TRAINING-FREE STRUCTURED DIFFUSION GUIDANCE FOR COMPOSITIONAL TEXT-TO-IMAGE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale diffusion models have achieved state-of-the-art results on text-to-image synthesis (T2I) tasks. Despite their ability to generate high-quality yet creative images, we observe that attribution-binding and compositional capabilities are still considered major challenging issues, especially when involving multiple objects. In this work, we improve the compositional skills of T2I models, specifically more accurate attribute binding and better image compositions. To do this, we incorporate linguistic structures with the diffusion guidance process based on the controllable properties of manipulating cross-attention layers in diffusion-based T2I models. We observe that keys and values in cross-attention layers have strong semantic meanings associated with object layouts and content. Therefore, by manipulating the cross-attention representations based on linguistic insights, we can better preserve the compositional semantics in the generated image. Built upon Stable Diffusion, a SOTA T2I model, our structured cross-attention design is efficient that requires no additional training samples. We achieve better compositional skills in qualitative and quantitative results, leading to a very significant 5-8% advantage in head-to-head user comparison studies. Lastly, we conduct an in-depth analysis to reveal potential causes of incorrect image compositions and justify the properties of cross-attention layers in the generation process.

1 INTRODUCTION

Text-to-Image Synthesis (T2I) is to generate natural and faithful images given a text prompt as input. Recently, there has been a significant advancement in the quality of generated images by the extremely large-scale vision-language models, such as DALL-E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), and Parti (Yu et al., 2022). In particular, Stable Diffusion (Rombach et al., 2022) is the state-of-the-art open-source implementation showing superior evaluation metric gains after training over billions of text-image pairs.

In addition to generating high-fidelity images, the ability to compose multiple objects into a coherent scene is also essential. Achieving such ability requires the model to understand both the full prompt and individual linguistic concepts from the prompt. As a result, the model should be able to combine multiple concepts and generate novel objects that have never been included in the training data. In this work, we mainly focus on improving the compositionality of the generation process, as it is essential to achieve controllable and generalized text-to-image synthesis with multiple objects in a complex scene.

Attribute binding is a critical compositionality challenge (Ramesh et al., 2022; Saharia et al., 2022) to existing large-scale diffusion-based models. Despite the improvements in generating multiple objects in the same scene, existing models still fail when given a prompt such as "a brown bench in front of a white building" (see Fig. 1). The output images contains "a white bench" and "a brown building" instead, potentially due to strong training set bias or imprecise language understanding. From a practical perspective, explaining and solving this type of two-object binding challenge is a primary step to understanding more complex prompts with multiple objects. Therefore, how to bind the attributes to the correct objects is a fundamental problem for a more complicated and reliable compositional generation. While previous work has addressed compositional T2I (Park et al., 2021), our work tackles open-domain foreground objects with counterfactual attributes, such as color and material.

Figure 1: **Three challenging phenomena in the compositional generation.** Attribute leakage: The attribute of one object is (partially) observable in another object. Interchanged attributes: the attributes of two or more objects are interchanged. Missing objects: one or more objects are missing. With slight abuse of attribute binding definitions, we aim to address all three problems in this work.

Even though state-of-the-art (SOTA) T2I models are trained on large-scale text-image datasets, they can still suffer from inaccurate results for simple prompts similar to the example above. Hence, we are motivated to seek an alternative, data-efficient method to improve the compositionality. We observe that the attribute-object relation pairs can be obtained as text spans for free from the parsing tree of the sentence. Therefore, we propose to combine the structured representations of prompts, such as a constituency tree or a scene graph, with the diffusion guidance process. Text spans only depict limited regions of the whole image. Conventionally, to map their semantics into corresponding images, we need spatial information such as coordinates (Yang et al., 2022) as input. However, coordinate inputs cannot be interpreted by T2I models. Instead, we make use of the observations that attention maps provide free *token-region associations* in trained T2I models (Hertz et al., 2022). By modifying the key-value pairs in cross-attention layers, we manage to map the encoding of each text span into attended regions in 2D image space.

