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

Yasi Zhang, Peiyu Yu, and 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, including human evaluation, on several challenging benchmarks demonstrate the superior performance of our method over previous strong counterparts. With better aligned attention maps, our approach shows great promise in further enhancing the text-controlled image editing ability of diffusion models. The code is available at https://github.com/YasminZhang/EBAMA.

Keywords: Attention Map Alignment· Energy-Based Models · Text-to-Image Diffusion Models

1 Introduction

Recently, large-scale text-to-image diffusion models [2,11,19,22,26,27] 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 [5, 25]. Some well-known issues include omission, hallucination, or duplication of details [30], semantic leakage of attributes between entities [25], and miscomprehension of intricate textual descriptions [27].

Many previous works have focused on addressing the semantic misalignment issues, particularly concerning multiple-object generation and attribute binding. Composable Diffusion (CD) [16] composes multiple output noises guided

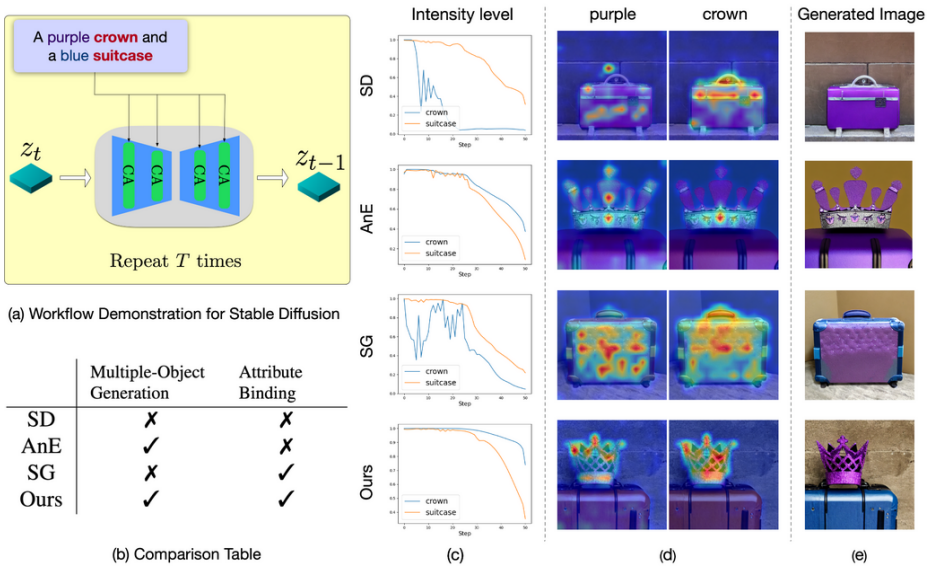

Fig. 1: Key observations of the generation process of diffusion models. The given prompt is "a purple crown and a blue suitcase". In panel (c), we hypothesize that if the intensity level of any object in the prompt does not remain high during the first half of the denoising process, e.g. the crown in SD and SG, the model would fail to generate the object in the final image. The panel (d) 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 panel (e). All methods share the same random seed.

by different objects in a text prompt during the generation process. However, this approach often results in a blend of objects, failing to distinctly separate them. Prompt-to-Prompt (PtP) [8] observes a strong correlation between cross-attention maps and the layout of an image. Building on this, Structured Diffusion (StrD) [7] experiments with averaging attention maps generated by different noun phrases for the same queried image latent representation. Yet, a simple average is inadequate for consistently generating images with multiple objects possessing complex attributes. Attend-and-Excite (AnE) [4] 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. Similarly, A-star (A*) [1] aims to minimize the intersection of different objects' attention maps. In response, SynGen (SG) [24] proposes an attribute-object pair-centric objective, aiming to minimize the distribution distance within the pair while maximizing it from other tokens, based on the assumption that normlized attention maps follow a multinomial distribution. Our findings (see Figs. 1 and 4-6) indicate that this approach

