Supplementary Material: Divide & Bind Your Attention for Improved Generative Semantic Nursing

Yumeng Li[1,2]
yumeng.li@de.bosch.com
Margret Keuper[2,3]
margret.keuper@uni-siegen.de
Dan Zhang[1,4]
dan.zhang2@de.bosch.com
Anna Khoreva[1,4]
anna.khoreva@de.bosch.com

[1] Bosch Center for Artificial Intelligence
[2] University of Siegen
[3] Max Planck Institute for Informatics
[4] University of Tübingen

This supplementary material to the main paper is structured as follows:

- In Appendix S.1, more visual comparison is provided.
- In Appendix S.2, we provide additional quantitative evaluation using more metrics and with other methods.
- In Appendix S.3, we ablate on the binding loss L_{bind}.
- In Appendix S.4, we present the algorithm overview and more details on the TIFA evaluation.

More attention visualization in mp4 video format can be found in the zip file or our project page.

S.1 Additional Qualitative Results

We provide more visual comparison using additional novel prompts in Fig. S.1 and across different benchmarks using the same random seed in Fig. S.2. As can be seen, Divide & Bind can handle various complex prompts well and outperform the other methods in different scenarios.

S.2 Additional Quantitative Evaluation

In Table S.1, we compare our Divide & Bind with Stable Diffusion and Attend & Excite using Full Prompt similarity and Minimum Object Similarity used in [3]. Full Prompt Similarity represents the average CLIP cosine similarity between the full text prompt and

Figure S.1: Qualitative comparison using novel prompts with the same random seeds. Tokens used for optimization are highlighted in blue. Compared to others, Divide & Bind can better comply with the input prompt while maintaining a high level of realism.

Method	Animal-Animal		Animal-Scene		COCO-Subject	
	Full Prompt	Min. Obj.	Full Prompt	Min. Obj.	Full Prompt	Min. Obj.
Stable Diffusion	0.312	0.220	0.348	0.206	0.324	0.229
Attend & Excite	0.333	0.249	0.344	0.240	0.328	0.236
Divide & Bind	0.331	0.246	0.345	0.236	0.329	0.236

Table S.1: Quantitative comparison using Full Prompt Similarity and Minimum Object Similarity. The differences between methods are minor, which may due to the suboptimality of the evaluation metric as pointed in [8].

the generated images. And the Minimum Object Similarity is the minimum value of the object CLIP similarity among all objects mentioned in the prompt. For instance, for the prompt "a cat and a dog", we compute the similarity between the image and the sub-phrase "a dog" and "a cat" and take the smaller value as the final result. The difference among methods using CLIP similarities are minor, due to the fact that CLIP similarity may not be accurate to evaluate the faithfulness of Text-to-Image synthesis [8, 13]. Therefore, we employed more recent evaluate metrics, TIFA score [8] and Text-Text similarity, for more reliable evaluation, as reported in Fig. 6 and Table 2 in the main paper.

In Table S.2, we additionally compare with two more text-to-image methods, Composable Diffusion [12] and Structure Diffusion [5] using Text-Text similarity. We outperform the other methods on both Animal-Animal and Color-Object benchmarks.

S.3 Ablation Study

We ablate the effect of the proposed binding loss L_{bind} qualitatively and quantitatively, as shown in Fig. S.3 and Table S.3. We observe that the binding loss introduce minor differ-