In this work, we discover similar observations in Stable Diffusion (Rombach et al., 2022) and utilize the property to build structured cross-attention guidance. Specifically, we use language parsers to obtain hierarchical structures from the prompts. We extract text spans across all levels that include visual concepts or entities and encode them separately to disentangle the attribute-object pairs from each other. Compared to using a single sequence of text embedding for guidance, we improve the compositionality by multiple sequences where each emphasizes an entity or a union of entities from multiple hierarchies in the structured language representations. We refer to our method as Structured Diffusion Guidance (StructureDiffusion). Our contributions can be summarized as three-fold:

- We propose an intuitive and effective method to improve compositional text-to-image synthesis by utilizing structured representations of language inputs. Our method is efficient and training-free that requires no additional training samples.
- Experimental results show that our method achieves more accurate attribute binding and compositionality in the generated images. We also propose a benchmark named **A**ttribute **B**inding **C**ontrast set (ABC-6K) to measure the compositional skills of T2I models.
- We conduct extensive experiments and analysis to identify the causes of incorrect attribute binding, which points out future directions in improving the faithfulness and compositionality of text-to-image synthesis.

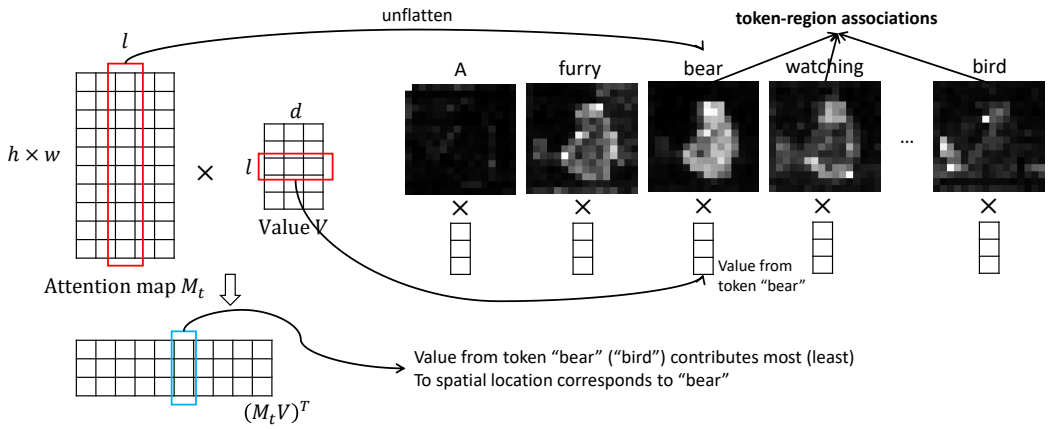

Figure 2: An illustration of cross attention operations and the token-region associations from attention maps. We omit some tokens for simplicity.

2 Diffusion Models & Structured Guidance

In this section, we propose a simple yet effective approach incorporating structured language representations into the cross-attention layers. We briefly introduce the Stable Diffusion model and its critical components in Sec. 2.1. Then, we present our method in detail in Sec. 2.2.

2.1 Background

Stable Diffusion We implement our approach and experiments on the state-of-the-art T2I model, Stable Diffusion (Rombach et al., 2022). It shares similar architecture with the Latent Diffusion Model (Rombach et al., 2022) and belongs to the class of Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020b). Stable Diffusion mainly consists of an autoencoder and a diffusion model. The pre-trained autoencoder encodes images into lower-resolution latent maps for diffusion model training. It decodes outputs from the diffusion model into images during inference. The diffusion model generates lower-resolution latent maps. Given a random Gaussian noise map z^T, it outputs a noise estimation ϵ at each step t and subtracts it from z^t. The final noise-free latent map prediction z^0 is fed into the autoencoder to generate images. Stable Diffusion adopts a modified UNet (Ronneberger et al., 2015) for noise estimation and a frozen *CLIP text encoder* (Radford et al., 2021) to generate a conditional embedding sequence from text input. The interactions between the image feature space and the textual embeddings are achieved through multiple *cross-attention layers* in both downsample and upsampling blocks.

CLIP Text Encoder Given a tokenized prompt input $\mathcal{T} = [t_1, t_2, \ldots, t_l]$, the CLIP text encoder outputs the sequence embeddings $\mathcal{W}_{\text{full}}$ from the last hidden state, i.e. $\mathcal{W}_{\text{full}} = \text{CLIP}_{\text{text}}(\mathcal{T})$ where c_{text} is the embedding dimension and l is the sequence length. Our key observation is that the contextualized CLIP embedding sequence is a potential cause of incorrect attribute binding. Due to the causal attention mask, attributes and object tokens in the later part of a sequence are contaminated with the attribute information before them. For example, When the user indicates some rare color for the second object (e.g. "a yellow apple and red bananas"), Stable Diffusion is prone to generate "banana" in "yellow" instead, as the embeddings of "yellow" is attended by token "banana".

