Object-Conditioned Energy-Based Model for Attention Map Alignment in Text-to-Image Diffusion Models

Yasi Zhang, Peiyu Yu, Ying Nian Wu
Department of Statistics and Data Science
University of California, Los Angeles
yasminzhang@ucla.edu, yupeiyu98@g.ucla.edu, ywu@stat.ucla.edu

Abstract

Text-to-image diffusion models have shown great success in generating high-quality text-guided images. Yet, these models may still fail to semantically align generated images with the provided text prompts, leading to problems like incorrect attribute binding and/or catastrophic object neglect. Given the pervasive object-oriented structure underlying text prompts, we introduce a novel object-conditioned Energy-Based Attention Map Alignment (EBAMA) method to address the aforementioned problems. We show that an object-centric attribute binding loss naturally emerges by approximately maximizing the log-likelihood of a z-parameterized energy-based model with the help of the negative sampling technique. We further propose an object-centric intensity regularizer to prevent excessive shifts of objects attention towards their attributes. Extensive qualitative and quantitative experiments on the AnE benchmark demonstrate the superior performance of our method over previous strong counterparts.

1. Introduction

Recently, large-scale text-to-image diffusion models [1, 8, 12, 15, 18, 19] have showcased remarkable capabilities in producing diverse, imaginative, high-resolution visual content based on free-form text prompts. Despite their revolutionary progress, however, these models may not consistently capture and convey the full semantic meaning of the provided text prompts [4, 16]. Some well-known issues include omission, hallucination, or duplication of details [22], semantic leakage of attributes between entities [16], and miscomprehension of intricate textual descriptions [19].

Many previous works have focused on addressing the semantic misalignment issues, particularly concerning multiple-object generation and attribute binding. Composable Diffusion (CD) [10] composes multiple output noises guided by different objects in a text prompt during the generation process. Prompt-to-Prompt (PtP) [6] observes a strong correlation between cross-attention maps and the layout of an image. Building on this, Structured Diffusion (StrD) [5] experiments with averaging attention maps generated by different noun phrases for the same queried image latent representation. Attend-and-Excite (AnE) [3] proposes a novel approach of maximizing the attention map scores of object tokens by updating the latent at each sampling step. However, we note that artifacts and incorrect attribute binding are likely when AnE maximizes the attention weights of object tokens without any concerns on attributes. In response, SynGen (SG) [17] proposes an attribute-object pair-centric objective, aiming to minimize

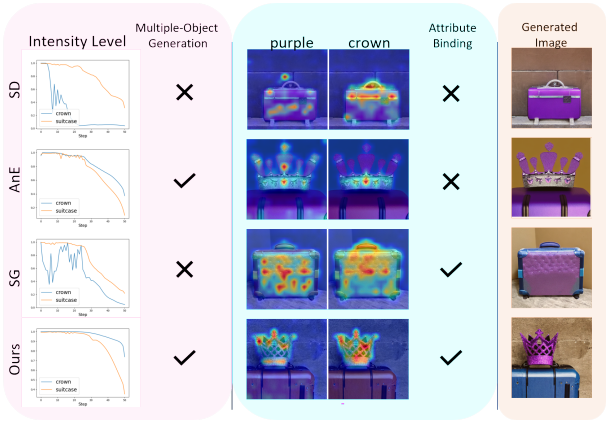

Figure 1. **Key observations of the generation process of diffusion models**. The given prompt is "a purple crown and a blue suitcase". In the left panel, we hypothesize that if the intensity level of any object in the prompt does not remain high during the first half of the denoising process, e.g. the crown in SD and SG, the model would fail to generate the object in the final image. The middle panel suggests that if the attention map distributions of any attribute-object pair are not aligned, the model would struggle to correctly bind attributes to their respective objects, e.g. 'purple' and 'crown' in SD and AnE. The generated images are displayed in the right panel. All methods share the same random seed.

the distribution distance within the pair while maximizing it from other tokens, based on the assumption that normlized attention maps follow a multinomial distribution. Diverging from these methods, Energy-Based Cross Attention (EBCA) [13] introduces an Energy-Based Model (EBM) framework [21, 23–25] for queries and keys within cross-attention mechanisms, proposing updates to text embeddings instead of latent noise representations.

A closer look at both the fluctuations of attention intensities and the attention distributions of attribute-object pairs in these methods shed light on the root cause of the misalignment issues. As illustrated in Fig. 1, alignment in attribute-object attention maps (e.g., 'purple crown' in SG) encourages attribute binding. However, attention map alignment alone does not guarantee complete semantic alignment, as the intensity levels of object attention maps are crucial in determining the presence of an object in the final image.

