,* evaluation of our finetuned Stable Diffusion model).

2. Conditional Downstream Metrics

As multi-category instance composition serves as a prerequisite for successful downstream text-conditioned compositional generation, we conjecture that the improvements in downstream metrics are improved by higher chances of generating all instances mentioned in the prompt. In this section, we provide an additional object accuracy metric in Table 2, and show that TokenCompose improves object accuracy in all downstream metrics being evaluated. Furthermore, in Figure 1, we demonstrate how better multi-category instance composition capabilities can improve these metrics by plotting the improvement curve from ***(1)*** a frozen Stable Diffusion to ***(2)*** a Stable Diffusion finetuned with only $\mathcal{L}_{LDM}$ objective, and finally to ***(3)*** a Stable Diffusion model finetuned with both $\mathcal{L}_{LDM}$ and our grounding objectives conditioned and unconditioned on successful generation of all instances from the compositional prompts.

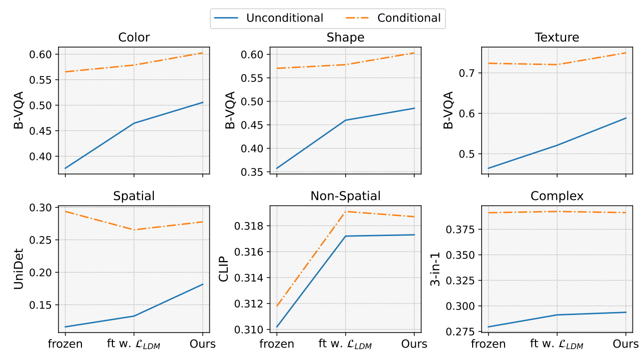

Figure 1. **Improvement comparison of unconditional and conditional downstream compositional metrics.** We illustrate that a significant margin of downstream compositional metrics is improved due to enhanced capabilities of multi-category instance composition. In this figure, we calculate the evaluation metrics from T2I-CompBench [8] by conditioning on the successful generation of all instances mentioned in the prompts, and compare the amount of quantitative improvement with the same metrics that are *not* conditioned on successfull generation of all instances.

From the results, we observe that the improvements in all attribute binding metrics (*e.g.,* color, shape, texture) and the majority of object relation metrics (*e.g.,* spatial, complex) are much more significant in the unconditional case than in the conditional case. The only downstream benchmark where the difference in improvement is insignificant is the *non-spatial* compositionality. We believe that this insignificance can be explained by two factors: (1) lowest amount of improvement in object accuracy for this specific downstream benchmark, as shown in Table 2 and (2) relatively low correlation between automatic scores (*i.e.,* CLIP Score [6, 15]) and human ratings among all compositional benchmarks from the T2I-CompBench [8]; this indicates that the evaluation model may have a comparably weak discriminative capability for this specific task.

Model	Color	Shape	Texture	Spatial	Non-Spat.	Complex
SD 1.4						
frozen	47.40	25.33	15.27	19.40	45.10	26.60
ft. w. $\mathcal{L}_{LDM}$	55.70	28.49	18.95	27.05	47.33	29.33
Ours	**62.92**	**32.95**	**25.59**	**33.30**	**48.94**	**32.10**

Table 2. **Object accuracy in downstream compositional metrics.** We calculate the object accuracy metric (*i.e.,* success rate of generating all instances in the prompt based on a detection model [13]) for each of the compositional benchmarks from T2I-CompBench [8].

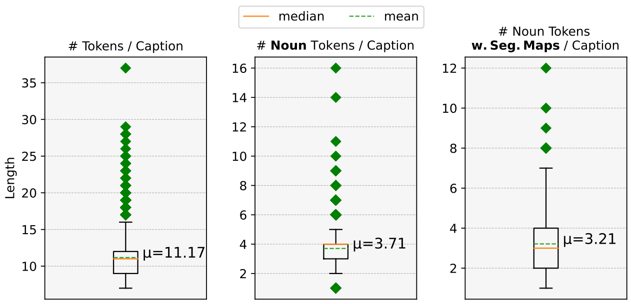

Figure 2. **Approximations of the average number of tokens with grounding objectives per training prompt.** From left to right, we show the distribution of number of tokens per caption, number of noun tokens per caption, and number of noun tokens that have generated segmentation maps per caption.