still struggles with object neglect due to its pair-centric nature. We argue that multiple-object generation is more critical than attribute binding, as attributes cannot manifest without the presence of objects. Furthermore, in scenarios with multiple objects and no explicit attributes in a prompt, SG is degraded to standard Stable Diffusion Models (SD) [26]. Diverging from these methods, Energy-Based Cross Attention (EBCA) [20] introduces an Energy-Based Model (EBM) framework [29, 31–34] 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. For example, in the image generated by SG, the crown is notably absent. Conversely, despite successful generation of both objects with strong intensities, AnE binds the attribute 'purple' incorrectly to the suitcase resulting from misaligned attention distributions of attribute-object pairs. 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. Notably, we show that approximately maximizing the log-likelihood for a z-parameterized EBM effectively leads to optimizing an object-centric binding loss, which emphasizes both the object attention map intensity levels and the attribute-object attention map alignment. We further develop an object-centric intensity regularizer to prevent excessive shifts of objects towards their attributes, providing an extra degree of freedom balancing the trade-off between correct attribute binding and the necessary presence of objects.

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; ii) extensive qualitative and quantitative experiments, including human evaluation, on several challenging benchmarks demonstrate the superior performance of our method over strong previous approaches. iii) We showcase that our approach has great promise in further enhancing the text-controlled image editing ability of diffusion models.

2 Related Work

EBM Framework for Attention Mechanisms Recent advancements in the theoretical exploration of attention mechanisms have increasingly embraced the EBM framework [13, 17, 23]. Modern Hopfield Networks [23] showcases that one of the proposed energy minima is equivalent to the attention mechanism. Building on this groundwork, Energy Transformer [12] designs an engineered energy

function to extract the relationships between tokens. Furthering this approach, EBCA [20] first formulates EBMs of query values condtioned on key values in each cross-attention layer. Similarly, our method seeks to exploit the theoretical potential of EBMs, focusing on the unique formulation of object-conditioned EBMs for attention maps.

Text-to-Image Diffusion Models Most large-scale text-to-image diffusion models [2, 11, 19, 22, 26, 27] utilize classifier-free guidance [9] for improved conditional synthesis results. However, due to its strong linguistic and visual priors injected from the training dataset, these models suffer from diverse semantic misalignment issues related to the objects in the provided text prompts and their attribute(s) [5, 25, 27, 30, 37]. Our approach better aligns images with its provided texts by mitigating the issues of object neglect and incorrect attribute binding without fine-tuning the diffusion models or additional training datasets.

Attention-Based Enhancement PtP [8] identifies a correlation between cross-attention maps and image layout. Expanding on this, StrD [7] experiments with averaging attention maps from different noun phrases to mitigate object neglect and attribute leakage. AnE [4] introduces a method to enhance object presence by maximizing attention map weights for object tokens. SG [24] proposes minimizing distribution distances of the attention maps within attribute-object pairs and maximizing the distances between the pairs and the other tokens. Different from the previous approaches, EBCA [20] adopts an EBM framework, focusing on updating text embeddings within cross-attention mechanisms. Our work also introduces an energy-inspired attention map alignment objective, while our objective has a specific emphasis on the object tokens.

3 Background

3.1 Stable Diffusion Models

For fair comparison with previous approaches, we also conduct experiments on open-sourced state-of-the-art Stable Diffusion Models (SD) [26]. SD first encodes an image x into the latent space using a pretrained encoder [6], i.e., $z = \mathcal{E}(x)$. Given a text prompt y, SD optimizes the conditional denoising autoencoder ϵ_θ by minimizing the objective

$$\mathcal{L}_\theta = \mathbb{E}_{t,\epsilon\sim\mathcal{N}(0,1),z\sim\mathcal{E}(x)}||\epsilon - \epsilon_\theta(z_t, t, \phi(y))||^2, \tag{1}$$

where ϕ is a frozen CLIP text encoder [21], z_t is a noised version of the latent z, and the time step t is uniformly sampled from $\{1, \ldots, T\}$. During sampling, z_T is randomly sampled from standard Gaussian and denoised iteratively by the denoising autoencoder ϵ_θ from time T to 0. Finally, a decoder $\mathcal{D}$ reconstructs the image as $\tilde{x} = \mathcal{D}(z_0)$.

