Supplementary Material

A Limitations

We list the limitations of our model as follows: (1) The effectiveness of our method is limited by the expressive power of the standard Stable Diffusion model. (2) Our energy-based model is conditioned on the object token $s \in S$. Typically, $|S| \neq 0$. However, when $|S| = 0$, it indicates that there are no objects in the prompt, and our method is degraded to regular diffusion model generation.

B Societal Impact and Ethical Concerns

This paper introduces a novel method for aligning attention maps to improve the compositional generation of images from text prompts, using diffusion models. Our approach aims to advance AI-assisted visual content creation in the creative, media, and communication sectors. Although we foresee no significant ethical issues unique to our method, it inherits general concerns associated with generative models, such as privacy, copyright infringement, misuse for deceptive content, and content bias.

C Object-Conditioned Energy-Based Model

In this section, we present the proof for Equ. (4). Specifically, we elaborate on the derivation of the gradient of the log-likelihood for the EBM as defined in Equ. (3):

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

D 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. (7). 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. (7)
    $z'_t \leftarrow z_t - \nabla_{z_t} L$
    $z_{t-1}, \_, \_ \leftarrow SD(z'_t, t, y)$
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

E 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.

E.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 H.

E.2 Parser

Following [24], we utilize the spaCy parser [10], 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.

E.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 ⟨SOT⟩ token at the beginning of y to signify the start of the text. It has been observed that in Stable Diffusion, the ⟨SOT⟩ token consistently receives the highest attention among all the tokens. Following [4], we exclude the attention allocated to ⟨SOT⟩ and then apply a softmax operation to the remaining tokens to obtain attention scores $\tilde{A}_t$.

E.4 Augmented Attribute Editing Setup

We utilize and slightly adapt the official repository from [8] for conducting attention editing. A cross-replace step of 0.8, i.e. replacing the first 80% steps of cross-attention maps, is employed for all editing tasks. Additionally, in line with the repository's provisions, self-attention maps also play a role in preserving the image's shape. For this purpose, we set the self-replace step at 0.4, i.e. replacing the first 40% steps of self-attention maps, for all the experiments.

E.5 A-Star Comparison

As A-STAR has not released their official code of implementation, we display their reported numeric results of T-C Sim. in Tab. 4. The table shows that our method demonstrates significantly superior performance when it comes to more complicated datasets. Note that A-A, A-O, and O-O include 0, 1, and 2 attributes in their prompts, respectively.

Table 4: Comparison with A-Star. * Values copied from the A-Star paper.

Method	A-A	A-O	O-O
AnE*	0.80 (-2.4%)	0.82 (-3.5%)	0.81 (-3.6%)
SG	0.77 (-6.1%)	0.83 (-2.4%)	0.81 (-3.6%)
A-STAR*	0.82 (-0.0%)	0.84 (-1.2%)	0.82 (-2.4%)
Ours	**0.82**	**0.85**	**0.84**

F Computational Efficiency

We randomly sampled a total of 100 prompts from the ABC-6K dataset and compared the time required for each method. Since there are no updates in SD, it provides us with the lower bound of the time needed for generating the images. As shown in Tab. 5, we can observe that AnE takes the longest time, approximately 47.0 minutes, to complete the generation process, while our method and SG require less than half of that time. The reason for this discrepancy is that AnE employs a technique called iterative latent refinement, which may require multiple updates at certain steps when the loss objective does not meet specific thresholds. We emphasize that our method is computationally efficient while still achieving top performance.

Table 5: Computational time comparison. We compare the time required for generating images for 100 prompts in ABC-6K dataset for each method.

	SD	AnE	SG	Ours
Time(min)	13.2	47.0	21.9	20.9

G External modifiers incorporation

We modify our workflow shown in Fig. 2 to include external modifiers without affecting the original prompt's guidance: (a) We first append the modifier(s) to the prompt to obtain attention maps for both original and external tokens. (b) We calculate the final loss using these attention maps. (c) After updating z_t, we drop the external modifier(s) and proceed with the original prompt. Tab. 6 presents the results of incorporating arbitrary syntactically unrelated modifiers, with their number denoted as n. The modifiers chosen are 'red'($n = 1$), and 'red blue green'($n = 3$). The performance decreases with n increasing. This decline may be due to external tokens interfering with the repulsion of other non-modifier tokens, as the attribute binding loss has to balance the repulsion of external tokens and that of non-modifier tokens.

n	A-A	A-O	O-O
0	**0.340/0.256/0.817**	**0.362/0.270/0.851**	**0.366/0.274/0.836**
1	0.337/0.254/0.812	0.360/0.269/0.844	0.364/0.272/0.826
3	0.331/0.248/0.822	0.357/0.266/0.846	0.362/0.270/0.823

Table 6: Results on external modifiers incorporation under the same setting as Tab. 1 . n denotes the number of external modifiers appended.

H Additional Ablation Experiments

Intensity Weight λ We explore various settings of the intensity weight parameter λ as illustrated in Fig. 10, 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.