3. More Examples

Multi-category Instance Composition. We provide more visualizations in Figure. 3 to illustrate capabilities of TokenCompose in multi-category instance composition. In addition to being able to generate multiple categories of instances successfully, we show that object affordance (*e.g.,* attach, sit, support, *etc*) can be maintained, which indicates that TokenCompose is able to implicitly learn the basic "physical rules" via the object token and segmentation consistency constraints.

Downstream Applications. We show in Figure. 4 positive impacts in downstream applications. As expected, TokenCompose shows its effectiveness in tasks beyond text-to-image generation.

Benchmark. MultiGen benchmark plays a crucial role in evaluating multi-category instance composition capabilities of text-to-image models. To offer a more intuitive sense in this compositional task, we visualize the generated images from various prompts in this benchmark (*e.g.,* MG COCO Instances and MG ADE20K Instances). We use the same latent for each comparison for fairness. The images are presented in Figure 7.

4. Grounded COCO Dataset

As seen in Figure 2 from the main paper, we adopt a POS Tagger and the Grounded SAM [1, 10, 11] to extract binary segmentation maps from noun tokens from the image-text pair dataset. We aim to expand the visualizations of the generated data in Figure 8. Each row and column represents a single data that the model is trained with. From the left to the right of the data are the caption, the input image, and the grounded binary segmentation maps.

The bold and underlined **text** in the caption represents noun tokens captured by the POS tagger where their corresponding segmentation maps are extracted from the Grounded SAM. Italicized and red *text* in the caption represents noun tokens captured by the POS tagger but does *not* have segmentation maps extracted from the Grounded SAM due to the model not being able to locate the objects.

Grounded segmentation maps on the right are paired with their correpsonding tokens. For tokens with a green background, their segmentation maps successfully capture the aligned contents in the image, whereas for a small fraction of tokens with an orange background, their segmentation maps do not capture the aligned contents in the image.

We also provide reference on approximations (*i.e.,* word split by space) of (1) average number of tokens per caption, (2) average number of *noun tokens* per caption, and (3) average number of noun tokens that *have their corresponding segmentation maps* per caption in Figure 2. As shown in the figure, our training dataset contains an average of 3.71 noun tokens overall and 3.21 noun tokens that have their corresponding segmentation maps.

5. Analysis

Attention Visualizations. To gain a better understanding of how incorporating grounding objectives into text-to-image models during training affects cross-attention maps for image reconstruction [14] & generation tasks at inference time, we provide token-level cross-attention map visualizations on three axes: ***(1)*** different cross-attention layers with various resolutions (Figure 9); ***(2)*** different heads of the multi-head cross-attention (Figure 10; and ***(3)*** different timesteps during the denoising process (Figure 11).

Multi-category Instance Composition Success Rate. Given a set of compositional prompts, we calculate the count of each category that appears in these prompts along with images where the instance(s) of this category is/are detected by a detection model [13]. We divide the number of images that contain a specific category of the instance by the number of prompts that contains this category to acquire the success rate and report the numbers in Figure 12.

Failure Cases. We provide generated and detected samples for categories with low and high success rates in Figure. 5. We believe that one explanation for the poor performance categories have variants and viewpoints with drastic visual differences, so learning and generating them is harder. Further, we aggregate patterns of common failures in our multi-category instance composition in Figure. 6. We believe that visual commonsense reasoning aspect of the generative model would be an area of improvement.

Figure 3. **More samples** in multi-category instance composition.

