Divide & Bind Your Attention for Improved Generative Semantic Nursing

Yumeng Li[1,2] Margret Keuper[2,3] Dan Zhang[1,4] Anna Khoreva[1,4]
[1]Bosch Center for AI [2] University of Siegen [3]MPI for Informatics [4]University of Tübingen
{yumeng.li, dan.zhang2, anna.khoreva}@de.bosch.com margret.keuper@uni-siegen.de

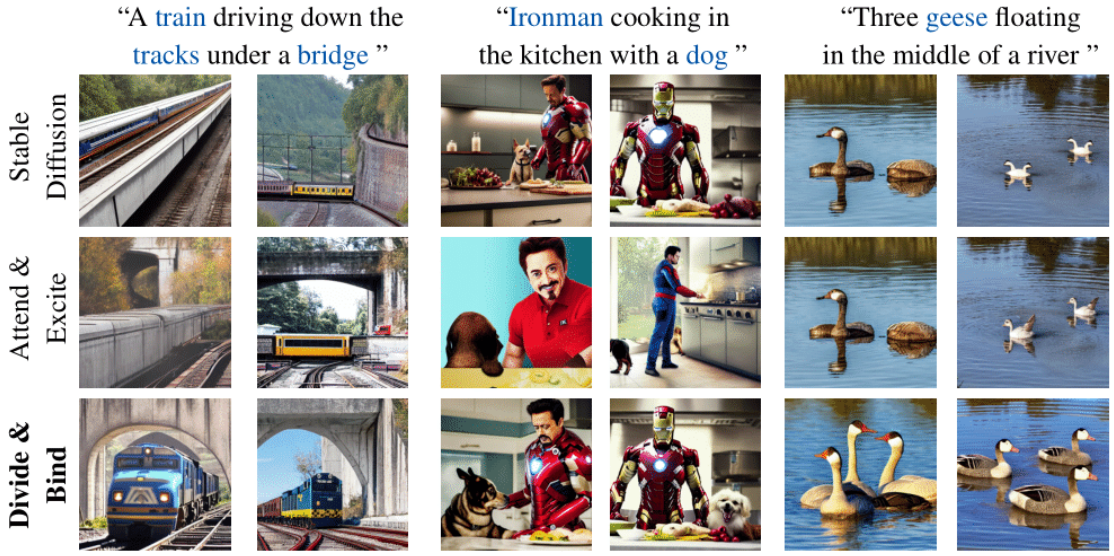

Figure 1: Our **Divide & Bind** can faithfully generate multiple objects based on detailed textual description. Compared to prior state-of-the-art semantic nursing technique for text-to-image synthesis, Attend & Excite [1], our approach exhibits superior alignment with the input prompt and maintain a higher level of realism.

Abstract

Emerging large-scale text-to-image generative models, e.g., Stable Diffusion (SD), have exhibited overwhelming results with high fidelity. Despite the magnificent progress, current state-of-the-art models still struggle to generate images fully adhering to the input prompt. Prior work, Attend & Excite, has introduced the concept of Generative Semantic Nursing (GSN), aiming to optimize cross-attention during inference time to better incorporate the semantics. It demonstrates promising results in generating simple prompts, e.g., "a cat and a dog". However, its efficacy declines when dealing with more complex prompts, and it does not explicitly address the problem of improper attribute binding. To address the challenges posed by complex prompts or scenarios involving multiple entities and to achieve improved attribute binding, we propose Divide & Bind. We introduce two novel loss objectives for GSN: a novel attendance loss and a binding loss. Our approach stands out in its ability to faithfully synthesize desired objects with improved attribute alignment from complex prompts and exhibits superior performance across multiple evaluation benchmarks. Our project page is here and full paper can be found in arXiv.

1. Introduction

In the realm of text-to-image (T2I) synthesis, large-scale generative models [4] have recently achieved significant progress and demonstrated exceptional capacity to generate stunning photorealistic images. However, it remains challenging to synthesize images that fully comply with the given prompt input [1].

There are two well-known semantic issues in text-to-image synthesis, i.e., "missing objects" and "attribute binding". "Missing objects" refers to the phenomenon that not all objects mentioned in the input text faithfully appear in the image. "Attribute binding" represents the critical compositionality problem that the attribute information, e.g.,

color or texture, is not properly aligned to the corresponding object or wrongly attached to the other object. To mitigate these issues, recent work Attend & Excite (A&E) [1] has introduced the concept of Generative Semantic Nursing (GSN). The core idea lies in updating latent codes on-the-fly such that the semantic information in the given text can be better incorporated within pretrained synthesis models.