Motivated by these key observations, we introduce a novel *object-conditioned* Energy-Based Attention Map Alignment (EBAMA) method to hopefully address both the incorrect attribute binding and the catastrophic object neglect problems in a unified framework. We summarize our **contributions** as follows: i) we introduce a novel object-conditioned EBAMA method to address both the incorrect attribute binding and the catastrophic object neglect problems in text-controlled image generation; and ii) extensive qualitative and quantitative experiments on the AnE benchmark demonstrate the superior performance of our method over strong previous approaches.

2. Background

For fair comparison with previous methods, we also conduct all experiments with open-sourced Stable Diffusion Models (SD) [18]. In the cross-attention mechanism, K is the linear projections of W_y, the CLIP-encoded text embeddings of text prompt y. Q is the linear projection of the intermediate image representation parameterized by latent variables z. Given a set of queries Q and keys K, the (unnormalized) attention features and (softmax-normalized) scores between these two matrices are

$$A = \frac{QK^T}{\sqrt{m}}, \quad \tilde{A} = \text{softmax}\left(\frac{QK^T}{\sqrt{m}}\right), \tag{1}$$

where m is the feature dimension. We consider both attention features and scores for our modeling here, which we denote as A_s and $\tilde{A}_s$ for token s, respectively.

3. Method

Following the pre-processing step in [17], we parse the prompt using Spacy's [7] transformer-based dependency parser to extract the object-oriented structure. We identify a set S of object tokens s from the prompt, whose tag is either

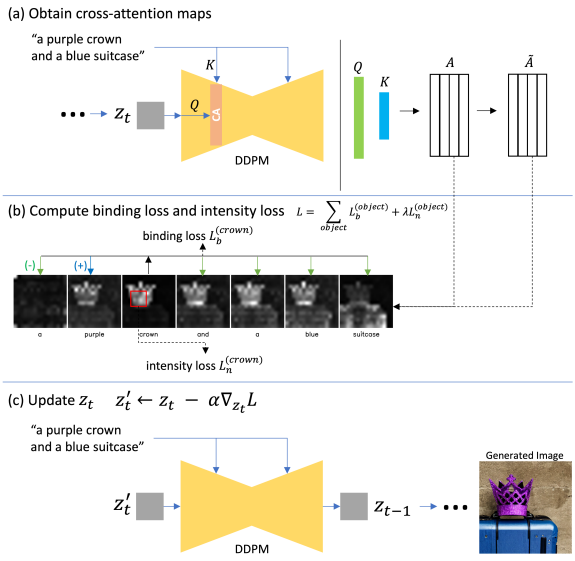

Figure 2. **An overview of our workflow for optimizing diffusion models.** It includes aggregation of attention maps, computation of object-centric attention loss, and updates to z_t.

NOUN (noun) such as 'backpack' or PROPN (proper noun) such as 'Tesla company' using the parser; we exclude nouns that serve as direct modifiers of other nouns. The remaining modifiers are grouped by their corresponding object tokens, denoted as the modifier sets for each object token s, i.e., $\mathcal{M}(s)$. Note that $\mathcal{M}(s) = \emptyset$ if there are no modifiers corresponding to the object token s.

3.1. Object-Conditioned Energy-Based Model

We assume that the distribution of the modifier tokens $l \in \bigcup_s \mathcal{M}(s)$ given the object token s is

$$p_z(l|s) = \frac{1}{Z(s)} \exp\left(f(A_l, A_s)\right), \tag{2}$$

where $Z(s) = \sum_l \exp\left(f(A_l, A_s)\right)$ is the normalizing constant and f is the negative energy function. We choose $f(A_l, A_s)$ as cosine similarity and consider attention features in Eqn. (1) as its input. Eqn. (2) therefore defines a multinomial token distribution as a z-parameterized conditional energy-based model, where z is the latent variables of SD. The inference-time optimization over the latent variables z is then equivalently maximizing the log-likelihood of this EBM, which increases the probabilities of the syntatically related modifier tokens of the given object s. To be specific, it can be shown that

$$\nabla_z \log p_z(l|s) = \nabla_z f(A_l, A_s) - \mathbb{E}_{p_z(l|s)}\left[\nabla_z f(A_l, A_s)\right]. \tag{3}$$

Since the vocabulary size of modifier tokens can be large in practice (in the order of 10^4), we consider resorting to nega-

tive sampling [11] for the approximation of the expectation term, where we uniformly sample tokens unrelated to the object token and calculate the Monte Carlo average. This particular implementation choice of Eqn. (3) then leads to the object-centric attribute binding loss below.