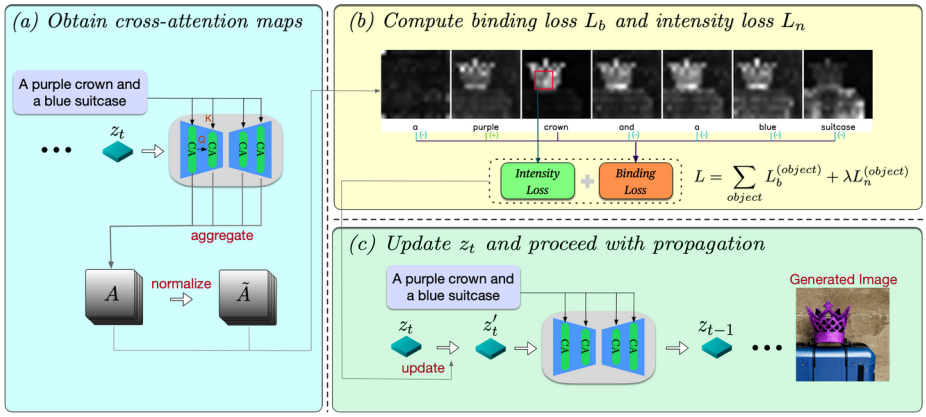

Fig. 2: An overview of our workflow for optimizing diffusion models. It includes aggregation of attention maps, computation of object-centric attention loss, and updates to z_t.

3.2 Cross-Attention Mechanism

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

$$A = \frac{QK^T}{\sqrt{m}}, \quad \tilde{A} = \text{softmax}\left(\frac{QK^T}{\sqrt{m}}\right), \tag{2}$$

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.

4 Method

The key idea of the proposed method derives from the *object-oriented structure* that underlies most prompts for text-to-image generation. To be specific, syntactically the majority of prompts can be parsed as the modifiers and the entity-nouns, i.e., the nouns that correspond to *objects* in the generated image, such as "A red metal *crown*", and "A *girl* in red", etc. Based on the observation, we propose to exploit the object-oriented structure by employing an ensemble of object-centric cross-attention losses for inference-time optimization. Our aim is to address the semantic misalignment issues including both the attribute binding (e.g., semantic leakage and attribute neglect [24]) and catastrophic object neglect [4], with the principally derived and deliberately designed optimization objective. We then discuss the key components of our method as follows.

4.1 Extraction of the Object-Oriented Structure

Following the pre-processing step in [24], we parse the prompt using Spacy's [10] 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 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. We refer to supplementary material for more details about the parsing process.

4.2 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

$$p_z(l|s) = \frac{1}{Z(s)} \exp\left(f(A_l, A_s)\right), \tag{3}$$

where $Z(s) = \sum_l \exp\left(f(A_l, A_s)\right)$ is the normalizing constant and f is the negative energy function. How to choose the energy function for attention maps remains an interesting and open problem. Prior works [20,23] utilize a log-sum-exp term to model the exponential interaction and alignment between state patterns (query) and stored patterns (key). However, the alignment measurement for attentions maps across different tokens is unclear. To bridge the gap, we propose the application of a non-crafted yet effective energy function — cosine similarity defined as $f(A_l, A_s) = \langle A_l, A_s \rangle / (||A_l|| \cdot ||A_s||)$. The efficacy of this choice of energy function is validated in the experimental section. Eqn. (3) 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 (see the supplementary material)

$$\nabla_z \log p_z(l|s) = \nabla_z f(A_l, A_s) - \mathbb{E}_{p_z(l|s)} \left[\nabla_z f(A_l, A_s)\right]. \tag{4}$$

Since the vocabulary size of modifier tokens can be large in practice (in the order of 10^4), we consider resorting to negative sampling [18] 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. (4) then leads to the object-centric attribute binding loss below.

