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

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. Project page and code.

1 Introduction

In the realm of text-to-image (T2I) synthesis, large-scale generative models [1, 5, 12, 22, 23, 25, 30] 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 [6, 7, 18, 29]. 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) [6] has introduced the concept of Generative Semantic

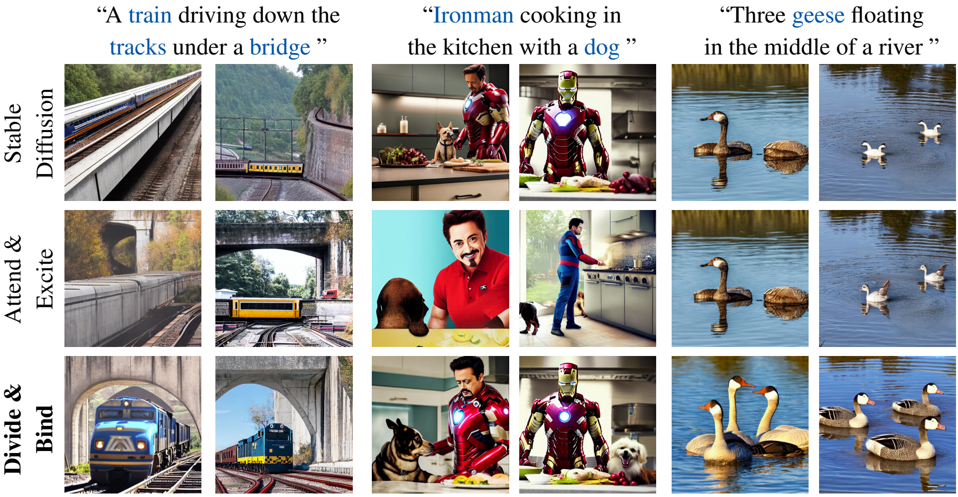


Figure 1: Our **Divide & Bind** can faithfully generate multiple objects based on detailed textual description. Compared to prior state-of-the-art semantic nursing technique for text-to-image synthesis, Attend & Excite [6], our approach exhibits superior alignment with the input prompt and maintain a higher level of realism.

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 [6], building upon the powerful open-source T2I model Stable Diffusion (SD) [23], 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 Fig. 3). Similar phenomenon has been observed in [28] as well.

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

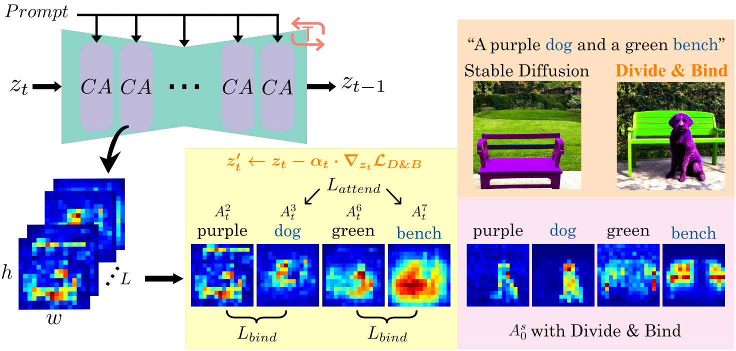


Figure 2: Method overview. We perform latent optimization on-the-fly based on the attention maps of the object tokens with our TV-based L_{attend} and JSD-based L_{bind}.

approach exhibits outstanding capability of generating images fully adhering to the prompt, outperforming A&E on several benchmarks involving complex descriptions.

2 Related Work

Text-to-Image Synthesis. With the rapid emergence of diffusion models [9, 19, 26], recent large-scale text-to-image models such as eDiff-I [1], Stable Diffusion [23], Imagen [25], or DALL·E 2 [22] have achieved impressive progress. Despite synthesizing high-quality images, it remains challenging to produce results that properly comply with the given text prompt. A few recent works [6, 7] aim at improving the semantic guidance purely based on the text prompt without model fine-tuning. StructureDiffusion [7] used language parsers for hierarchical structure extraction, to ease the composition during generation. Attend & Excite (A&E) [6] optimizes cross-attention maps during inference time by maximizing the maximum attention value of each object token to encourage object presence. However, we observed that A&E struggles with more complex prompts. In contrast, our Divide & Bind fosters the stimulation of multiple excitations, which aids in holding the position amidst competition from other tokens. Additionally, we incorporate a novel binding loss that explicitly aligns the object with its corresponding attribute, yielding more accurate binding effect.