Cross Attention Layers The cross attention layers take the embedding sequences from the CLIP text encoder and fuse them with latent feature maps to achieve classifier-free guidance. Denote a 2D feature map $\mathcal{X}^t$, it is projected into a query tensor by a linear layer $f_Q(\cdot)$ and reshaped as $Q^t \in R^{(n,h\times w,d)}$ where n denotes the number of attention heads, d is the feature dimension. Similarly $\mathcal{W}_{\text{full}}$ is projected as key and value tensors $K_{\text{full}}, V_{\text{full}} \in R^{(n,l,d)}$ by linear layers $f_K(\cdot), f_V(\cdot)$. The

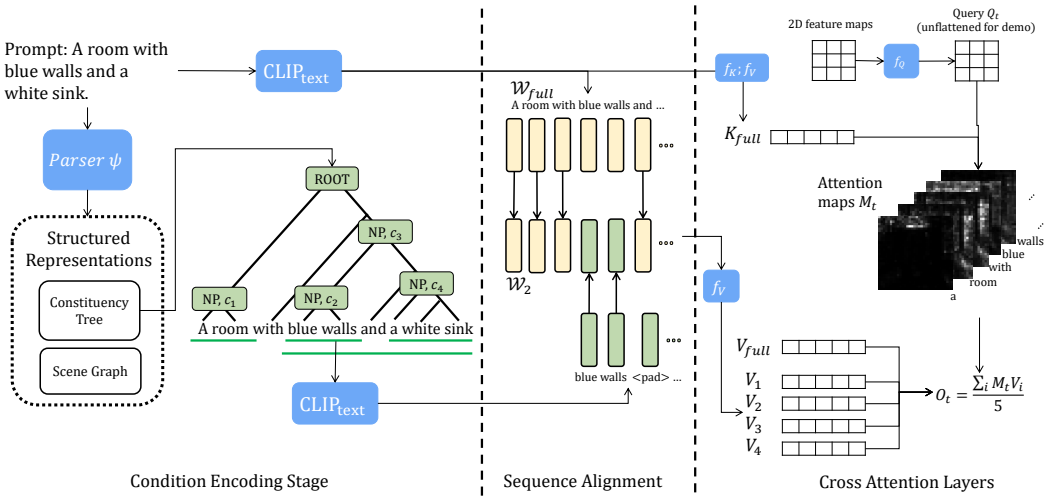

Figure 3: An illustration of our cross-attention design with structured representations. We unflatten the query and attention maps and omit the feature dimension d of all query, key, and value tensors for demonstration purposes. Note that noun phrases at multiple hierarchies are extracted and encoded through the frozen CLIP text encoder and projected to value vectors.

attention maps refer to the output from function $f_M(\cdot)$

$$M^t = f_M(Q^t, K_{full}) = \text{Softmax}(\frac{Q^t K_{\text{full}}^T}{\sqrt{d}}),\ M^t \in R^{(n,h\times w,l)}. \tag{1}$$

Cross Attention Controls Hertz et al. (2022) observe a clear separation in functionality between attention maps and token semantics in Imagen Saharia et al. (2022). The cross attention maps control the spatial layout of the image, while the value tensors map token semantics into attended regions. We observe a similar phenomenon while V_{full} also slightly contributes to the final layout. The difference potentially attributes to architecture discrepancy in hybrid cross attentions or diffusion input.

2.2 Structured Diffusion Guidance

Given the challenging prompts in Fig. 1, the attribute-object pairs are available **for free**[1] in many structured representations, such as a constituency tree or a scene graph. As is shown in Fig. 3, we can extract "blue walls" and "a white sink and door" into noun phrases (NP) and map their semantics into separate regions. The problem boils down to guiding the diffusion process of different regions within an image. We seek an implicit way of combining language structures with the cross-attention layers. Based on the observations in cross-attention controls, we interpret $M^t V_{\text{full}}$ as incorporating the token semantics into attended regions in M^t. This interpretation motivates us to multiply M^t with value tensors from different NP embeddings to achieve region-wise semantic fusion and guidance.