As an initial attempt A&E [1], building upon the powerful open-source T2I model Stable Diffusion (SD) [4], leveraged cross-attention maps for optimization. Since cross-attention layers are the only interaction between the text prompt and the diffusion model, the attention maps have significant impact on the generation process. To enforce the object occurrence, A&E defined a loss objective that attempts to maximize the maximum attention value for each object token. Although showing promising results on simple composition, e.g., "a cat and a frog", we observed unsatisfying outcomes when the prompt becomes more complex, as illustrated in Fig. 1. A&E fails to faithfully synthesize the "train" or "dog" in the first two examples, and miss one "goose" in the third one. We attribute this to the suboptimal loss objective, which only considers the single maximum value and does not take the spatial distribution into consideration. As the complexity of prompts increases, token competition intensifies. The single excitation of one object token may overlap with others, leading to the suppression of one object by another (e.g., missing "train" in Fig. 1) or to hybrid objects, exhibiting features of both semantic classes (e.g., mixed dog-turtle.

In this work, we propose a novel objective function for GSN. We maximize the total variation of the attention map to prompt multiple, spatially distinct attention excitations. By spatially distributing the attention for each token, we enable the generation of all objects mentioned in the prompt, even under high token competition. Intuitively, this corresponds to *dividing* the attention map into multiple regions. Besides, to mitigate the attribute *binding* issue, we propose a Jensen-Shannon divergence (JSD) based binding loss to explicitly align the distribution between excitation of each object and its attributes. Thus, we term our method Divide & Bind. Our main contributions can be summarized as: (i) We propose a novel total-variation based attendance loss enabling presence of multiple objects in the generated image. (ii) We propose a JSD-based attribute binding loss for faithfull attribute binding. (iii) Our approach exhibits outstanding capability of generating images fully adhering to the prompt, outperforming A&E on several benchmarks involving complex descriptions.

2. Method

Given the recognized significance of the cross-attention maps in guiding semantic synthesis, our method aims at optimizing the latent code at inference time to excite them based on the text tokens. We employ the generative semantic nursing (GSN) method (Sec. 2.1) for latent code optimization, and propose a novel loss formulation (Sec. 2.2). It consists of two parts, i.e. *divide* and *bind*, which encourages object occurrence and attribute binding respectively.

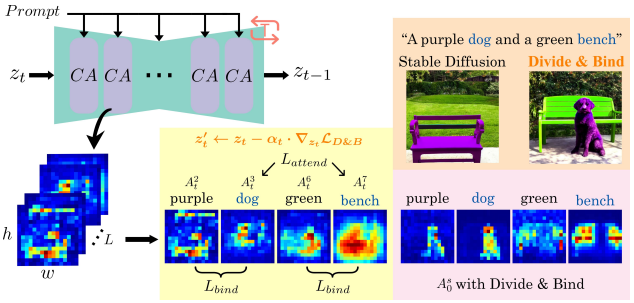

Figure 2: Method overview. We perform latent optimization on-the-fly based on the attention maps of the object tokens with our TV-based L_{attend} and JSD-based L_{bind}.

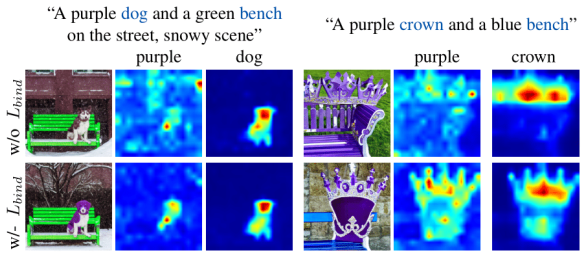

Figure 3: Binding loss ablation. L_{bind} aligns the excitation of attribute and object attention.

2.1. Generative Semantic Nursing (GSN)

We implement our method based on the open-source state-of-the-art T2I model Stable Diffusion(SD) [4]. To improve the semantic guidance in SD during inference, one pragmatic way is via latent code optimization at each time step of sampling, i.e. GSN [1], without any fine-tuning: $z'_t \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}$, where z_t is the latent variable at the timestep t, α_t is the updating rate and $\mathcal{L}$ is the loss to encourage the faithfulness between the image and text description, e.g. object attendances and attribute binding.

2.2. Divide & Bind

Our proposed method Divide & Bind consists of a novel objective for GSN

$$\min_{z_t} \mathcal{L}_{D\&B} = \min_{z_t} \mathcal{L}_{attend} + \lambda \mathcal{L}_{bind} \tag{1}$$

which has two parts, the attendance loss $\mathcal{L}_{attend}$ and the binding loss $\mathcal{L}_{bind}$ that respectively enforce the object attendance and attribute binding. λ is the weighting factor.

