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

Supplementary Material

A. Object-Conditioned Energy-Based Model

In this section, we present the proof for Equ. (3). Specifically, we elaborate on the derivation of the gradient of the log-likelihood for the EBM as defined in Equ. (2):

$$
\begin{aligned}
&\nabla_z \log p_z(l|s) \\
=&\nabla_z \log(\exp(f(A_l, A_s))) - \nabla_z \log(\sum_l \exp(f(A_l, A_s))) \\
=&\nabla_z f(A_l, A_s) - \sum_l \frac{\exp(f(A_l, A_s))}{\sum_l \exp(f(A_l, A_s))} \nabla_z f(A_l, A_s) \\
=&\nabla_z f(A_l, A_s) - \sum_l p_z(l|s) \nabla_z f(A_l, A_s) \\
=&\nabla_z f(A_l, A_s) - \mathbb{E}_{p_z(l|s)} \left[\nabla_z f(A_l, A_s)\right].
\end{aligned}
$$

B. Algorithm

Our workflow can be outlined in Algo. 1. Initially, for the first half of the denoising steps, the latent variable z_t is updated using the gradient of the loss function, i.e. Equ. (6). The latter half of the denoising steps follows the standard generation process of diffusion models.

Algorithm 1 Energy-Based Attention Map Alignment

Input: A text prompt y, a set of object tokens S, a set of modifier tokens $\{\mathcal{M}(s)\}_{s\in S}$, a pretrained Stable Diffusion model SD, total sampling steps T, an image decoder $\mathcal{D}$
Output: An image x aligned with the prompt y

Initialize $z_T \sim \mathcal{N}(0, 1)$
for t in $T : [T/2] + 1$ **do**
  $\_, A, \tilde{A} \leftarrow SD(z_t, t, y)$
  Compute attention loss L according to Equ. (6)
  $z'_t \leftarrow z_t - \nabla_{z_t} L$
  $z_{t-1}, \_, \_ \leftarrow SD(z'_t, t, y)$
end for
for t in $[T/2] : 1$ **do**
  $z_{t-1}, \_, \_ \leftarrow SD(z_t, t, y)$
end for
$x \leftarrow \mathcal{D}(z_0)$
return x

C. Implementation Details

Experiments were conducted on a Linux-based system equipped with 4 Nvidia R9000 GPUs, each of them has 48GB of memory. To ensure a fair comparison with previous methods, we utilized the official Stable Diffusion v1.4 text-to-image model with the CLIP ViT-L/14 text encoder.

C.1. Hyperparameters

In our approach, we utilize a default fixed guidance scale of 7.5. The update step size is selected as $\alpha = 20$. We employ a DDIM sampler with a total of 50 steps. The update of the latent variable z_t is confined to the first half of the denoising process, which, in this context, corresponds to the initial 25 steps. Further discussion regarding the step size and the updated timesteps is in Appendix D.

C.2. Parser

Following [17], we utilize the spaCy parser [7], specifically employing the transformer-based `en_core_web_trf` model. Initially, we identify tokens within the prompt that are tagged as either NOUN or PNOUN , thereby constituting our object set. Subsequently, we extract all modifiers within this set based on a predefined set of syntactic dependencies, which include `amod`, `nmod`, `compound`, `npadvmod`, and `conj`. Finally, any NOUN or PNOUN that functions as a modifier for other entity-nouns within the object set is excluded.

C.3. Attention Map Extraction

The aggregated attention features A_t comprises N spatial attention maps, each corresponding to a token of the input prompt y. The CLIP text encoder appends a specialized $\langle$`SOT`$\rangle$ token at the beginning of y to signify the start of the text. It has been observed that in Stable Diffusion, the $\langle$`SOT`$\rangle$ token consistently receives the highest attention among all the tokens. Following [3], we exclude the attention allocated to $\langle$`SOT`$\rangle$ and then apply a softmax operation to the remaining tokens to obtain attention scores $\tilde{A}_t$.

D. Ablation Experiments

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.

Intensity Weight We demonstrate the impact of different choices for the intensity weight λ, which plays a role in

Table 2. **Ablation results on repulsive term.** Both Ours and Ours($\lambda = 0$) benefit from the repulsive term as defined in Eqn. (4).