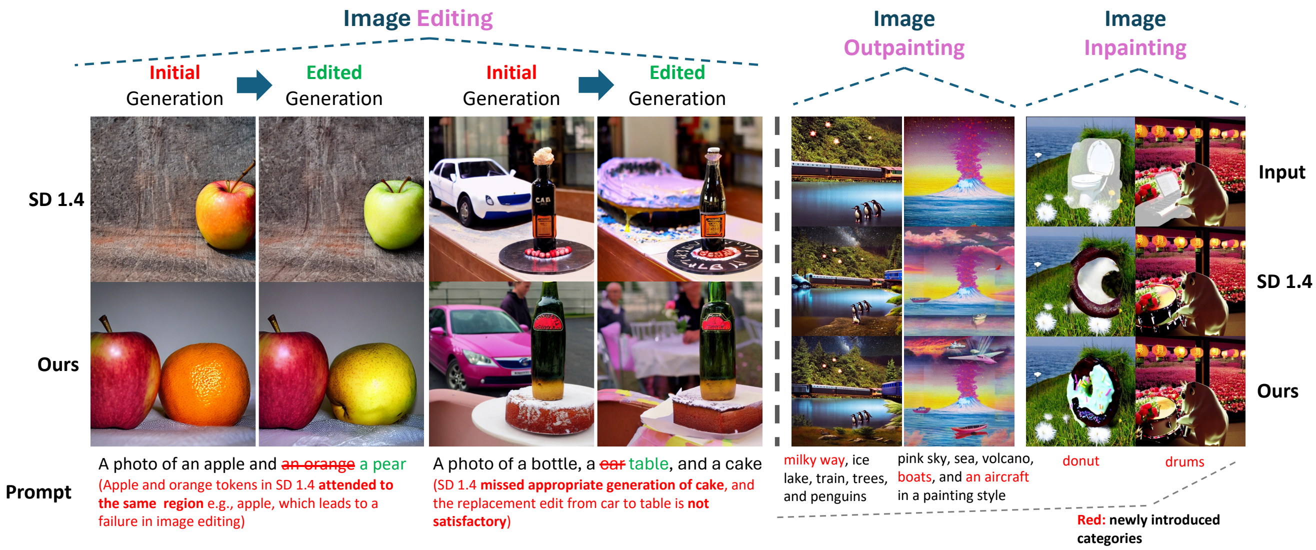

Figure 4. **Downstream applications** in prompt-to-prompt [5] image editing and zero-shot outpainting and inpainting.

