Improving Compositional Generation of Diffusion Models via Adversarial Inference-Time Optimization

Wenxuan Tan*, Sonia Cromp, Chenghui Li, Jitian Zhao, Samuel Guo

CS 839 Project Report

1 Introduction

While demonstrating remarkable performance on a variety of tasks, Vision-Language foundation models (VLM) continue to struggle with compositional reasoning and generation. VLMs, regardless of model or training dataset size, show bag-of-word behaviors when matching complicated texts with images (see [8], [10]). For example, with the query "red fox and grey dog", VLMs may mismatch colors and attributes by generating a grey fox and red fox, or even ignore certain objects during generation when the query includes multiple objects. SOTA text-to-image (T2M) latent diffusion models (LDM) largely struggle to avoid: **(1) attribute mixing and (2) missing objects** (see fig 5).

The aforementioned issues are well known in many works (see a discussion in section 2). To provide a concrete example, we contrast our proposed adversarial inference-time optimization model with stable diffusion and attend-and-excite model [2], as an improved stable diffusion model in Figures 1, 2 and 3. We consider two specific contrastive loss functions including cosine similarity and distance correlation. The images generated from the attend-and-excite model [2] misses some objects in fig 1 and mixes the different color attributes in figures 2 and 3. On the contrary, our proposed adversarial inference-time optimization model with cosine similarity in fig 1, 2 and 3 successfully generate different objects with correct attributes.

2 Related Work

There were some potential solutions in the literature, while none of them can fully solve those generation issues.

Extra training/conditioning T2M models use text prompts as input and condition image embeddings on it. To address the missing attributes, [11] increased the size of the model to a 3-times larger, two-stage LDM with multi-resolution images with size (height, width) as an extra token. However, this method still suffers from attribute mixing. Other works explicitly control

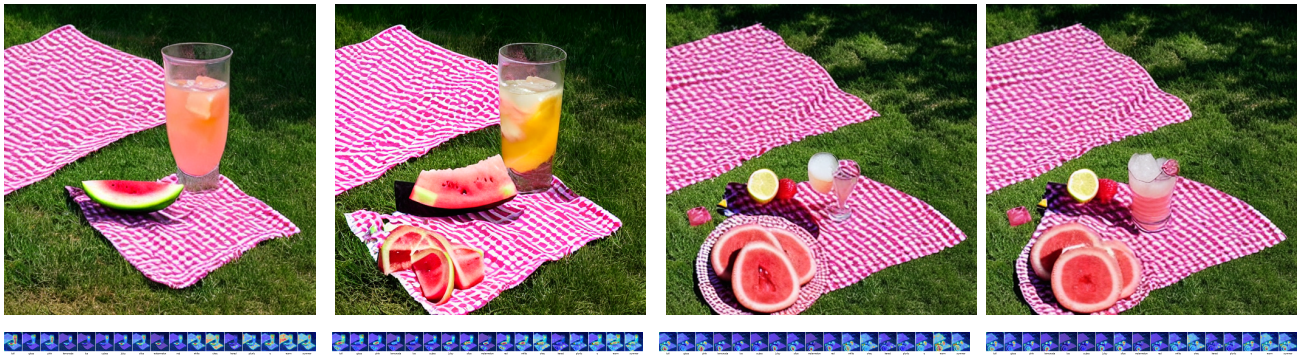
*Project lead

(a) Cosine similarity (b) Distance correlation (c) Attend and excite (d) Stable diffusion

Figure 1: Prompt: A red apple, a yellow banana, and a green pear in a fruit basket

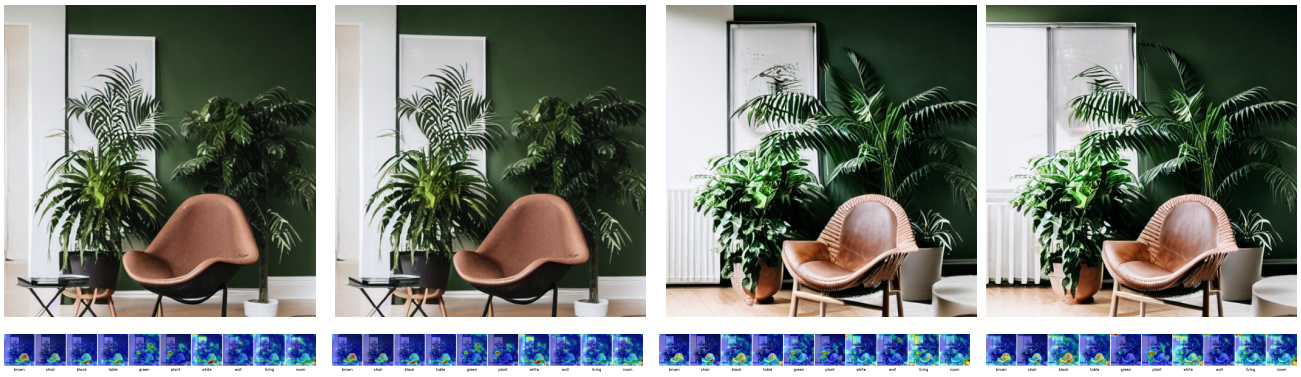
(a) Cosine similarity (b) Distance correlation (c) Attend and excite (d) Stable diffusion

Figure 2: Prompt: A yellow flower in a blue vase



(a) Cosine similarity (b) Distance correlation (c) Attend and excite (d) Stable diffusion

Figure 3: Prompt: A tall glass of pink lemonade with ice cubes, a juicy slice of watermelon, a red and white checkered picnic blanket, and a warm summer sun.