Total Variation. Total variation (TV) measures the differences between neighbors. Thus, minimization encourages smoothness that was used in different tasks, *e.g.*, denoising [3], image restoration [4], and segmentation [27], just to name a few. Here, we use TV for a different purpose. We seek to divide attention maps into multiple excited regions. Thus, we choose TV *maximization* to enlarge the amount of local changes in attention maps over the image such that diverse object regions are encouraged to emerge. As a result, we enhance the chance of generating each desired object while concurrently competing with other objects.

3 Preliminaries

Stable Diffusion (SD). We implement our method based on the open-source state-of-the-art T2I model SD [23], which belongs to the family of latent diffusion models (LDMs). LDMs are two-stage methods, consisting of an autoencoder and a diffusion model trained in the latent space. In the first stage, the encoder $\mathcal{E}$ transforms the given image x into a latent code

$z = \mathcal{E}(x)$, then z is mapped back to the image space by the decoder $\mathcal{D}$. The autoencoder is trained to reconstruct the given image, i.e. $\mathcal{D}(\mathcal{E}(x)) \approx x$. In the second stage, a diffusion model [9, 19] is trained in the latent space of the autoencoder. During training, we gradually add noise to the original latent z_0 with time, resulting in z_t. Then the UNet [24] denoiser ε_θ is trained with a denoising objective to predict the noise ε that is added to z_0:

$$\mathcal{L} = \mathbb{E}_{z\sim\mathcal{E}(x),\varepsilon\sim N(0,I),c,t}\left[\|\varepsilon - \varepsilon_\theta(z_t, c, t)\|^2\right], \tag{1}$$

where c is the conditional information, *e.g*. text. During inference, given z_T randomly sampled from Gaussian distribution, UNet outputs noise estimation and gradually removes it, finally producing the clean version z_0.

Cross-Attention in Stable Diffusion. In SD, a frozen CLIP text encoder [21] is adopted to embed the text prompt $\mathcal{P}$ into a sequential embedding as the condition c, which is then injected into UNet through cross-attention (CA) to synthesize text-complied images. The CA layers take the encoded text embedding and project it into queries Q and values V. The keys K are mapped from the intermediate features of UNet. The attention maps are then computed by $A_t = Softmax(\frac{QK^T}{\sqrt{d}})$, where t indicates the time step, Softmax is applied along the channel dimension. The attention maps A_t can be reshaped into $\mathbb{R}^{h\times w\times L}$, where h, w is the resolution of the feature map, L is the sequence length of the text embedding. Further, we denote the cross-attention map that corresponds to the sth text token as $A_t^s \in \mathbb{R}^{h\times w}$, see an illustration in Fig. 2. One known issue of SD is that not all objects are necessarily present in the final image [6, 17, 29], while, as shown in [1, 8], the high activation region of the corresponding attention map strongly correlates to the appearing pixels belonging to one specific object in the final image. Hence, the activation in the attention maps is an important signal and an influencer in the semantic guided synthesis.

4 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. 4.1) for latent code optimization, and propose a novel loss formulation (Sec. 4.2). It consists of two parts, i.e. *divide* and *bind*, which encourages object occurrence and attribute binding respectively.

4.1 Generative Semantic Nursing (GSN)

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 [6]

$$z_t' \leftarrow z_t - \alpha_t \cdot \nabla_{z_t}\mathcal{L}, \tag{2}$$

where 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. GSN has the advantage of avoiding fine-tuning SD.

As the text information is injected into the UNet of SD via cross attention layers, it is natural to set the loss $\mathcal{L}$ with the cross attention maps as the inputs. Given the text prompt

Figure 3: Cross-attention visualization in different timesteps for each object token and predicted clean image $\hat{x}_0^{(t)}$. Note that this is GIF, video version is provided in the supp. material and our project page.