Method	Repul.	Animal-Animal			Animal-Object			Object-Object		
		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**

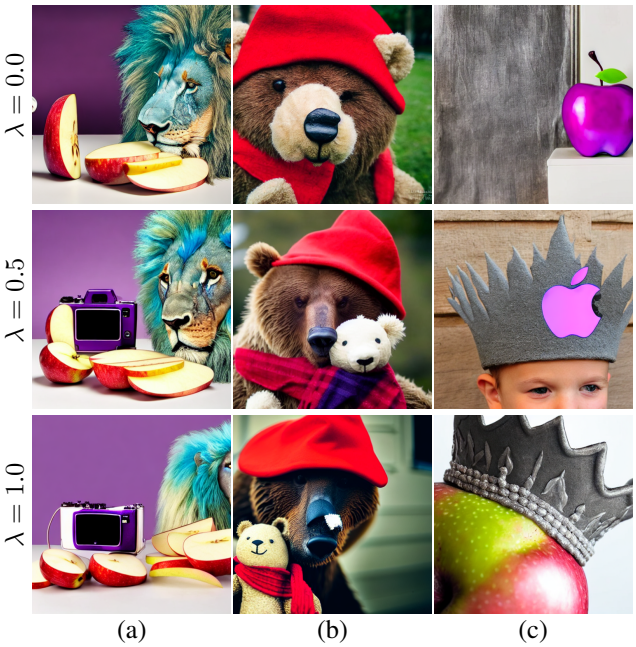

Figure 4. **Ablation demonstration for intensity weight λ.** (a) a sliced apple and a purple camera and a teal lion; (b) a brown bear with red hat and scarf and a small stuffed bear; (c) a gray crown and a purple apple. We have selected one prompt from each dataset to showcase the stability of our method. Each column shares the same random seed.

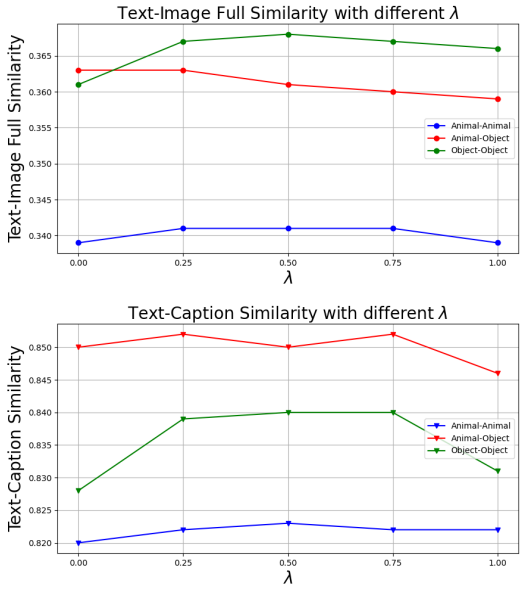

Figure 5. **Ablation study for λ.** We generated 10 images for each prompt with the same seed across all methods. The results indicate that for datasets with Animal-Animal and Object-Object pairings, a setting of $\lambda = 0.5$ is optimal; whereas for the Animal-Object dataset, $\lambda = 0.25$ yields the best performance.

enhancing the intensity level. In Fig. 4, 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, e.g. the camera in (a), the stuffed bear in (b), and the crown in (c). Conversely, when $\lambda = 1.0$, artifacts are likely to appear, e.g. the camera in (a), the brown bear in (b), and the apple in (c); attribute binding becomes less effective, e.g. the purple camera in (a), the stuffed bear in (b) and the purple apple in (c).

We explore various settings of the intensity weight parameter λ as illustrated in Fig. 5, where the metrics are computed across 10 images for each prompt. The values of Text-Image Full Similarity (Full. Sim.) and Text-Caption Similarity (T-C Sim.) are presented as functions of varying λ. At $\lambda = 0$, the intensity level is disregarded by the method. Conversely, increasing λ shifts the focus more towards the intensity level, at the expense of distribution alignment in attention maps.

For Animal-Animal and Object-Object , both metrics peak at $\lambda = 0.5$. For the Animal-Object dataset, the Text-Image similarity attains its highest score at $\lambda = 0$ or $\lambda = 0.25$. Given that Text-Caption Similarity is maximal at $\lambda = 0.25$, this value is selected for the Animal-Object dataset.

Our analysis indicates that λ effectively balances the trade-off between intensity level and attribute binding. Extremes of λ (e.g., 1.0 or 0.0) yield suboptimal generation results.