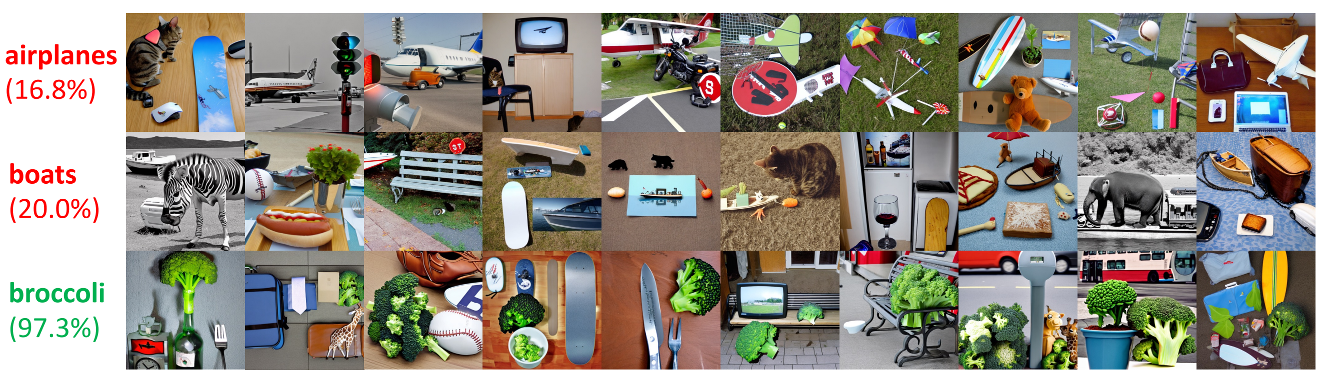

Figure 5. Categories with poor & strong performance

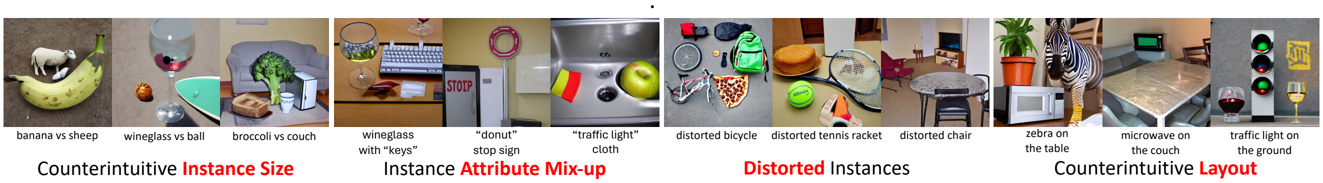

Figure 6. **Failure case analysis** in multi-category compositionality.

Figure 7. **Sampled images with prompts from our MULTIGEN benchmark.** To facilitate understanding of our multi-category instance composition benchmark as well as further qualitative comparisons among different baselines, we provide sampled images from MULTIGEN with COCO instances as well as ADE20K instances.

Figure 8. **Grounding dataset visualization.** We provide visualizations of 50 selected image-text pairs (number of noun tokens with segmentation maps ≥ 5) and their corresponding token-level binary segmentation maps from the training data to facilitate understanding of our training dataset.