Specifically, given a prompt $\mathcal{P}$ and a parser $\psi(\cdot)$, we first extract a collection of concepts from all hierarchical levels as $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$. For constituency parsing, we extract all noun phrases (NP) from the tree structure. For the scene graphs, we extract object and their relation with another object as text segments. As is shown in Fig. 3, we encode each NP separately:

$$\mathbb{W} = [\mathcal{W}_{\text{full}}] + [\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_k],\ \mathcal{W}_i = \text{CLIP}_{\text{text}}(c_i). \tag{2}$$

We use $\mathcal{W}_{\text{full}}$ to obtain K_{full} and M^t as in Eq. 1, assuming that the full-prompt key tensor is able to generate layouts without missing objects. We obtain a set of value tensors from $\mathbb{W}$ and multiply each with M^t to achieve a conjunction of k NPs in $\mathcal{C}$:

$$\mathbb{V} = [f_V(\mathcal{W}_{\text{full}})] + [f_V(\mathcal{W}_1), \ldots, f_V(\mathcal{W}_k)] = [V_{\text{full}}, V_1, \ldots, V_k]. \tag{3}$$

[1]For free means that the extra computational cost introduced here is trivial in the whole diffusion process.

Algorithm 1 StructureDiffusion Guidance.

Require:
Input: Compositional prompt $\mathcal{P}$, a parsing function ξ, initialized code z^t, decoder ψ of pre-trained autoencoder, trained diffusion model ϕ.
Output: Generated image x.

Retrieve concept set $\mathcal{C} = [c_1, \ldots, c_k]$ by traversing $\xi(\mathcal{P})$;
$\mathcal{W}_{\text{full}} \leftarrow \text{CLIP}_{\text{text}}(\mathcal{P}),\ \mathcal{W}_i \leftarrow \text{CLIP}_{\text{text}}(c_i),\ i = 1, \ldots, k$;
for $t = T, T-1, \ldots, 1$ **do**
  Obtain feature maps $\mathcal{X}^t$ from previous layer output.
  for each cross attention layer **do**
    $Q^t \leftarrow f_Q(\mathcal{X}^t)$; {$\mathcal{X}^t$ denotes 2D feature maps from previous layer}
    $K_{\text{full}} \leftarrow f_K(\mathcal{W}_{\text{full}}),\ V_i \leftarrow f_V(\mathcal{W}_i), i = \text{full}, 1, \ldots, k$;
    Obtain attention map M^t from Q^t, K_{full}; {Eq. 1}
    Obtain O^t from $M^t, \mathbb{V} = \{V_i\}$; {Eq. 4}
  end for
  Feed O^t to following layers and output z^{t-1}
end for
Feed z^0 to decoder $\xi(\cdot)$ to generate x.

$$O^t = \frac{1}{(k+1)} \sum_i (M^t V_i), i = \text{full}, 1, 2, \ldots, k. \tag{4}$$

Empirically, Eq. 4 maintains the image layout while improving the attribute binding since M^t remains unchanged. However, Stable Diffusion tends to omit objects in generated images (Fig. 1), especially for concept conjunction prompts that connect two objects with the word "and". For this type of prompt, we devise a variant of our method that computes a set of attention maps $\mathbb{M}$ from $\mathcal{C}$ and multiply them to $\mathbb{V}$:

$$\mathbb{K} = \{K_i\},\ K_i = f_K(\mathcal{W}_i),\ i = \text{full}, 1, 2, \ldots, k. \tag{5}$$

$$\mathbb{M}^t = \{M_i^t\},\ M_i^t = f_M(Q^t, K_i),\ i = \text{full}, 1, 2, \ldots, k. \tag{6}$$

$$O^t = \frac{1}{(k+1)} \sum_i (M_i^t V_i), i = \text{full}, 1, 2, \ldots, k. \tag{7}$$

O^t is the output of a certain cross-attention layer and then fed into downstream layers to generate latent maps and final image x. Our approach only modifies the cross-attention layers in the diffusion model and leave other layers unchanged. Our algorithm can be summarized as 1, which requires no training or additional data.