(a) Cosine similarity (b) Distance correlation (c) Attend and excite (d) Stable diffusion

Figure 4: Prompt:

the cross-attention maps by predicting bounding boxes (masks) for each object (e.g. [12]), or train the LDM to follow editing instructions (e.g. [1]). In general, these methods are prohibitive to scale up due to the cost of annotating text-image pairs.

Training-free control of attention maps In stable diffusion and most of T2M models, text conditioning is imposed via cross-attention layers (see architecture in fig 6). The attention score of token l with image patch j is computed as

$$\mathbf{A}_{lj} = \text{Softmax}\left(\frac{\mathbf{Q}_j \mathbf{K}_l^T}{\sqrt{d}}\right),$$

Given an image divided into n patches, where $n = \frac{h}{\text{patch\_size}} \times \frac{w}{\text{patch\_size}}$, the attention map for the l^{th} token can be represented as a matrix $\mathbf{M}_l^{h/\text{patch\_size} \times w/\text{patch\_size}}$ where the element $\mathbf{M}_{l,ij}$ represents the attention score of the token at patch (i, j). During the optimization, $\mathbf{M}_l$ is reshaped into a vector $\mathbf{A}_l^n$.

It is known that missing objects occurs when one object's map overlaps with and is suppressed by another, and attribute mixing happens when one token's map attends to the wrong object (see fig 6). Structured guidance [4] first proposed to parse the prompt into phrases and linearly combine the Keys or Values of objects/attributes to strengthen their effect in the cross-attention layers. Unfortunately, this work largely cherry-picks results so their experimental results are unreliable. Prompt-to-prompt editing [7] empirically scales (e.g. multiply by 6) the cross-attention maps of attributes to effectively avoid mixing/missing attributes, but does not fix missing object issues. Attend and excite [2] is a popular method to avoid missing objects by simultaneously penalizing two maps. However, it simply performs gradient ascent on the weakest map at each denoising step, so it could be unstable for the attribute mixing issues. Recently, [6] (see fig 7) combines various controls on attention maps to allow scaling and attribute editing on segmented map areas, but its multi-region-based denoising is costly. Another line of works [3, 5, 13] use LLMs to predict bounding box coordinates as extra or segment the salient regions in the maps to enforce concentration. However, we argue that these methods are either ad-hoc, rely on auxiliary modules, or can't handle both challenges.

3 Method

In this project, we aim to develop a method that

- addresses both attribute mixing and missing objects.
- does not rely on extra annotated data, denoising multiple segmented regions, or auxiliary modules, all of which can slow down inference.
- can be interpreted theoretically.

We propose adding an adversarial inference-time optimization process (similar to that in attend and excite) to handle both **intra-phrase** and **inter-phrase** overlap/correlation. Given prompts in the form of "a <adjectives> <obj 1>, a <adjectives> <obj 2>, and a <adjectives> <obj 3>,..." we first use a syntax parser to separate the prompt into phrases:

`"a <adj 1>..<adj n_1> <obj 1> | a <adj 2>..<adj n_2> <obj 2>..."`

In general, our parser would give us m phrase groups where the i^{th} phrase has an *object* word (also referred to as the *anchor token*) and n_m associated *adjective/attribute tokens*. Intuitively, there are several behaviors from inter- and intra-phrase groups we'd expect:

- Each object should have some minimum presence in the attention map; this is same as original Attend-and-Excite, namely discourages neglect.
- The attention maps of an object and its associated attributes should have small distance (high overlap) with each other; namely, it encourages proper attribute binding.
- Different objects' attention maps should have large distance (low overlap) with each other; namely, it discourages mixing objects.

To facilitate these additional constraints when computing the overall loss term, we operate on the flattened attention maps v averaged over all layers for each token, rather than the maximal attention per token used in the original Attend-and-Excite work. After parsing the phrase groups and adjective/object tokens, our overall loss function can be expressed as follows:

$$\mathcal{L} = \sum_{i=1}^{m} \sum_{j=1}^{n_m} d\left(v_{\text{adj}_j}, v_{\text{obj}_i}\right) + \alpha \sum_{i=1}^{m} 1 - \max v_{\text{obj}_i} + (1-\alpha) \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} 1 - d\left(v_{\text{obj}_i}, v_{\text{obj}_k}\right)$$

where d is any distance metric used to compare attention maps; we experiment with L1 loss, cosine similarity, distance correlation [14] and Wasserstein Distance for multivariate gaussians.

The hyperparameter α controls a trade-off in the loss between exciting each individual object's attention map and enforcing non-overlap between the attention maps of different objects, respectively.