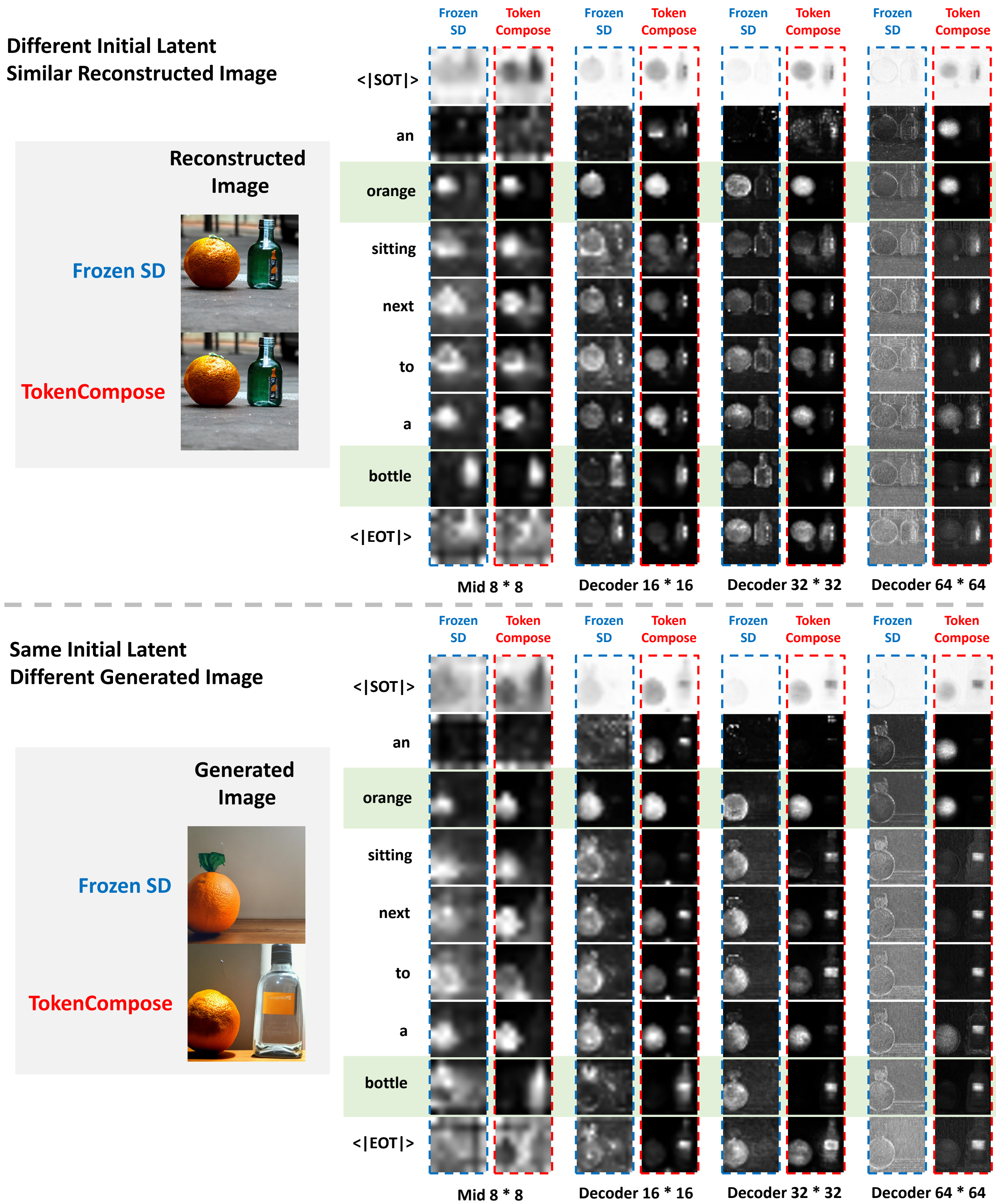

Figure 9. **Cross-attention map visualizations by different cross-attention layers at denoising U-Net.** We visualize the cross-attention map between a frozen Stable Diffusion [16] model and our model at the middle block and decoder layers of the denoising U-Net, where cross-attention at these layers are trained with our grounding objectives. In the upper example, we leverage null text inversion [14] to let two models reconstruct similar images using different latents for comparable cross-attention maps to demonstrate stronger grounding capabilities of our model. In the lower example, we use the same initial latent for two different models to generate images to demonstrate how stronger grounding capabilities lead to better compositionality.