$\mathcal{P}$ and a list of object tokens S, we will have a set of attention maps $\{A_t^s\}$ for $s \in S$. Ideally, if the final image contains the concept provided by the object token s, the corresponding cross-attention map A_t^s should show strong activation. To achieve this, A&E [6] enhances the single maximum value of the attention map, i.e. $L_{A\&E} = -\min_{s\in S}(\max_{i,j}(A_t^s[i,j]))$. However, it does not facilitate with multiple excitations, which is increasingly important when confronted with complex prompts and the need to generate multiple instances. As shown in Fig. 3, a single excitation can be easily taken over by the other competitor token, leading to missing objects in the final image. Besides, it does not explicitly address the attribute binding issue. Instead, our Divide & Bind promotes the allocation of attention across distinct areas, enabling the model to explore various regions for object placement. Moreover, we introduce an attribute binding regularization which explicitly encourages attribute alignment.

4.2 Divide & Bind

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

$$\min_{z_t} \mathcal{L}_{D\&B} = \min_{z_t} \mathcal{L}_{attend} + \lambda \mathcal{L}_{bind} \tag{3}$$

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. Detailed formulation of both loss terms is presented as follows.

Divide for Attendance. The attendance loss L_{attend} is to incentivize the presence of the objects, thus is applied to the text tokens associated with *objects* S,

$$\mathcal{L}_{attend} = -\min_{s\in S} TV(A_t^s),\ TV(A_t^s) = \sum_{i,j} |A_t^s[i+1,j] - A_t^s[i,j]| + |A_t^s[i,j+1] - A_t^s[i,j]| \tag{4}$$

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. (4) 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

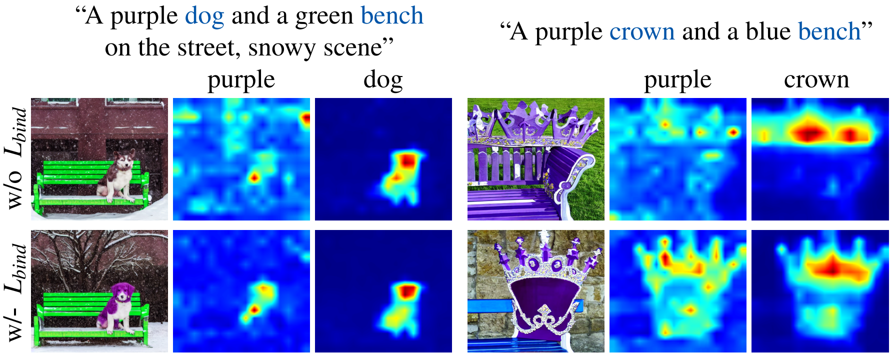


Figure 4: Binding loss ablation. L_{bind} aligns the excitation of attribute and object attention.

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. (3). 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. Furthermore, from an optimization perspective, it allows the model to search among different options for converging to the final solution. The loss is applied at the initial sampling steps. As can be seen from the GIF in Fig. 3, for the "dog" token, regions on both left and right sides are explored in the initial phase. In the end, the left side is taken over by the "turtle" but the "dog" token covers the right side. While for SD, the "dog" token has a single weak activation, and for Attend & Excite, it only has one single high activation region on the right that is taken over by the "turtle" later.

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:

$$\mathcal{L}_{bind} = JSD\left(\widetilde{A_t^r} \| \widetilde{A_t^s}\right). \tag{5}$$

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. 4, after applying L_{bind}, the attribute token (e.g. "purple") is more localized to the correct object region (e.g. "dog" or "crown").