3.2. Object-Conditioned Energy-Based Attention Map Alignment

For each object token $s \in S$, we design the following two components that consist of the object-centric attention loss:

Object-centric attribute binding With the help of negative sampling, the attribute binding loss is:

$$L_b^{(s)} = -\frac{1}{|\mathcal{M}(s)|}\sum_{l\in\mathcal{M}(s)} f(A_s, A_l) + \frac{1}{N-|\mathcal{M}(s)|-1}\sum_{l\notin\mathcal{M}(s), l\neq s} f(A_s, A_l), \quad (4)$$

whose negative gradient w.r.t. z could be seen as the Monte Carlo approximation of Eqn. (3). The goal of $L_b^{(s)}$ is to: i) maximize the cosine similarity between the given object s and its syntactically-related modifier tokens, while ii) enforcing the repulsion of grammatically unrelated ones in the feature space. Note that the loss above only applies to the cases where $\mathcal{M}(s)$ is a non-empty set. For the case where $\mathcal{M}(s) = \emptyset$, only the second term of Eqn. (4) is used.

Object-centric intensity regularizer We observe that the object-related attention feature can still be overly shifted when there are multiple modifier tokens in the $\mathcal{M}(s)$ or multiple object tokens in a prompt; this could again potentially leads to the object neglect phenomenon. To address this issue, we follow [3] and propose an object-centric intensity regularizer to maintain the attention intensity level of object s:

$$L_n^{(s)} = -||\mathcal{K}(\tilde{A}_s)||_\infty, \quad (5)$$

where $\mathcal{K}$ is a 3x3 Gaussian kernel, and $||\cdot||_\infty$ denotes the maximum value of a vector.

The final object-centric attention loss L is the linear combination of the binding loss and the regularizer, i.e.

$$L = \sum_{s\in S} L^{(s)} = \sum_{s\in S} L_b^{(s)} + \lambda L_n^{(s)}, \quad (6)$$

where intensity weight λ is a hyper-parameter to specify. $\lambda > 0$ enforces the presence of object s, but excessively intensified object attention can hinder the attribute binding performance and lower visual image quality.

3.3. Workflow

Our workflow is illustrated in Fig. 2. To begin, at each time step t, we aggregate the attention map features denoted as A at a resolution of 16x16. Subsequently, we calculate the object-centric attention loss, as described in Eqn. (6). Finally, we backpropagate the computed loss and update z_t for each time step, following the formula $z_t' \leftarrow z_t - \alpha\nabla_{z_t} L$, where α represents the step size.

4. Experiments

We compare our generation results with previous methods including SD, CD, StrD, EBCA, AnE, and SG.

Datasets The AnE dataset [3] comprises three benchmarks: Animal-Animal, Animal-Object, and Object-Object. Each benchmark varies in complexity and incorporates a combination of potentially colored animals and objects. The prompt patterns for these benchmarks include two unattributed animals, one unattributed animal and one attributed object, and two attributed objects, respectively.

Metrics Full Sim. is the CLIP [14] cosine similarity score between the text prompt and the generated image. Furthermore, we assess CLIP similarity for the most neglected object independently from the full text by computing the CLIP similarity scores between each sub-prompt and the generated image. The smaller score is denoted as Min. Sim.. T-C Sim. is the average CLIP similarity between the prompt and all captions generated by a pre-trained BLIP image-captioning model [9] with the generated image as input. Recent work [2, 26] has found that large Vision-and-Language Models (VLMs) [9, 14, 20, 27] demonstrate a significant lack of compositional understanding, failing to reflect human preferences accurately. We suggest that the development of better metrics be a future research direction.

Quantatitive Comparison We generate 64 images for each prompt using the same seed across all methods and compute the average score between each prompt and its corresponding images. Our method consistently demonstrates superior performance across all datasets, as shown in Tab. 1. We stress the following advantages of our method: (1) Our method distinguishes itself from SG by its adaptability to the Animal-Animal dataset, even when the prompts lack specific attributes; (2) Our method with $\lambda = 0$ surpasses AnE and SG in all cases, underscoring the effectiveness of our object-centric attribute binding loss; (3) As the dataset becomes more complicated, our method with hyper-picked λ gains a more significant advantage over that with $\lambda = 0$.