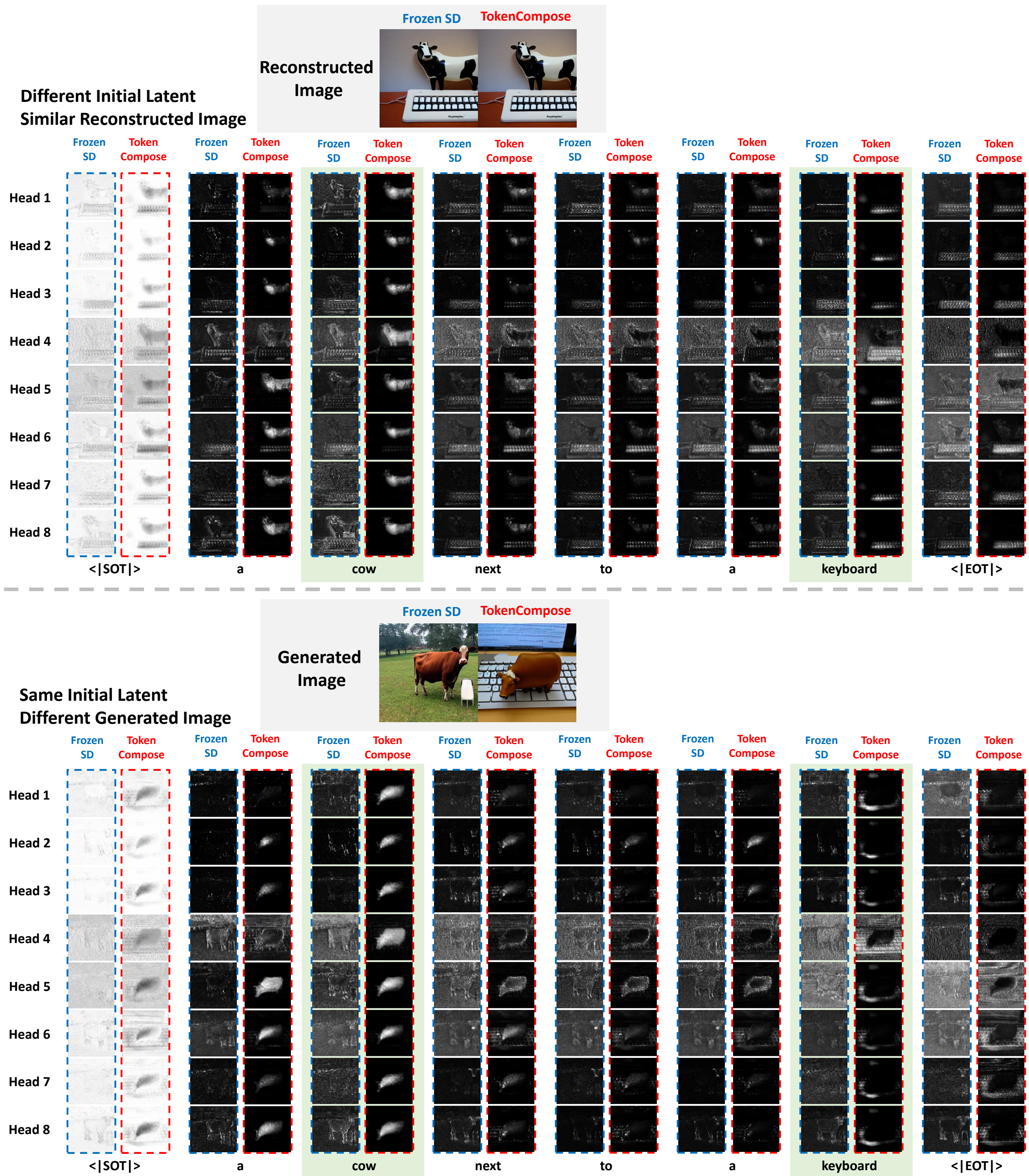

Figure 10. **Cross-attention map visualizations by different cross-attention heads at the denoising U-Net.** We visualize the cross-attention map of each head at the last layer (*i.e.,* $U_{\mathrm{D}}^{64\times 64}$) of the U-Net decoder between a frozen Stable Diffusion [16] model and our model. We use the same setting as in Figure 9 but with a different prompt for a more diverse visualization. We show that our grounding objective *allows* flexibility of different heads to attend to different regions of the latent.