3 Experiment

3.1 Experiment Settings

Datasets To address attribute binding and compositional generation, we propose a new benchmark, **A**ttribute **B**inding **C**ontrast set (ABC-6K). We first extract MSCOCO captions that include at least two color words modifying different subjects. For each caption, we switch the position of two color words to create a contrast caption (Gardner et al., 2020). We end up with 6.4K captions or 3.2K contrastive pairs. ABC-6K investigates the capability of counterfactual prompts, and the model should perform well in both original captions and their contrast set.

In addition to natural compositional prompts, we challenge our method with less detailed prompts that conjunct two concepts together. These prompts follow the sentence pattern of "a red apple and a yellow banana" and conjunct two objects with their attribute descriptions. We refer to this set of prompts as **C**oncept **C**onjunction 500 (CC-500). These prompts test the ability to compose two objects that rarely appear together in reality. Moreover, since the relation between two objects is not given, these prompts also validate the ability to imagine and place objects coherently within the scene. We also evaluate our method on 10K randomly sampled captions from MSCOCO (Lin et al., 2014). The purpose is to show that our method can generalize beyond attribute binding and does not introduce quality degradation for general-purpose prompts.

Benchmark	StructureDiffusion (ours) v.s.	Alignment			Fidelity		
		Win(↑)	Tie	Lose(↓)	Win(↑)	Tie	Lose(↓)
ABC-6K	Stable Diffusion	**42.2**	35.6	22.2	**48.3**	39.1	12.6
CC-500	Stable Diffusion	**31.8**	27.7	38.9	**37.8**	30.6	31.6
	Composable Diffusion	**46.5**	30.1	22.8	**61.4**	19.8	18.8

Table 1: Percentage of generated images of StructureDiffusion that are better than (win), tied with, or worse than (lose) the compared model in terms of text-image alignment and image fidelity. Note that Composable Diffusion cannot be applied to ABC-6K as those prompts may not contain explicit "and" word that separate concepts.

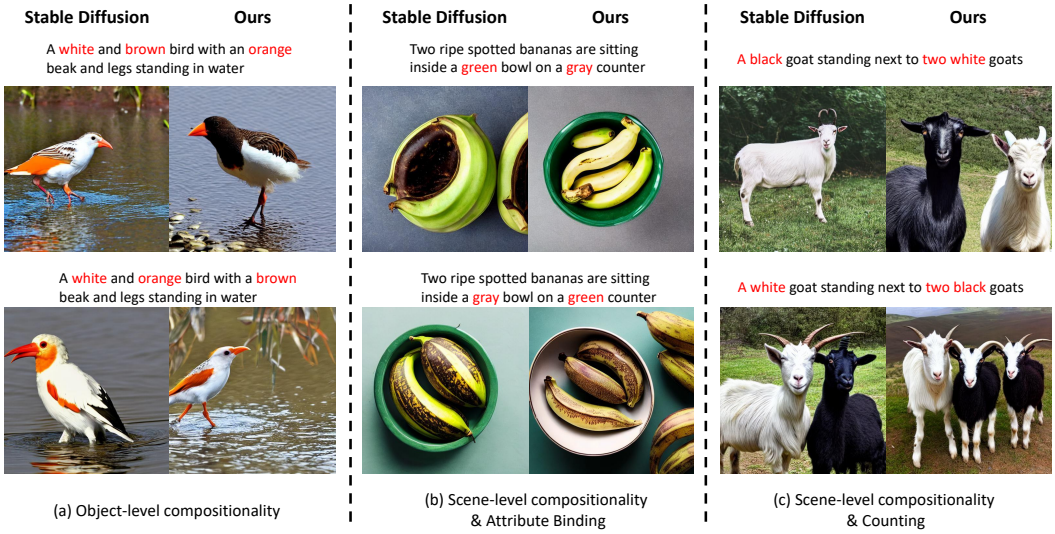

Figure 4: Qualitative results on ABC-6K. Our method improves both object-level and scene-level compositionality.

Evaluation Metrics We mainly rely on human evaluations for compositional prompts and concept conjunction since automatic evaluations are not sensitive to interchanged attributes or robust to counterfactual prompts. For instance, CLIP can be unreliable in matching the compositions between image and text inputs (Thrush et al., 2022). For each prompt, we ask human annotators to compare two generated images, one from Stable Diffusion and one from our method, and indicate which image demonstrates better image-text alignment or image fidelity. We fix the random seed for all image generation to have a fair comparison with the baseline. As for MSCOCO, we follow previous work to utilize Inception Score (IS) (Salimans et al., 2016), Frećhet Inception Distance (FID) (Heusel et al., 2017) and CLIP R-precision (R-prec.) (Park et al., 2021). IS and FID mainly measure the image bank's systematic quality and diversity, while R-prec measures image-level alignment.