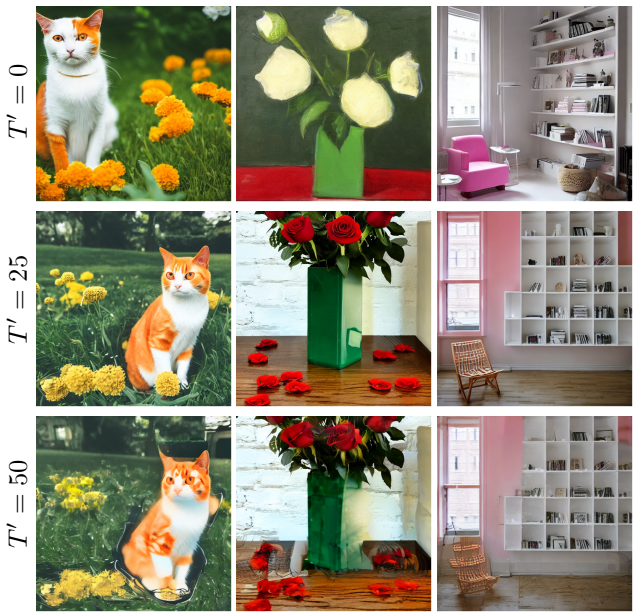

Figure 6. **Ablation demonstration for updated timesteps** T'. (a) an orange and white cat sitting in the grass near some yellow flowers; (b) red roses in a square green vase; (c) A room with pink walls and white display shelves and chair. Each column shares the same random seed.

Number of updated timesteps T' We explore different settings for the updated timesteps, denoted as T', which refer to the timestep numbers of updating the latent variable z_t. This exploration is depicted in Fig. 6. When $T' = 0$, our method defaults to the standard stable diffusion generation, with no updates applied to the model. In this configuration, due to the lack of interventions during the generation process, the generated images often exhibit semantic misalignments. Examples include the yellow flowers in (a), the red roses in (b), and the pink walls in (c). Conversely, setting $T' = 25$ implements our proposed method, which produces images better aligned with the input text. However, increasing T' to 50, where z_t is updated throughout the generation process, can introduce artifacts. Notable instances of these artifacts are visible in the representations of the cat in (a), the vase in (b), and the chair in (c).

Step Size α We investigate different settings for the step size α, as depicted in Fig. 7. When α is set to 1, the step size is too small, leading to insufficient attribute binding and the inability to generate multiple objects effectively. This is evident from the examples of blue walls in (b), a green pole in (c), and the missing crown in (a). Conversely, with α set to 40, the step size becomes excessively large, causing an overemphasis on certain attributes, e.g. blue in (a), blue in (b), black in (c) (note that the building behind is also black).

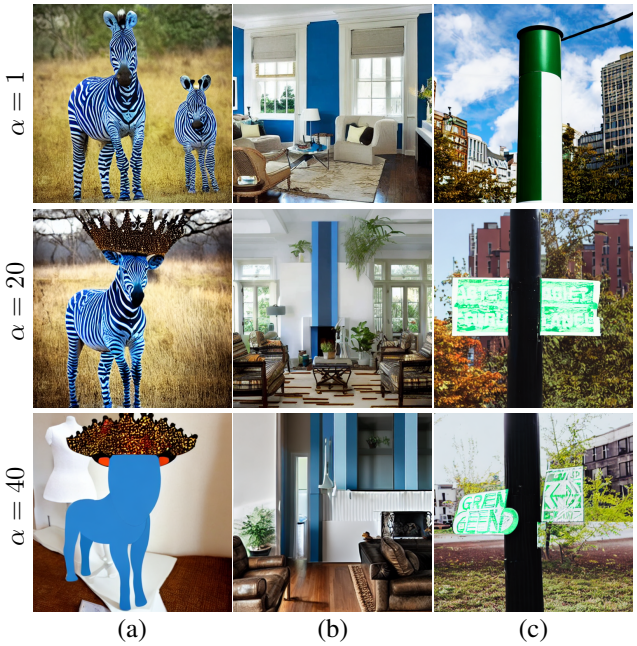

Figure 7. **Ablation demonstration for step size** α. (a) a blue zebra and a spotted crown; (b) a living room with white walls and blue trim; (c) a green and white sign on a **black** pole and some buildings. Each column shares the same random seed.