4.3 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 First, instead of operating on the noun-modifier normalized attention score pairs as in [24], we focus on optimizing the log-likelihood of the object-conditioned EBM using negative sampling. This gives us the attribute binding loss:

$$L_b^{(s)} = -\frac{1}{|\mathcal{M}(s)|}\sum_{l\in\mathcal{M}(s)} f(A_s, A_l) + \frac{1}{N-|\mathcal{M}(s)|-1}\sum_{l\notin\mathcal{M}(s), l\neq s} f(A_s, A_l), \quad (5)$$

whose negative gradient w.r.t. z could be seen as the Monte Carlo approximation of Eqn. (4). 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 **repulsive term** of Eqn. (5) is used.

Object-centric intensity regularizer Although the proposed attribute binding loss mitigates the catastrophic object neglect problem (see $\lambda = 0$ entries in Tab. 1), 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 [4] and propose an object-centric intensity regularizer to maintain the attention intensity level of object s:

$$L_n^{(s)} = -||\mathcal{K}(\tilde{A}_s)||_\infty, \quad (6)$$

where $\mathcal{K}$ is a 3x3 Gaussian kernel, and $||\cdot||_\infty$ denotes the maximum value of a vector. We use the attention scores in Eqn. (2) as its input. We refer to $||\mathcal{K}(\tilde{A}_s)||_\infty$ as the **intensity level** of the object token s.

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

$$L = \sum_{s\in S} L^{(s)} = \sum_{s\in S} L_b^{(s)} + \lambda L_n^{(s)}, \quad (7)$$

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. We provide empirical analysis on how to tune the weight in practice (see Section 5.4 and supplementary material for details).

4.4 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. This aggregation is performed after one step of propagating z_t through the denoising model. Subsequently, we calculate the object-centric attention loss, as described in Eqn. (7). Finally, we backpropagate the computed loss and update z_t for each time

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 [26]	0.311	0.213	0.767	0.340	0.246	0.793	0.335	0.235	0.765
CD [16]	0.284	0.232	0.692	0.336	0.252	0.769	0.349	0.265	0.759
StrD [7]	0.306	0.210	0.761	0.336	0.242	0.781	0.332	0.234	0.762
EBCA [20]	0.291	0.215	0.722	0.317	0.229	0.732	0.321	0.231	0.726
AnE [4]	0.332	0.248	0.806	0.353	0.265	0.830	0.360	0.270	0.811
SG [24]	0.311	0.213	0.767	0.355	0.264	0.830	0.355	0.262	0.811
Ours($\lambda = 0$)	**0.340**	<u>0.255</u>	<u>0.814</u>	<u>0.362</u>	**0.271**	**0.851**	<u>0.360</u>	<u>0.270</u>	<u>0.823</u>
Ours	**0.340**	**0.256**	**0.817**	**0.362**	<u>0.270</u>	**0.851**	**0.366**	**0.274**	**0.836**

step, following the formula $z_t' \leftarrow z_t - \alpha \nabla_{z_t} L$, where α represents the step size. In our experimental setup, we set $\alpha = 20$. Note that we only perform updates on z_t during the first half of the sampling steps, which corresponds to 25 steps since we use a DDIM sampler with a total of 50 steps. More details about the workflow can be found in the supplementary material.

5 Experiments

We compare our generation results with SD, CD, StrD, EBCA, AnE, and SG on two artificial datasets, AnE dataset [4] and DVMP [24], and one natural-language dataset, ABC-6K [7]. We refer to supplementary material for implementation details and computational efficiency comparison.