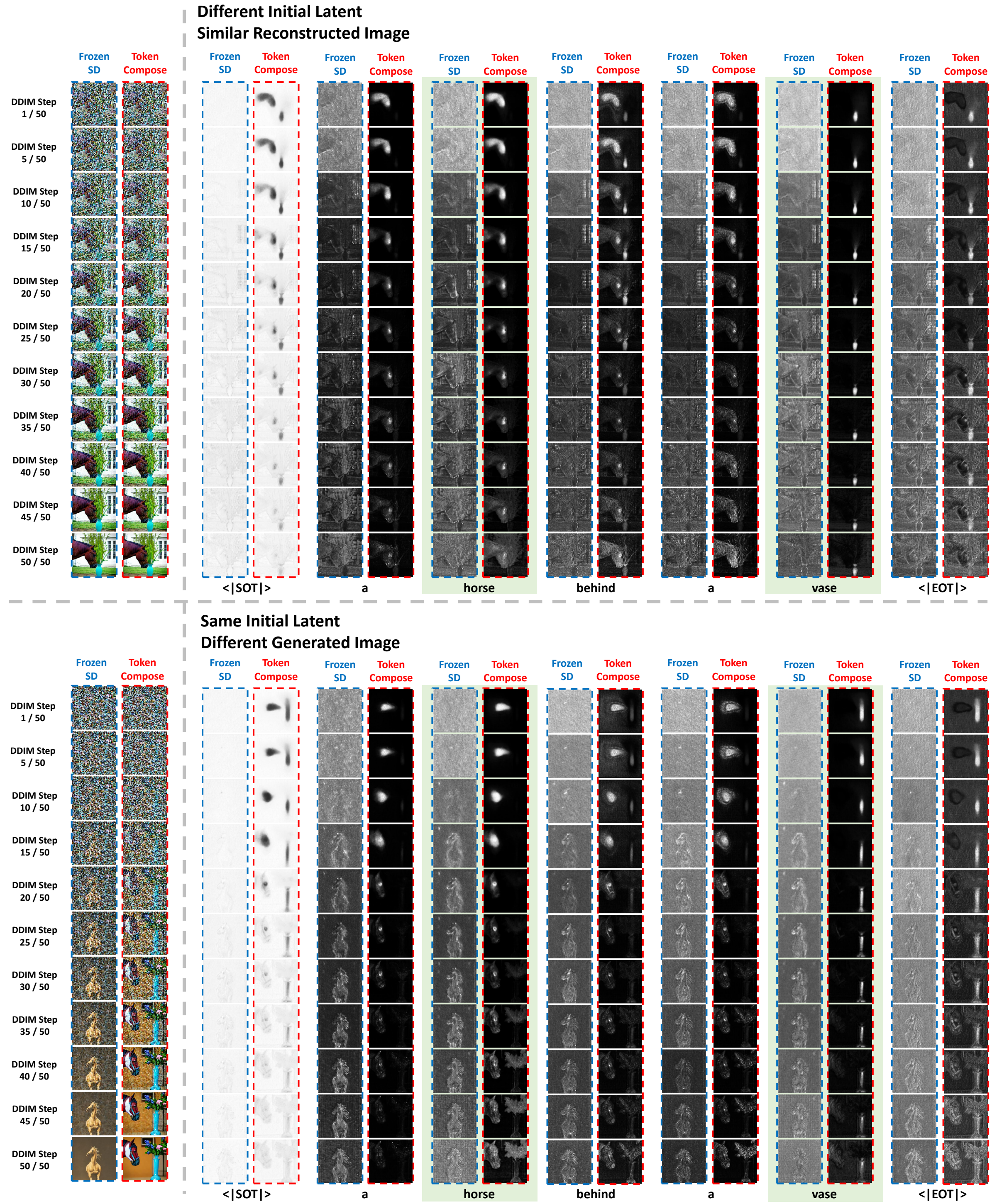

Figure 11. **Cross-attention map visualizations by different denoising time steps at the denoising U-Net.** We visualize the cross-attention map of the last layer (*i.e.,* $U_{\mathrm{D}}^{64\times64}$) of the U-Net decoder between a frozen Stable Diffusion [16] model and our model at time step 1 and every 5 steps based on a 50-step DDIM [17] scheduler. We use the same setting as in Figure 9 with a different prompt for a more diverse visualization. We show that our grounding objective enables cross-attention of different object tokens to aggregate at different regions of the noisy latent *early* during inference. This enables the model to generate different categories of instances more successfully, leading to better multi-category instance composition capabilities.