Implementation Detail Throughout the experiments, we implement our method upon Stable Diffusion v1.4. For all comparisons between our method and Stable Diffusion, we fix the seed to generate the same initial gaussian map and use 50 diffusion steps with PLMS sampling (Liu et al., 2021a). We fix the guidance scale to 7.5 and equally weight the key-value matrices in cross-attention layers if not otherwise specified. We do not add hand-crafted prompts such as "a photo of" to the text input. We use the Stanza Library (Qi et al., 2020) for constituency parsing and obtain noun phrases if not otherwise specified.

3.2 Compositional Prompts

Here we show the quantitative and qualitative evaluation results on ABC-6K. We observe that our method sometimes generates very similar images to Stable Diffusion. Hence, we first generate

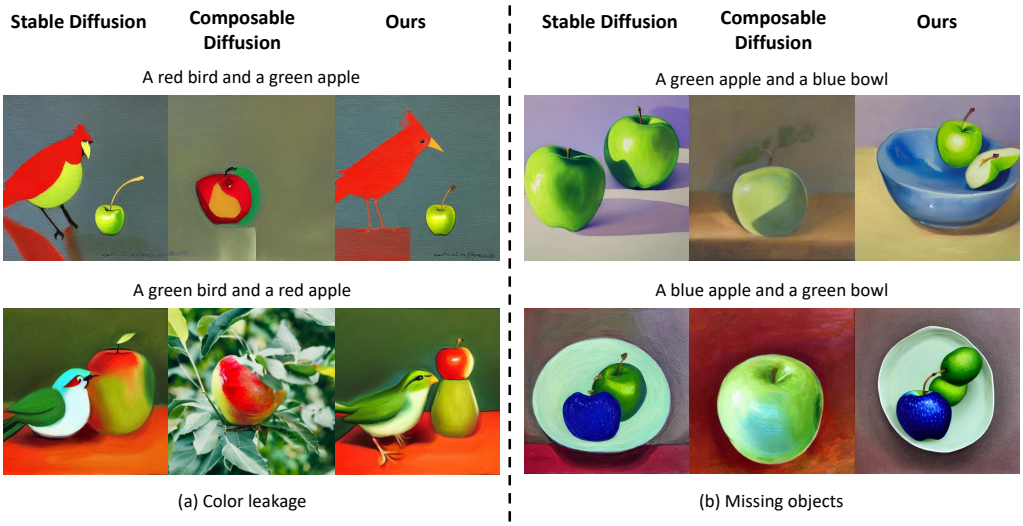

Figure 5: Qualitative results on CC-500 prompts that emphasize two aspects. (a) Color leakage: our method prevents the green color from invading the bird or apple. (b) Missing objects: our method completes the "blue bowl" and improves the quality of the "blue apple".

two images per prompt for our method and Stable Diffusion, involving around 12K image pairs to compare. Then, we filter out 20% of the most similar pairs and then randomly sampled 1500 pairs for human evaluations. As shown in Table 1, annotators indicate around a 42% chance of our method winning the comparison, 7% higher than losing the comparison. There is still a 22% of chance that our images are compatible with Stable Diffusion.

We show qualitative examples characterizing three different perspectives in Fig. 4. The first example shows instance-level compositionality, where the prompt describes different colors of the different parts of a single bird. Our method is better than Stable Diffusion to handle body parts separately and fill in the correct color semantics. The second column showcases an example of incorrect attribute binding from Stable Diffusion. In contrast, our method successfully associates "green" with the bowl and "gray" with the background counter. Our method manages to bind colors correctly even for the contrastive prompt that the colors are switched. Stable Diffusion also tends to drop out the "two goats" as shown in the third column, potentially due to a contextualized sequence of text embeddings. With the structured representations of prompts, our method encodes the spans of different goats separately and improves the number of entities in this case.