Divide for Attendance. The attendance loss L_{attend} is to incentivize the presence of the objects, thus is applied to the

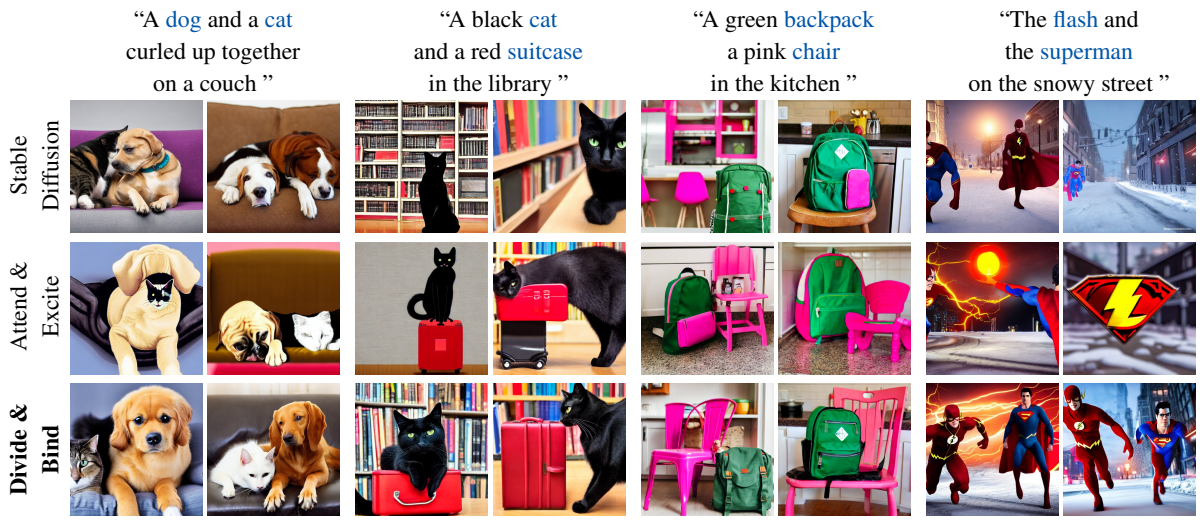

Figure 4: Qualitative comparison in different settings with the same random seeds. Tokens used for optimization are highlighted in blue. Compared to others, Divide & Bind shows superior alignment with the input prompt while maintaining a high level of realism.

text tokens associated with *objects* S,

$$\mathcal{L}_{attend} = -\min_{s \in S} TV(A_t^s), \tag{2}$$

$$TV(A_t^s) = \sum_{i,j} |A_t^s[i+1,j] - A_t^s[i,j]| + |A_t^s[i,j+1] - A_t^s[i,j]|$$

where $A_t^s[i,j]$ denotes the attention value of the s-th token at the specific location $[i,j]$ and time step t. The loss formulation in Eq. (2) is based on the the finite differences approximation of the total variation (TV) $|\nabla A_t^s|$ along the spatial dimensions. It is evaluated for each object token and we take the smallest value, i.e., representing the worst case among the all object tokens. Taking the negative TV as the loss, we essentially maximize the TV for latent optimization in Eq. (1). It encourages large activation differences across many neighboring spatial positions, thus not only having one high activation region but also many of them. Such an activation pattern in the space resembles to dividing it into different regions. The model can select some of them to display the object with single or even multiple attendances. This way, conflicts between different objects that compete for the same region can be more easily resolved.

Attribute Binding Regularization. In addition to the object attendance, the given attribute information, e.g. color or material, should be appropriately attached to the corresponding object. We denote the attention map of the object token and its attribute token as A_t^s and A_t^r, respectively. For attribute binding, it is desirable that A_t^r and A_t^s are spatially well-aligned, i.e. high activation regions of both tokens are largely overlapped. To this end, we introduce $\mathcal{L}_{bind}$. After proper normalization along the spatial dimension, we can view the normalized attention maps $\widetilde{A_t^r}$ and $\widetilde{A_t^s}$ as two probability mass functions whose sample space has size $h \times w$. To explicitly encourage such alignment, we can then minimize the symmetric similarity measure Jensen–Shannon divergence (JSD) between these two distributions:

$$\mathcal{L}_{bind} = JSD\left(\widetilde{A_t^r} \| \widetilde{A_t^s}\right). \tag{3}$$

Specifically, we adopt the Softmax-based normalization along the spatial dimension. When performing normalization, we also observe the benefit of first aligning the value range between the two attention maps. Namely, the original attention map of the object tokens A_t^s have higher probability values than the ones of the attribute tokens A_t^r. Therefore, we first re-scale A_t^r to the same range as A_t^s. As illustrated in Fig. 3, after applying L_{bind}, the attribute token (e.g. "purple") is more localized to the correct object region (e.g. "dog" or "crown").