Implementation Details. The token identification process can either be done manually or automatically with the aid of GPT-3 [2] as shown in [11]. Taking advantage of the in-context learning [10] capability of GPT-3, by providing a few in-context examples, GPT-3

Evaluation Set	Description	Example	# Prompt
Animal-Animal	a [animalA] and a [animalB]	"a cat and a frog"	66
Color-Object	a [colorA] [subjectA] and a [colorB] [subjectB]	"a green backpack and a pink chair"	66
Animal-Scene	a [animalA] and a [animalB] [scene]	"a bird and a bear in the kitchen"	56
Color-Obj-Scene	a [colorA] [subjectA] and a [colorB] [subjectB] [scene]	"a black cat and a red suitcase in the library"	60
Multi-Object	more than two instances in the image	"two cats and two dogs" "three sheep standing in the field"	30
COCO-Subject	filtered COCO captions containing subject related questions from TIFA	"a dog and a cat curled up together on a couch"	30
COCO-Attribute	filtered COCO captions containing attribute related questions from TIFA	"a red sports car is parked beside a black horse"	30

Table 1: Description of benchmarks used for the experimental evaluation.

can automatically extract the desired nouns and adjectives for new input prompts.

We inherit the choice of optimization hyperparameters from the initial attempt for GSN - Attend & Excite (A&E) [6]. The optimization is operated on the attention map at 16×16 resolution, as they are the most semantically meaningful ones [8]. Based on the observation that the image semantics are determined by the initial denoising steps [13, 16], the update is only performed from $t = T$ to $t = t_{end}$, where $T = 50$ and $t_{end} = 25$ in all experiments. The weight of binding loss $\lambda = 1$, if the attribute information is provided. Otherwise, $\lambda = 0$, i.e., using only the attendance loss.

5 Experiments

5.1 Experimental Setup

Benchmarks. We conduct exhaustive evaluation on seven prompt sets as summarized in Table 1. Animal-Animal and Color-Object are proposed in [6], 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. Besides the designed templates, we also filtered the COCO captions used in the TIFA benchmark [11] and categorize them into COCO-Subject and COCO-Attribute. There are up to four objects without any attribute assigned in COCO-Subject and two objects with attributes COCO-Attribute, respectively. Note that the attributes in COCO-Attribute contain not only color, but also texture information, such as "a wooden bench".

Evaluation metrics. To quantitatively evaluate the performance of our method, we used the text-text similarity from [6] and the recently introduced TIFA score [11], which is more accurate than CLIPScore [21] and has much better alignment with human judgment on text-to-image synthesis. To compute the text-text similarity, we employ the off-the-shelf image captioning model BLIP [15] to generate captions on synthesized images. We then measure the CLIP similarity between the original prompt and all captions. Evaluation of the

Figure 5: 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.

TIFA metric is based on a performance of the visual-question-answering (VQA) system, e.g. mPLUG [14]. By definition, the TIFA score is essentially the VQA accuracy. More detailed description of the TIFA evaluation protocol and evaluation on the full prompt text-image similarity and minimum object similarity from [6] can be found in the supp. material.

5.2 Main Results

As shown in Fig. 6, we first quantitatively compare Divide & Bind with Stable Diffusion (SD) [23] and Attend & Excite (A&E) [6] on Animal-Animal and Color-Object, originally proposed in [6], 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 Divide & Bind 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

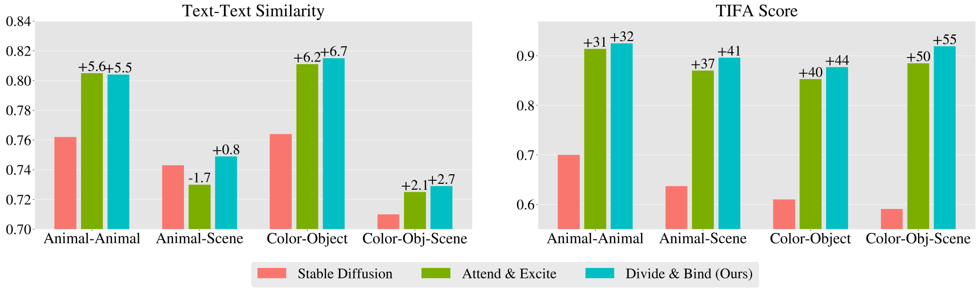


Figure 6: Quantitative comparison using Text-Text similarity and TIFA Score. Divide & Bind achieves comparable performance to A&E on the simple Animal-Animal and Color-Object, and shows superior results on more complex text descriptions, i.e., Animal-Scene and Color-Obj-Scene. Improvements over SD in % are reported on top of the bars.