3.3 Concept Conjunction

Here we address challenging concept conjunction prompts and evaluate our method on CC-500. Apart from Stable Diffusion, we also compare to Composable Diffusion (Liu et al., 2022) implemented on top of Stable Diffusion. For Composable Diffusion, we separate the prompts into text segments by the keyword "and" and feed each span into an independent diffusion process. We generate three images per prompt and use all images for human evaluation for Stable Diffusion. We randomly sampled 600 images for comparison to Composable Diffusion.

As shown in Table 1, our method outperforms Stable Diffusion by around 5-8% in terms of both image-text alignment and image fidelity. In addition, our method surpasses Composable Diffusion by 18% in alignment and 40% in fidelity. These results align with our observations that adding structured cross-attention matrices enhances the fine-grained details of images. We also discover that images from Composable Diffusion tend to oversaturate, resulting in unnatural visual textures. This could attribute to the summation of score estimations from multiple branches. In contrast, our method only modifies cross-attention layers without introducing oversaturation issues and inefficient parallel diffusion processes. The above analysis can be further illustrated in Fig. 5. Our approach revises the incorrect shapes (the first column) and prevents color bleeding (the second column). In addition, we also observe strengthened details such as the "blue apple" in the third column.

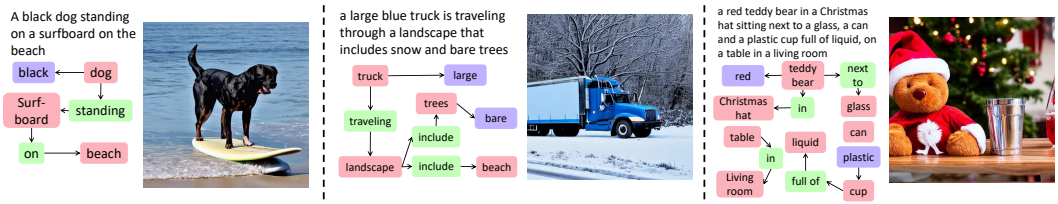

Figure 6: Qualitative results of using scene graph parser to generate structured representations.

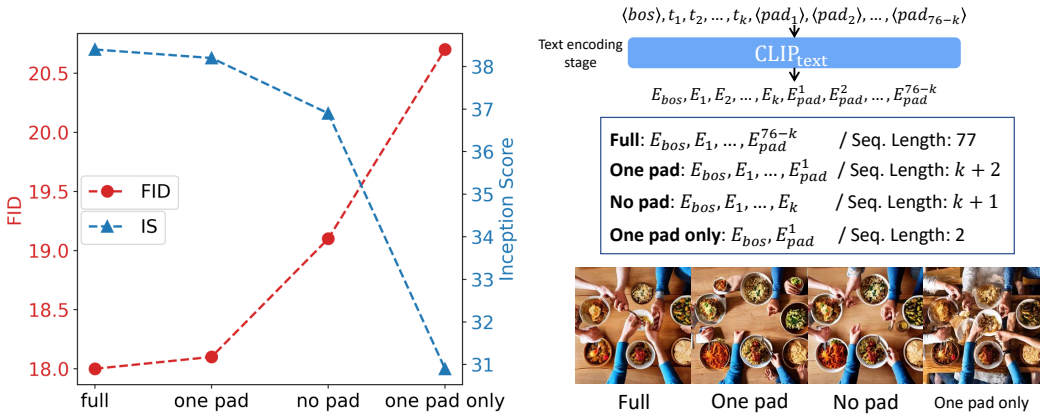

Figure 7: Ablation study on the text sequence embeddings. We find that the padding embeddings are fully contextualized, representing the prompt's high-level semantics. However, not all padding tokens are necessary to maintain a high-fidelity output from Stable Diffusion.

3.4 Other Prompts

We show that StructureDiffusion does not harm the overall image quality and diversity on MSCOCO. We follow the standard evaluation process of T2I and generate 10,000 images from randomly sampled MSCOCO captions. Stable Diffusion obtains 39.9 IS, 18.0 FID and 72.2 R-Precision. Our method achieves 40.9 IS, 17.9 FID and 72.3 R-Precision. Our StructureDiffusion maintains the image fidelity and diversity with a slightly better or comparable IS/FID/R-Prec. scores. Compared to Stable Diffusion, our method enhances fine-grained details and maintains the low-level image layout and structures.