3. Experiments

Benchmarks. We conduct exhaustive evaluation on multiple prompt sets. Animal-Animal and Color-Object are proposed in [1], which simply compose two subjects and alternatively assign a color to the subject. Building on top of this, we append a postfix describing the scene or scenario to challenge the methods with higher prompt complexity, termed as Animal-Scene and Color-Obj-Scene. Further, we introduce Multi-Object which aims to produce multiple entities in the image. Note that different entities could belong to the same category. For instance, "one cat and two dogs" contains in total three entities and two of them are dogs.

Evaluation metrics. For quantitative evaluation, we used

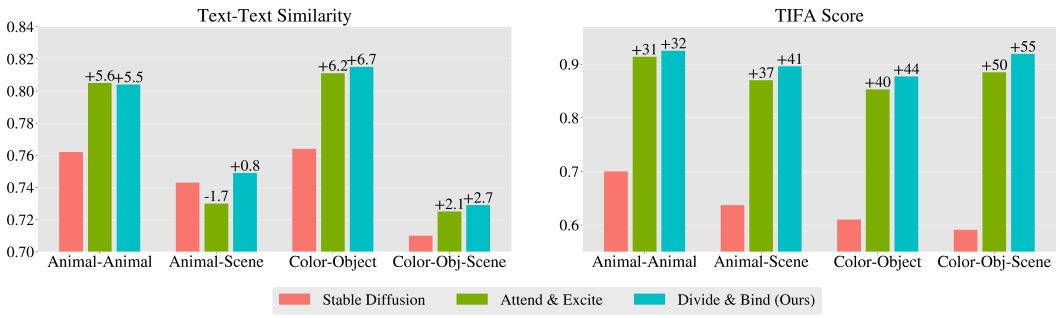

Figure 5: Quantitative comparison using Text-Text similarity and TIFA Score. Divide & Bind achieves comparable performance to A&E on the simple Animal-Animal and Color-Object, and shows superior results on more complex text descriptions, i.e., Animal-Scene and Color-Obj-Scene. Improvements over SD in % are reported on top of the bars.

the text-text similarity from [1] and the recently introduced TIFA score [2], which is more accurate than CLIPScore-and has much better alignment with human judgment on text-to-image synthesis. Text-text similarity measures the CLIP similarity between the original prompts and BLIP [3] generated captions of synthesized images. TIFA score is essentially the visual-question-answering accuracy based on a VQA system.

Main results. As shown in Fig. 5, we first quantitatively compare Divide & Bind (D&B) with Stable Diffusion (SD) [4] and Attend & Excite (A&E) [1] on Animal-Animal and Color-Object, originally proposed in [1], as well as our new benchmarks Animal-Scene and Color-Obj-Scene, which include scene description and has higher prompt complexity. It can be seen that D&B is on-par with A&E on Animal-Animal and achieves slight improvement on Color-Object. Due to the simplicity of the template, the potential of our method cannot be fully unleashed in those settings. In more complex prompts: Animal-Scene and Color-Obj-Scene, D&B outperforms the other methods more evidently, especially on the TIFA score (e.g., 5% improvement over A&E in Color-Obj-Scene). Qualitatively, both SD and A&E may neglect the objects, as shown in the "the flash and the superman on the snowy street" example in Fig. 4. Despite the absence of objects in the synthesized images, we found SD can properly generate the scene, while A&E tends to ignore it occasionally, e.g. the "library" and "kitchen" information in Fig. 4). In the "a green backpack and a pink chair in the kitchen" example, both SD and A&E struggle to bind the pink color with the chair only. In contrast, Divide & Bind, enabled by the binding loss, demonstrates a more accurate binding effect and has less leakage to other objects or background.

Next, we evaluate the methods on Multi-Object, where multiple entities should be generated. Visual comparison is presented in the third column of Fig. 1. In the "three geese floating in the middle of a river" example, both SD and A&E only synthesize two realistic looking geese, while the image generated by D&B fully complies with the prompt. We observe that often the result of A&E resembles the one of SD. This is not surprising, as A&E does not encourage attention activation in multiple regions. As long as one instance of the corresponding object token appears, the loss of A&E would be low, leading to minor update.

4. Conclusion

In this work, we propose a novel inference-time optimization objective Divide & Bind for semantic nursing of pretrained T2I diffusion models. Targeting at mitigating semantic issues in T2I synthesis, our approach demonstrates its effectiveness in generating multiple instances with correct attribute binding given complex textual descriptions. We believe that our regularization technique can provide insights in the generation process and support further development in producing images semantically faithful to the textual input.

References

[1] 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, 3, 4

[2] 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. 4

[3] 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. 4

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 4