Datasets The AnE dataset [4] 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. The DVMP dataset [24] features a diverse set of objects (e.g., daily objects, animals, fruits, etc.) and diverse modifiers including colors, textures and so on. It features more than three attribute descriptor per prompt. The ABC-6K dataset [7], derived from natural MSCOCO [15] captions, includes prompts with at least two color words modifying different objects. The first two datasets are artificial, while ABC-6K is composed of natural language captions.

5.1 Quantitative Comparison

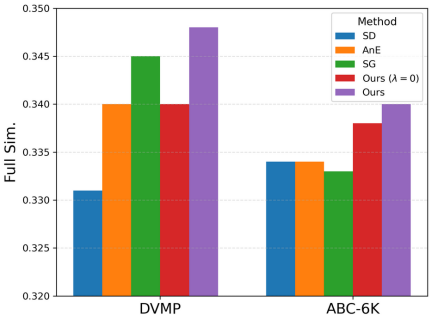

Fig. 3: Full Sim. results on DVMP and ABC-6K datasets. We randomly sample 200 prompts from each dataset and generate 4 images for each prompt.

Following the setting of [4], we compare the Text-Image Full Similarity (Full Sim.), Text-Image Min Similarity (Min. Sim.), and Text-Caption Similarity (T-C Sim.) on the AnE dataset. Additionally, we present the Full Sim. results on the DVMP and ABC-6K datasets.

Full Sim. is the CLIP [21] 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 [14] with the generated image as input.

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 $\lambda = 0$.

In Fig. 3, despite SG's deliberate design for multi-attribute prompts, our method consistently surpasses SG. Furthermore, in the ABC-6K dataset, AnE and SG exhibit performance levels similar to that of SD, while our method consistently achieves superior results. These advantages are further confirmed by our human evaluation in 5.3.

5.2 Qualitative Comparison

In Figs. 4-6, we identify recurrent failure modes in SG and AnE, attributable to the ineffectiveness of their objective design. AnE frequently struggles with incorrect attribute association, whereas SG often fails to generate multiple objects simultaneously. In contrast, our method attains high-quality semantic alignment with deliberately designed optimization objective. It also exhibits more stable performance across different random seed selections.

Fig. 4: Qualitative comparison on the AnE dataset. Each column shares the same random seed.

Fig. 5: Qualitative comparison on the DVMP dataset. Each column shares the same random seed.

Object Omission SG, due to its pair-centric approach, frequently omits objects, as evidenced by missing items like cars, apples, and crowns in Fig. 4, tomatoes, crowns, strawberries in Fig. 5, and earrings, ties, etc. in Fig. 6.

Fig. 6: Qualitative comparison on the ABC-6K dataset. Each column shares the same random seed.

Attribute Omission Due to a lack of concern for attribute tokens, AnE fails to overcome the strong visual priors over objects. e.g., the green apple in Fig. 4, the non-spotted dog in Fig. 5, and the brown cat in Fig. 6.

Attribute Leakage In the case of SG, examples include purple on the wall, blue spilled on the plants, and purple on the suitcase in Fig. 4, illustrating how attributes emerge as leakage when the respective object is absent. Additional examples include the tomato's color spilling onto the dog and red metal leaking onto the chair in Fig. 5, as well as blue color leaking and forming artifacts in Fig. 6. AnE, with its sole focus on intensity, also suffers significantly from attribute leakage, evident in the purple backpack and blue suitcase in Fig. 4, the red metal chair in Fig. 5, and the blue glasses and earrings, and red towel in Fig. 6.

We argue that addressing object neglect or attribute binding in isolation is insufficient, as these issues are intrinsically interconnected. Our method adeptly balances these two concerns with a chosen intensity weight λ, demonstrating its success in addressing the challenges above.

5.3 Human Evaluation

Recent work [3, 35] has found that large Vision-and-Language Models (VLMs) [14, 21, 28, 36] demonstrate a significant lack of compositional understanding, failing to reflect human preferences accurately. Given this, we conducted human evaluations across all three datasets to rigorously assess our model's performance.