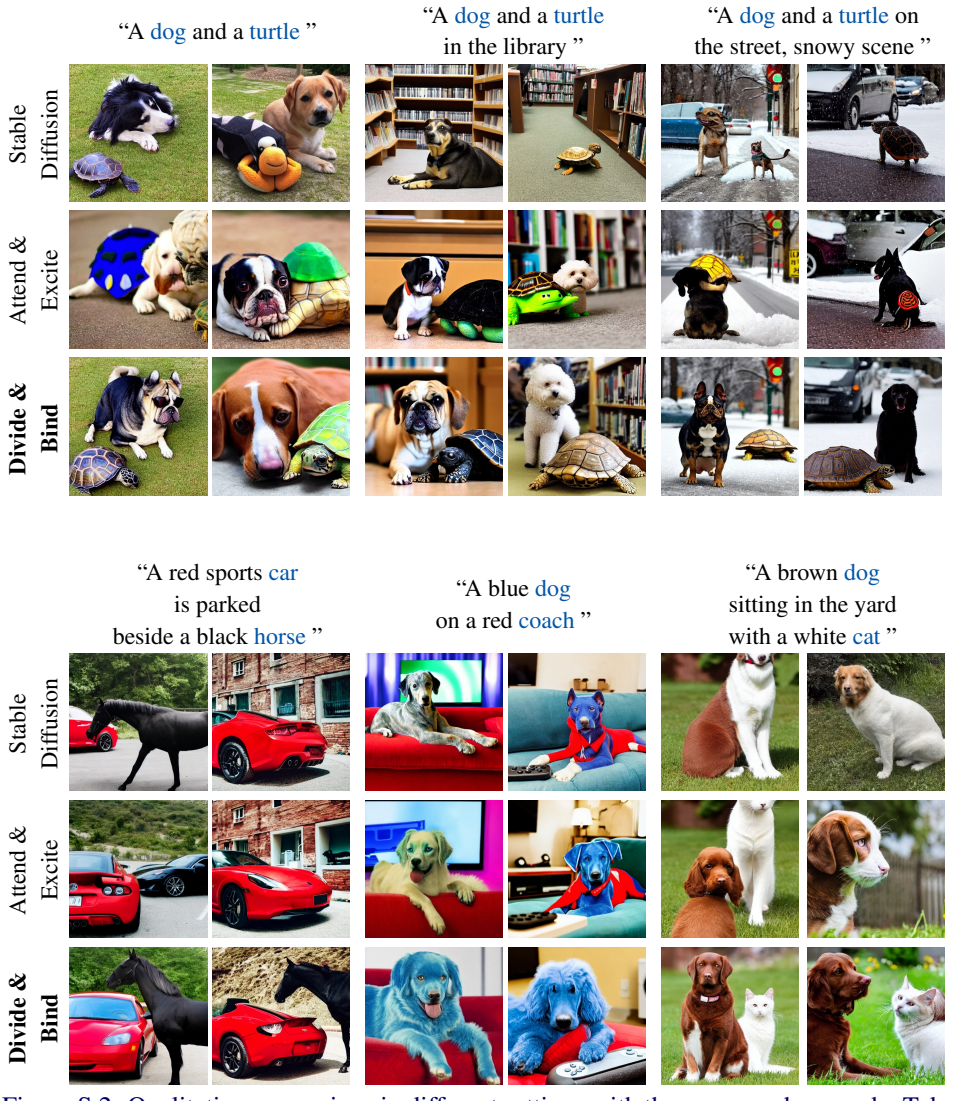

Figure S.2: Qualitative comparison in different settings with the same random seeds. Tokens used for optimization are highlighted in blue. Compared to others, Divide & Bind shows superior alignment with the input prompt while maintaining a high level of realism.

ence on the quantitative evaluation. We hypothesize that the coarse measurement of current evaluation metrics may not be able to reflect the advantage of our method and are not well aligned with human judgement [8, 13]. As illustrated in Fig. S.3, without the binding loss, the model is able to partially reflect the attribute but may mix with other attributes as well. For instance, in the second column, the front of the car is partially in green, which should be assigned to the balloon. While such imperfect results could still fool current evaluation metrics, as part of the car is indeed in pink. With L_{bind}, we can see the attributes can be more accurately localized at the corresponding object area. Therefore, we employ the binding loss by default, if the attributes are provided in the prompt.

Method	Animal-Animal	Color-Object
Stable Diffusion [25]	0.77	0.77
Composable Diffusion [20]	0.69	0.76
Structure Diffusion [8]	0.76	0.76
Attend & Excite [3]	0.80	0.81
Divide & Bind	**0.81**	**0.82**

Table S.2: Comparison with other Text-to-Image methods in Text-Text similarity. Divide & Bind surpasses the other methods on both evaluation sets.

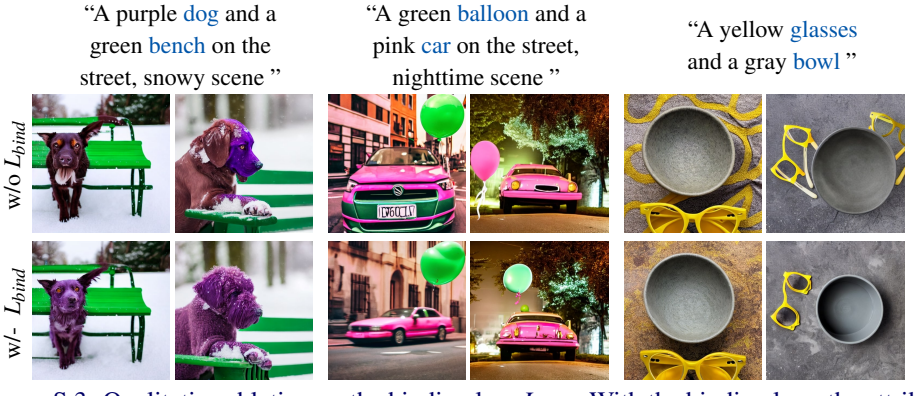

Figure S.3: Qualitative ablation on the binding loss L_{bind}. With the binding loss, the attribute can be more accurately assigned to the corresponding object.