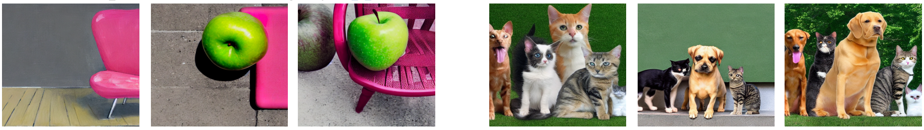
Method	Multi-Object		COCO-Subject		COCO-Attribute	
	Text-Text	TIFA	Text-Text	TIFA	Text-Text	TIFA
Stable Diffusion	0.786	0.647	0.823	0.791	0.790	0.752
Attend & Excite	**0.809**	0.755	0.818	0.824	0.793	0.798
Divide & Bind	0.805	**0.785**	**0.824**	**0.840**	**0.799**	**0.805**

Table 2: Quantitative comparison on complex COCO-captions and Multi-Object generation. Divide & Bind surpasses the other methods when it comes to handling complex prompts.

unleashed in those settings. In more complex prompts: Animal-Scene and Color-Obj-Scene, Divide & Bind 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 "bird and a bear on the street, snowy scene" example in Fig. 5. 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 the second column of Fig. 5). 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. We provide ablation on the binding loss in the supp. material.

Next, we evaluate the methods on Multi-Object, where multiple entities should be generated. Visual comparison is presented in the third column of Fig. 5. In the "three sheep standing in the field" example, both SD and A&E only synthesize two realistic looking sheep, while the image generated by Divide & Bind fully complies with the prompt. For the "one cat and two dogs" example, SD and A&E either miss one entity or generate the wrong species. 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. We also provide the quantitative evaluation in Table 2. Our Divide & Bind outperforms other methods by a large margin on the TIFA score, but only slightly underperforms A&E on Text-Text similarity. We hypothesize that this is due to the incompetence of CLIP on counting [20], thus leading to inaccurate evaluation, as pointed out in [11] as well.

We also benchmark on real image captions, i.e. COCO-Subject and COCO-Attribute, "A pink chair and a gray apple" "One dog and three cats"



Stable Diffusion Attend&Excite **Divide & Bind** Stable Diffusion Attend&Excite **Divide & Bind**

Figure 7: Limitations: challenging rare combinations (left) and instance miscounting (right).

where the text structure can be more complex than fixed templates. Quantitative evaluation is provided in Table 2, where Divide & Bind showcases its advantages on both benchmarks over SD and A&E. A visual example "a dog and a cat curled up together on a couch" is shown in Fig. 5. Consistent with the observation above: while A&E encourages the object occurrence, it may generate unnatural looking images. While SD, may neglect the object, its results are more realistic. Divide & Bind performs well with respect to both perspectives.

Limitations. Despite improved semantic guidance, it is yet difficult to generate extremely rare or implausible cases, e.g., unusual color binding "a gray apple". Our method may generate such objects together with the common one, e.g., generating a green apple and a gray apple in the same image, see Fig. 7. As we use the pretrained model without fine-tuning, some data bias is inevitably inherited. Another issue is miscounting: more instances may be generated than it should. We attribute the miscounting to the imprecise language understanding limited by the CLIP text encoder [20, 21]. This effect is also observed in other large-scale T2I models, e.g., Parti [30], making it an interesting case for future research.

6 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] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[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] Vicent Caselles, Antonin Chambolle, and Matteo Novaga. Total variation in imaging. *Handbook of mathematical methods in imaging*, 1(2):3, 2015.

[4] T Chan, Selim Esedoglu, Frederick Park, and A Yip. Total variation image restoration:

Overview and recent developments. *Handbook of mathematical models in computer vision*, 2006.

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[6] 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.

[7] 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.

[8] 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.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33, 2020.

[10] 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.

[11] 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.

[12] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023.

[13] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2023.

[14] 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.

[15] 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.

[16] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022.

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

[18] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022.

[19] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

[20] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*, 2023.

[21] 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.

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020.

[27] Dennis Sun and Matthew Ho. Image segmentation via total variation and hypothesis testing methods. 2011.

[28] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *ACL*, 2023.

[29] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

[30] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022.