Raters were enlisted online, with the requirement that each participant possessed an educational level of a bachelor's degree or higher. In the process of evaluation, they were presented with 2-way multiple choice problems consisting of a text prompt and two images generated by our method and one of four baselines, including SD, AnE, SG, and our method with $\lambda = 0$. For each dataset, 100 prompts were randomly sampled for evaluation. The effectiveness of prompt-image alignment was assessed by asking raters, "Which image better matches the given description?". More details are provided in supplementary material.

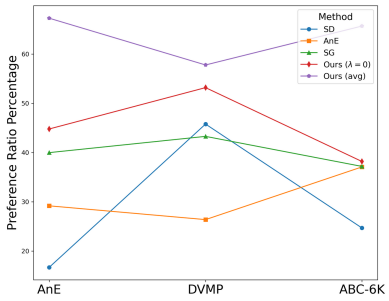

Fig. 7: Preference ratio percentage on text-image alignment by human evaluation. Ours(avg) represents the average preference ratio of our method compared with the other four methods.

The human evaluation results are shown in Fig. 7. We observe that: 1) our method consistently surpasses SD, AnE, and SG aligned with quantitative results in Tab. 1 and Fig. 3; 2) our method shows a more pronounced advantage over other methods on the natural-language ABC-6K dataset. We argue that our object-centric objective, in harmony with the object-oriented patterns prevalent in naturally occurring prompts, exhibits superior efficacy in handling complex real-world, natural-language-based prompts.

5.4 Ablation Study

Repulsive Term Tab. 2 presents the results of ours($\lambda = 0$) w/o and w/ the repulsive term in rows 1 and 2, and similarly, ours w/o and w/ this term in rows 3 and 4, under the same settings as Tab. 1. Row 2/4 demonstrates a significant performance increase than Row 1/3 due to the repulsive term, validating the effectiveness of negative sampling approximation.

Table 2: Ablation results on repulsive term. Both Ours and Ours($\lambda = 0$) benefit from the repulsive term as defined in Eqn. (5).

		Animal-Animal			Animal-Object			Object-Object		
Method	Repul.	Full Sim.	Min. Sim.	T-C Sim.	Full Sim.	Min. Sim.	T-C Sim.	Full Sim.	Min. Sim.	T-C Sim.
Ours($\lambda = 0$)	✗	0.311	0.213	0.767	0.343	0.246	0.794	0.334	0.237	0.765
	✓	**0.340**	0.255	0.814	0.362	**0.271**	**0.851**	0.360	0.270	0.823
Ours	✗	0.338	0.250	0.810	0.360	0.267	0.841	0.359	0.269	0.819
	✓	**0.340**	**0.256**	**0.817**	**0.362**	0.270	**0.851**	**0.366**	**0.274**	**0.836**

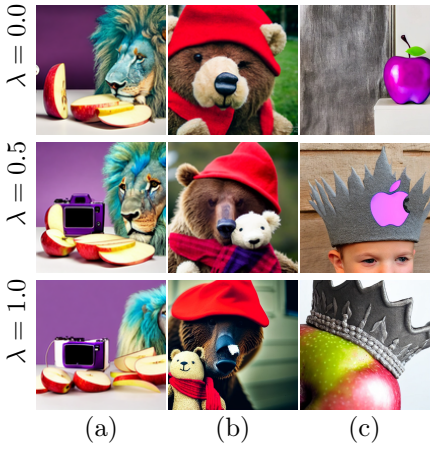

Fig. 8: Ablation demonstration for intensity weight λ. (a) a sliced apple and a purple camera and a teal lion; (b) a brown bear with red hat and scarf and a small stuffed bear; (c) a gray crown and a purple apple. We have selected one prompt from each dataset to showcase the stability of our method. Each column shares the same random seed.