S.4 Implementation & Evaluation Details

Algorithm 1 Simplified Algorithm Overview of Divide & Bind

Input: A text prompt $\mathcal{P}$ and a pretrained Stable Diffusion SD
Output: A noised latent z_{t-1} for the next denoising step

Determine object S and attribute R tokens by GPT with in-context learning as in TIFA [8]
Extract attention maps for the object tokens A_t^s and attribute tokens A^r
if A^r are not None **then**
    $L_{D\&B} = L_{attend} + \lambda L_{bind}$
else
    $L_{D\&B} = L_{attend}$
end if
$z_t' \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} L_{D\&B}$
$z_{t-1} \leftarrow SD(z_t', \mathcal{P}, t)$
return z_{t-1}

Algorithm Overview. We provide the algorithm overview in Algorithm 1. Given the text prompt $\mathcal{P}$, we firstly identify the tokens of interest, e.g., object tokens and attribute tokens. This process can either be done manually or with the aid of GPT-3 [1] can be done auto-

Method	Color-Object		Color-Obj-Scene		COCO-Subject	
	Text-Text	TIFA	Text-Text	TIFA	Text-Text	TIFA
w/o L_{bind}	0.815	0.876	0.729	0.919	0.796	0.800
w/- L_{bind}	0.814	0.877	0.727	0.918	0.799	0.805

Table S.3: Ablation study on the binding loss L_{bind}. Despite the approach with the binding loss achieved similar performance or minor improvement, we observed more accurate attribute localization as visualized in Fig. S.3.

matically as shown in [8]. Taking advantage of the in-context learning [2, 6] capability of GPT-3, by providing a few in-context examples, GPT-3 can automatically extract the desired nouns and adjectives for new input prompts. For instance, in our experiments on the COCO-Subject and COCO-Attribute benchmarks, we used the captions of COCO images without fixed templates as the prompts, where the object and attribute tokens were selected automatically using GPT-3. Based on the token indices, we can extract attention maps and apply our $L_{B\&D}$ to update the noised latent z_t.

CLIP-Based Evaluation. For computing the CLIP-based similarity metrics, e.g., Text-Text similarity, Full Prompt Similarity and Minimum Object Similarity, we employ the pretrained CLIP VIT-B/16 model [14]. To obtain the caption of generated images for Text-Text similarity evaluation, we use the BLIP [11] image captioning model finetuned on the MSCOCO Captions dataset [4] from the LAVIS library [10].

TIFA Evaluation. Evaluation of the TIFA metric is based on a performance of the visual-question-answering (VQA) system, e.g. mPLUG [9]. By definition, the TIFA score is essentially the VQA accuracy. Given the text input $\mathcal{T}$, we can generate $\mathcal{N}$ multiple-choice question-answer pairs $\{Q_i, C_i, A_i\}_{i=1}^{N}$, where Q_i is a question, C_i is a set of possible choices and A_i is the correct answer. These question-answer pairs can be designed manually or automatically produced by the large-scale language model, e.g. GPT-3 [1]. By providing a few in-context examples, GPT-3 can follow the instruction to generate question-answer pairs, and generalize to new text captions [2, 8].

Computational Complexity. Measured on a V100 GPU using 50 sampling steps, Stable Diffusion takes approximately 13 seconds to generate a single image. As we follow the hyperparameter settings as Attend & Excite [3], both A&E and our method have a similar average runtime of 17 seconds. The runtime slightly varies with the complexity of prompts.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[3] 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.

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[5] 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.

[6] Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A Smith, and Mari Ostendorf. In-context learning for few-shot dialogue state tracking. In *EMNLP*, 2022.

[7] Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A Smith, and Mari Ostendorf. In-context learning for few-shot dialogue state tracking. *arXiv preprint arXiv:2203.08568*, 2022.

[8] 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 preprint arXiv:2303.11897*, 2023.

[9] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *EMNLP*, 2022.

[10] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022.

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

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

[13] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. LLM-Score: Unveiling the power of large language models in text-to-image synthesis evaluation. *arXiv preprint arXiv:2305.11116*, 2023.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

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