Qualitative Comparison In Fig. 3, we identify recurrent failure modes in SG and AnE, attributable to the ineffectiveness of their objective design. AnE frequently strug-

Table 1. **Comparison of Full Sim., Min. Sim., and T-C Sim. across different methods on the AnE dataset.** Note that the performance of SG on Animal-Animal is degraded to SD, as the prompts do not contain any attribute-object pairs. The best and second-best performances are marked in bold numbers and underlines, respectively; tables henceforth follows this format.

	Animal-Animal			Animal-Object			Object-Object		
Method	Full Sim.↑	Min. Sim.↑	T-C Sim.↑	Full Sim.↑	Min. Sim.↑	T-C Sim.↑	Full Sim.↑	Min. Sim.↑	T-C Sim.↑
SD[18]	0.311	0.213	0.767	0.340	0.246	0.793	0.335	0.235	0.765
CD[10]	0.284	0.232	0.692	0.336	0.252	0.769	0.349	0.265	0.759
StrD[5]	0.306	0.210	0.761	0.336	0.242	0.781	0.332	0.234	0.762
EBCA[13]	0.291	0.215	0.722	0.317	0.229	0.732	0.321	0.231	0.726
AnE[3]	0.332	0.248	0.806	0.353	0.265	0.830	0.360	0.270	0.811
SG[17]	0.311	0.213	0.767	0.355	0.264	0.830	0.355	0.262	0.811
Ours($\lambda = 0$)	**0.340**	0.255	0.814	0.362	**0.271**	**0.851**	0.360	0.270	0.823
Ours	**0.340**	**0.256**	**0.817**	**0.362**	0.270	**0.851**	**0.366**	**0.274**	**0.836**

Figure 3. **Qualitative comparison on the AnE dataset.** Each column shares the same random seed.

gles with incorrect attribute association, whereas SG often fails to generate multiple objects simultaneously. Yet, our method attains high-quality semantic alignment with deliberately designed optimization objective, exhibiting more stable performance across different random seed selections.

5. Conclusion

We introduce an object-conditioned EBAMA framework to address the alignment issues in text-to-image diffusion models. We propose an object-centric attribute binding loss that maximizes the log-likelihood of the object-conditioned EBM in the attention feature space. An intensity regularizer is further designed to provide an extra degree of freedom balancing the trade-off between correct attribute binding and the necessary presence of objects. Extensive quantitative and qualitative comparisions demonstrate the superiority of our method in aligned text-to-image generation.

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. 1

[2] Yingshan Chang, Yasi Zhang, Zhiyuan Fang, Yingnian Wu, Yonatan Bisk, and Feng Gao. Skews in the phenomenon space hinder generalization in text-to-image generation. *arXiv preprint arXiv:2403.16394*, 2024. 3

[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. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1, 3, 4

[4] Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022. 1

[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 *The Eleventh International Conference on Learning Representations*, 2023. 1, 4

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[7] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1): 411–420, 2017. 2, 1

[8] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*, 2023. 1

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3

[10] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 1, 4

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 3

[12] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[13] Geon Yeong Park, Jeongsol Kim, Beomsu Kim, Sang Wan Lee, and Jong Chul Ye. Energy-based cross attention for bayesian context update in text-to-image diffusion models. *arXiv preprint arXiv:2306.09869*, 2023. 2, 4

[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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1

[16] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models. *arXiv preprint arXiv:2210.10606*, 2022. 1

[17] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment, 2023. 1, 2, 4

[18] 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 (CVPR)*, pages 10684–10695, 2022. 1, 2, 4

[19] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 1

[20] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 3

[21] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644. PMLR, 2016. 2

[22] 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, 2022. 1

[23] Peiyu Yu, Sirui Xie, Xiaojian Ma, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Unsupervised foreground extraction via deep region competition. *Advances in Neural Information Processing Systems*, 34:14264–14279, 2021. 2

[24] Peiyu Yu, Sirui Xie, Xiaojian Ma, Baoxiong Jia, Bo Pang, Ruiqi Gao, Yixin Zhu, Song-Chun Zhu, and Ying Nian Wu. Latent diffusion energy-based model for interpretable text modeling. *arXiv preprint arXiv:2206.05895*, 2022.

[25] Peiyu Yu, Yaxuan Zhu, Sirui Xie, Xiaojian Shawn Ma, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Learning energy-based prior model with diffusion-amortized mcmc. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[26] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-

language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022. 3

[27] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 3