Intensity Weight We demonstrate the impact of different choices for the intensity weight λ, which plays a role in enhancing the intensity level. In Fig. 8, we present some representative examples where the model needs to generate multiple objects with certain modifiers. When $\lambda = 0.5$, the generation is balanced. However, when $\lambda = 0.0$, all images more or less suffer from object neglect. Conversely, when $\lambda = 1.0$, artifacts are likely to appear and attribute binding becomes less effective.

Object-Conditioned Perspective Besides being able to handle more flexible prompts where no attribute-object pairs exist, our method is object-centric and thus, in the repulsive term, we only calculate the energy of objects and non-modifiers. The way SG handles it is to treat objects and modifiers equivalently when faced with non-modifiers. In the row '- obj cond.' of Tab. 3, we replace the energy of object and non-modifier $f(A_s, A_l)$ with the average energy of $f(A_s, A_l)$ and $f(A_s, A_m)$, where s, l, m represent object, non-modifier, and modifier, respectively. Note that as no modifiers exist in A-A, the results remain the same. In the other datasets, our object-conditioned perspective plays a vital role in the success of our method as the performance significantly decreases.

Energy function Choice To calculate KL div., SG assumes attention maps follow a multinomial distribution. Yet, cosine similarity does not pose any assumption on the distribution and achieves superior performance. In the row '- cos sim.' of Tab. 3, we replace cosine similarity with the average KL div..

Table 3: Ablations on object-conditioned perspective and energy function choice.

	A-A	A-O	O-O
ours($\lambda = 0$)	0.814	0.851	0.823
- obj cond.	0.814 (-0.0%)	0.832 (-2.2%)	0.813 (-1.2%)
- cos sim.	0.812 (-0.2%)	0.846 (-0.6%)	0.817 (-0.7%)
SG	0.767 (-5.8%)	0.830 (-2.5%)	0.811 (-1.5%)

5.5 Augmented Attribute Editing

PtP [8] allows local editing with word swap, adding new phrase, or re-weighting by replacing the attention

weights of unchanged tokens, or re-weighting the attention maps of target tokens. This heavily relies on the semantic coupling between tokens and their attention maps. In Fig. 9, we categorize the failure cases in PtP shown in the left panel into four situations: (a) ineffective editing with aligned text-to-image generation; (b) ineffective editing with incorrect attribute binding, e.g. the semantic leakage of 'pink'; (c) ineffective editing with object neglect, e.g. the 'apple'; and (d) insignificant editing with aligned text-to-image generation, e.g. the property 'metal' for the drum. In contrast, our method effectively enhances the semantic distribution of attention maps, allowing PtP with our approach to apply effective and significant local attribute editing to the original images.

6 Conclusion

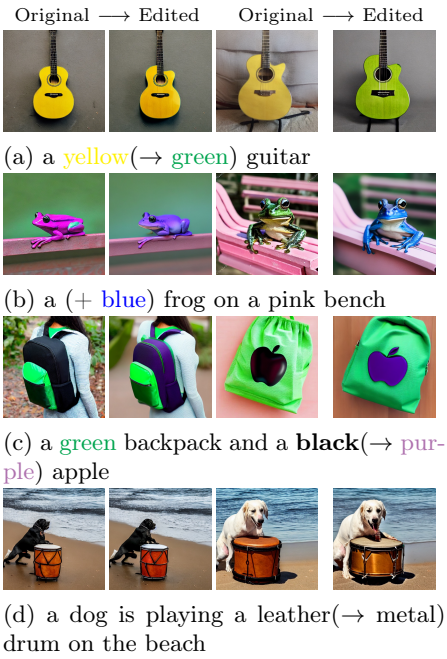

Fig. 9: Augmented attribute editing with our method. The left two columns demonstrates the attribute editing results from PtP, while the right two columns demonstrates the results from PtP w. ours.