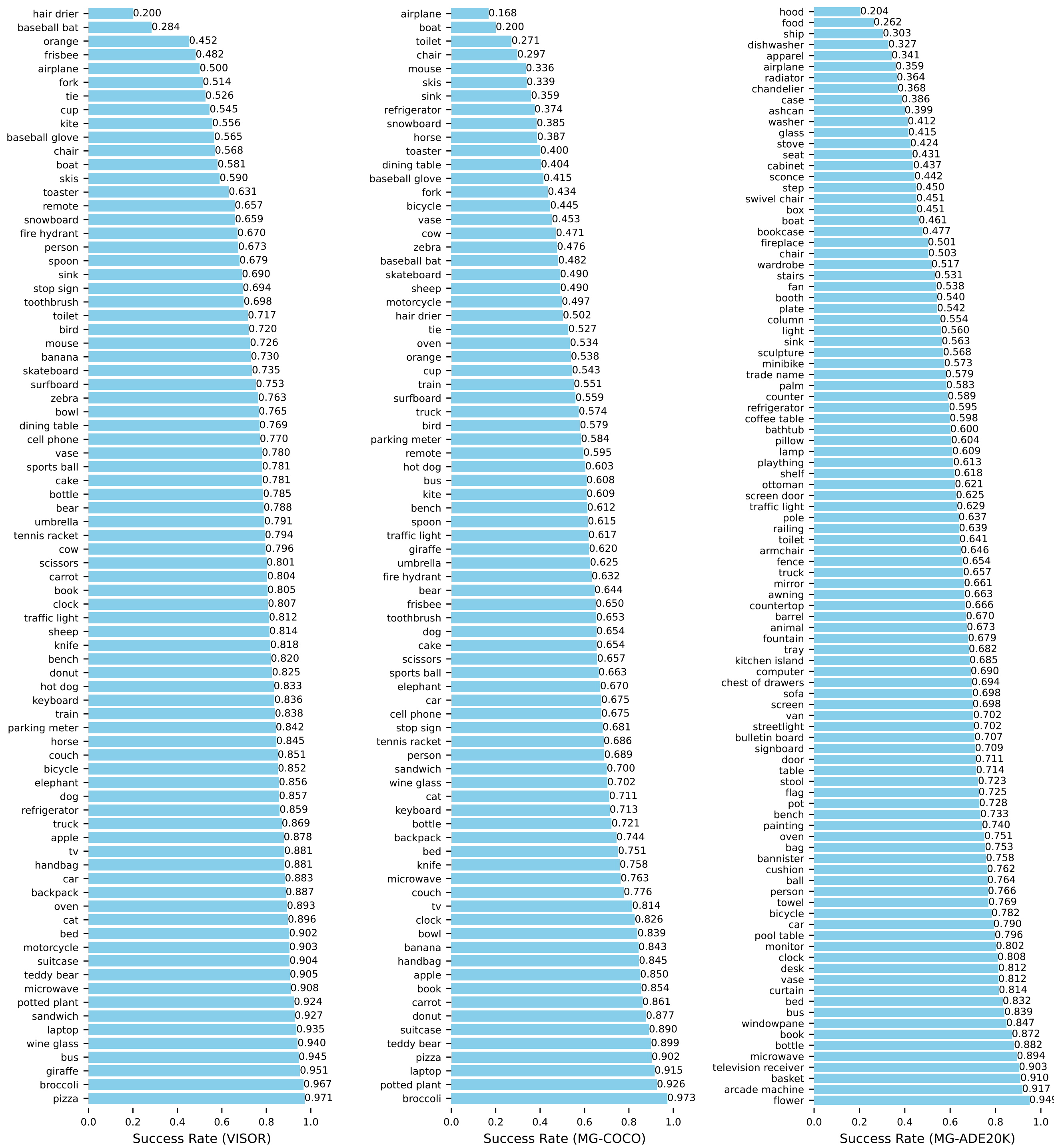

Figure 12. **Instance generation success rate in multi-category instance composition benchmarks.** We provide the success rate (from a 0-1 scale) of generating instances with our best-performing model for VISOR [4], MG-COCO, and MG-ADE20K benchmarks.

References

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019. 1, 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[4] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 9

[5] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[6] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 1

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1

[8] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1, 2

[9] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 1

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1, 2

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[13] Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[14] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2, 6

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6, 7, 8

[17] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 8