Note that in the first term $d\left(v_{\text{adj}_j}, v_{\text{obj}_i}\right)$, the gradient is *not* propagated through the object to update its attention map, as we only want update the adjectives' attention maps to match their object anchors.

3.1 Evaluation

Many works hired contract workers for subjective evaluation. However, for subjective evaluation we intended to use the near SOTA object detection model DETA [9] trained on COCO. COCO has annotated text-image pairs, so we can use the text as prompt and detect how many objects are present. Since we haven't had time to tune the lr to achieve major gain on our toy prompts, we've decided to halt formal COCO evaluation.

4 Conclusion

In this report, we propose an adversarial inference-time optimization step to improve the diffusion's ability to prevent mixing attribute and missing objects. Instead of using maximum linking function between losses as done in [2], we tested cosine similarity, distance correlation and Wasserstein distance. This step makes the optimization step over a more regular function, which improves the efficiency and accuracy of the diffusion model. We observe partial improvement on the generated images compared with the attend and excite model, and the COCO accuracy report is put in the future work.

5 Discussion

Both cosine similarity and distance correlation are proven useful in the attend and excite structure. In addition to these linear linking loss function, the Wasserstein distance seems also useful to link the relationship among the attributes and objects. Currently Wasserstein distance is not working as we only have one sample per distribution(per token attention map) so the covariance matrix is 0, but we'll fix this by treating token in a phrase as belonging to the same distribution. We are also interested in developing a theoretical structure for the contrastive function..

Figure 5: Four random trials with Stable diffusion v2. Backpack is missing in 1, car is missing in 2 and the color is mixed.

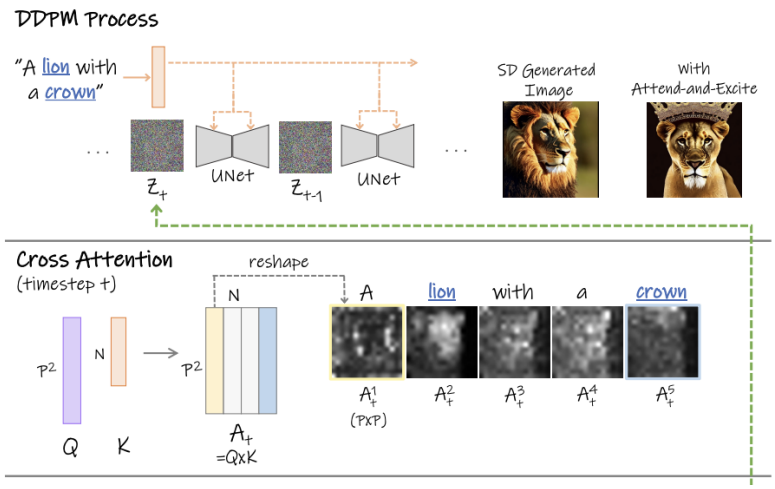


Figure 6: Cross-attention

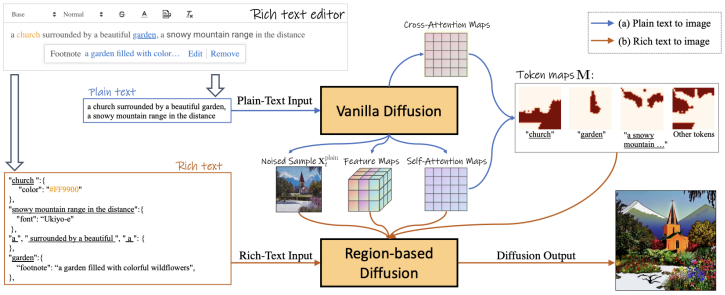


Figure 2. **Rich-text-to-image framework.** First, the plain-text prompt is processed by a diffusion model to collect self- and cross-attention maps, noised generation, and residual feature maps at certain steps. The token maps of the input prompt are constructed by first creating a segmentation using the self-attention maps and then labeling each segment using the cross-attention maps. Then the rich texts are processed as JSON to provide attributes for each token span. The resulting token maps and attributes are used to guide our region-based control. We inject the self-attention maps, noised generation, and feature maps to improve fidelity to the plain-text generation.

Figure 7

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[2] 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.

[3] Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2174–2183, 2023.

[4] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[5] Weixi Feng, Wanrong Zhu, Tsu jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models, 2023.

[6] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7545–7556, 2023.

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.

[8] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023.

[9] Jeffrey Ouyang-Zhang, Jang Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. Nms strikes back, 2022.

[10] Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Cross-modal attention congruence regularization for vision-language relation alignment. *arXiv preprint arXiv:2212.10549*, 2022.

[11] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[12] Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models, 2023.

[13] Jiayu Xiao, Liang Li, Henglei Lv, Shuhui Wang, and Qingming Huang. Rb: Region and boundary aware zero-shot grounded text-to-image generation, 2023.

[14] Xingjian Zhen, Zihang Meng, Rudrasis Chakraborty, and Vikas Singh. On the versatile uses of partial distance correlation in deep learning. In *European Conference on Computer Vision*, pages 327–346. Springer, 2022.