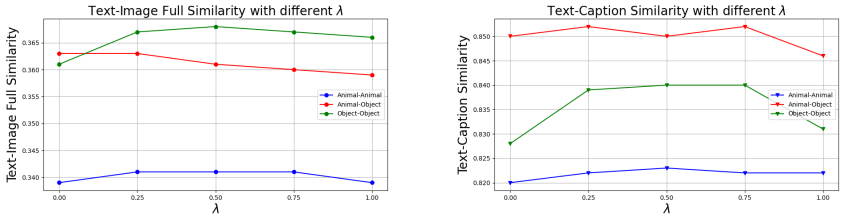

Fig. 10: Ablation study for λ. We generated 10 images for each prompt with the same seed across all methods. The results indicate that for datasets with Animal-Animal and Object-Object pairings, a setting of $\lambda = 0.5$ is optimal; whereas for the Animal-Object dataset, $\lambda = 0.25$ yields the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

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. Based on these findings, and considering that the ABC-6K dataset is more akin to the Object-Object dataset – notably, they both contain at least two attributes and two objects – we select $\lambda = 0.5$ for the ABC-6K dataset in all our experiments. The situation with the DVMP dataset is more intricate due to its potential for containing one to three objects and an unlimited number of attributes. As we suggest in the main text, selecting different intensity weights based on the number of objects in a prompt is recommended to optimize our method's performance.

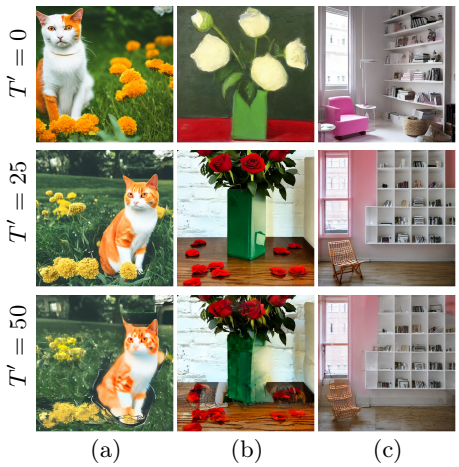

Fig. 11: Ablation demonstration for updated timesteps T'. (a) an orange and white cat sitting in the grass near some yellow flowers; (b) red roses in a square green vase; (c) A room with pink walls and white display shelves and chair. Each column shares the same random seed.

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. 11. 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).

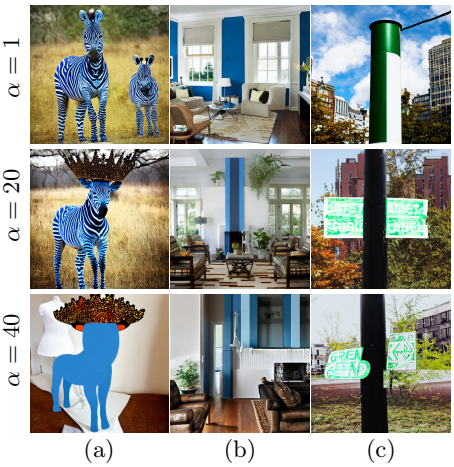

Fig. 12: Ablation demonstration for step size α. (a) a blue zebra and a spotted crown; (b) a living room with white walls and blue trim; (c) a green and white sign on a **black** pole and some buildings. Each column shares the same random seed.

Step Size α We investigate different settings for the step size α, as depicted in Fig. 12. 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).

I Additional Details on Human Evaluation

Raters were enlisted via an online platform under conditions of anonymity, with the requirement that each participant possessed an educational level of a bachelor's degree or higher. Additionally, they were assured of the protection of their privacy and the confidentiality of their identities throughout the process.

Fig. 13 provides a screenshot of the rating interface. The order of images was randomized.

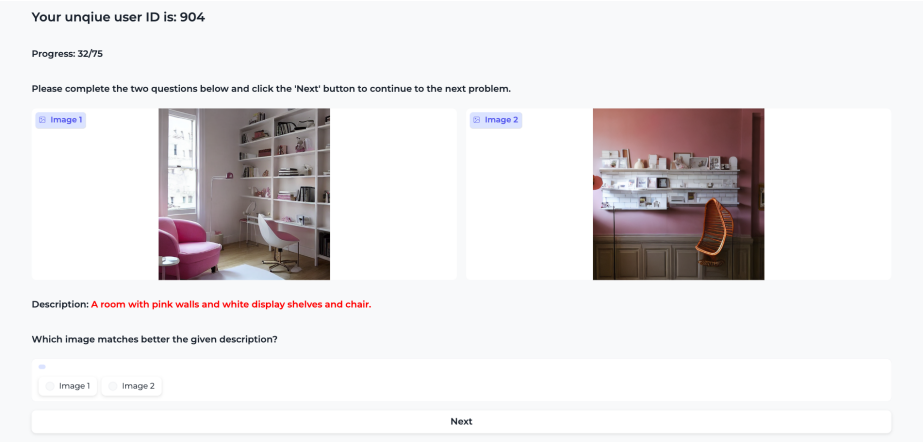

Fig. 13: A screenshot of the rating interface. The order of images was randomized.