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. This advancement promises great improvements in text-controlled attention-based image editing with semantically aligned attention maps.

Acknowledgements

The work was partially supported by NSF DMS-2015577 and a gift fund from Amazon. We truly thank the three anonymous reviewers for their valuable comments.

References

1. Agarwal, A., Karanam, S., Joseph, K., Saxena, A., Goswami, K., Srinivasan, B.V.: A-star: Test-time attention segregation and retention for text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2283–2293 (2023)
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
3. Chang, Y., Zhang, Y., Fang, Z., Wu, Y., Bisk, Y., Gao, F.: Skews in the phenomenon space hinder generalization in text-to-image generation. arXiv preprint arXiv:2403.16394 (2024)
4. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
5. Conwell, C., Ullman, T.: Testing relational understanding in text-guided image generation. arXiv preprint arXiv:2208.00005 (2022)
6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
7. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PUIqjT4rzq7
8. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=_CDixzkzeyb
9. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
10. Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear **7**(1), 411–420 (2017)
11. Hoogeboom, E., Heek, J., Salimans, T.: simple diffusion: End-to-end diffusion for high resolution images. arXiv preprint arXiv:2301.11093 (2023)
12. Hoover, B., Liang, Y., Pham, B., Panda, R., Strobelt, H., Chau, D.H., Zaki, M.J., Krotov, D.: Energy transformer. arXiv preprint arXiv:2302.07253 (2023)
13. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences **79**(8), 2554–2558 (1982)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)

15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
16. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
17. McEliece, R., Posner, E., Rodemich, E., Venkatesh, S.: The capacity of the hopfield associative memory. IEEE transactions on Information Theory **33**(4), 461–482 (1987)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)
19. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
20. Park, G.Y., Kim, J., Kim, B., Lee, S.W., Ye, J.C.: Energy-based cross attention for bayesian context update in text-to-image diffusion models. arXiv preprint arXiv:2306.09869 (2023)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
22. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
23. Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G.K., et al.: Hopfield networks is all you need. arXiv preprint arXiv:2008.02217 (2020)
24. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment (2023)
25. Rassin, R., Ravfogel, S., Goldberg, Y.: Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models. arXiv preprint arXiv:2210.10606 (2022)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
27. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
28. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022)
29. Xie, J., Lu, Y., Zhu, S.C., Wu, Y.: A theory of generative convnet. In: International Conference on Machine Learning. pp. 2635–2644. PMLR (2016)
30. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H.,

Baldridge, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation (2022)
31. Yu, P., Xie, S., Ma, X., Jia, B., Pang, B., Gao, R., Zhu, Y., Zhu, S.C., Wu, Y.N.: Latent diffusion energy-based model for interpretable text modeling. arXiv preprint arXiv:2206.05895 (2022)
32. Yu, P., Xie, S., Ma, X., Zhu, Y., Wu, Y.N., Zhu, S.C.: Unsupervised foreground extraction via deep region competition. Advances in Neural Information Processing Systems **34**, 14264–14279 (2021)
33. Yu, P., Zhang, D., He, H., Ma, X., Miao, R., Lu, Y., Zhang, Y., Kong, D., Gao, R., Xie, J., et al.: Latent energy-based odyssey: Black-box optimization via expanded exploration in the energy-based latent space. arXiv preprint arXiv:2405.16730 (2024)
34. Yu, P., Zhu, Y., Xie, S., Ma, X.S., Gao, R., Zhu, S.C., Wu, Y.N.: Learning energy-based prior model with diffusion-amortized mcmc. Advances in Neural Information Processing Systems **36** (2024)
35. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: 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)
36. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. arXiv preprint arXiv:2111.08276 (2021)
37. Zhang, Y., Yu, P., Zhu, Y., Chang, Y., Gao, F., Wu, Y.N., Leong, O.: Flow priors for linear inverse problems via iterative corrupted trajectory matching. arXiv preprint arXiv:2405.18816 (2024)