3.5 Scene Graph Input

We show that our method is not limited to constituency parsing but can also be extended to scene graphs. As shown in Fig. 6, we first adopt the scene graph parser (Wu et al., 2019) and obtain a graph like the ones next to each image from the input prompt. The parser returns basic entities and their relations in between. We extract text spans of basic entities with their attributes attached and text spans that include two related entities. Like the constituency tree, we encode these spans separately and re-align each with the entire prompt encoding sequence.

4 Ablation Study

4.1 Contextualized Text Embeddings

One limitation brought by our StructureDiffusion is that the cross-attention computation costs increase by the number of noun phrases. Yet we noticed that most of the attention maps are computed from padding embeddings, as Stable Diffusion adopts CLIP text encoders and automatically pads the sequence to 77 tokens. We conjecture that not all padding tokens are necessary for generating high-quality images. As is shown in Fig. 7, we study four different patterns of token embeddings.

We discover that leaving the nearest padding embeddings maintains a similar IS / FID score as the full sequence. Further removing this padding embedding results in apparent degradation. While only using the nearest padding embedding results in the worst image quality, we find that the high-level image layout and semantics are preserved (see bottom right of Fig. 7). This phenomenon indicates that the padding embeddings are fully contextualized with the full prompt semantics. This also justifies our re-alignment operation that preserves padding embeddings of the main sequence $\mathcal{W}_{\text{full}}$.

5 Related Work

Text-to-Image Synthesis The diffusion model is an emerging type of model that generate high-quality images with a much more stable training process (Song & Ermon, 2019; Ho et al., 2020a). Rombach et al. (2022) proposes to encode an image with an autoencoder and then leverage a diffusion model to generate continuous feature maps in the latent space. Stable Diffusion Rombach et al. (2022) adopts similar architecture but is trained on large-scale image-text datasets with fixed CLIP text encoder. Imagen (Saharia et al., 2022) addresses the importance of language understanding by using a frozen T5 encoder (Raffel et al., 2020), a dedicated large language model. We mainly focus on diffusion models and conduct our experiments on Stable Diffusion (Rombach et al., 2022), the SOTA open-sourced T2I model.

Compositional Generation The compositional or controllable generation has been an essential direction for T2I models to understand and disentangle basic concepts in the generation process. As text inputs are relatively weak conditions, previous work leverage layout or scene graph to enhance compositionality (Johnson et al., 2018; Hong et al., 2018; Yang et al., 2022; Gafni et al., 2022). More recently, Liu et al. (2022) proposes an approach where the concept conjunctions are achieved by adding estimated scores from a parallel set of diffusion processes. In contrast, our method can be directly merged into the cross-attention layers with much less computational overhead.

Diffusion Guidance Ho & Salimans (2022) develops classifier-free guidance where a single diffusion model is jointly trained under conditional and unconditional inputs. Most large-scale SOTA models, including autoregressive ones, adopt this technique for flexible and improved conditional synthesis results (Rombach et al., 2022; Ramesh et al., 2022; Gafni et al., 2022; Yu et al., 2022; Saharia et al., 2022). Hertz et al. (2022) discovers unique properties of cross attention maps on Imagen (Saharia et al., 2022) and achieves structure-preserving image editing by manipulating these maps. We observe similar properties in Stable Diffusion (Rombach et al., 2022) but propose a different algorithm for fine-grained, compositional text-to-image generation.

6 Conclusion

In this work, we propose a training-free method for compositional text-to-image generation. First, we observe that existing large-scale T2I diffusion models can still struggle in compositional image synthesis. We address this challenge by explicitly focusing on binding objects with the correct attributes. Second, we propose structured diffusion guidance incorporating language structures into the cross-attention layers. We propose two simple techniques to align the structured encoding with the attention maps. We implement our method on a state-of-the-art diffusion-based model and conduct experiments on a wide range of prompts. Using our structured guidance, we demonstrate that attributes can be bound more accurately while maintaining the overall image quality and diversity. In addition, we justify our approach by conducting an in-depth analysis of the frozen language encoder and attention maps. As we introduce an implicit way to improve image compositionality, future directions may explore explicit approaches to generate plausible image layouts without missing components. Developing more reliable sample-level automatic evaluation metrics for image compositionality is also essential for large-scale evaluations and model comparisons. We hope that our simple yet effective approach accelerates the development of more interpretable and principled methods for diffusion-based text-to-image models.

REPRODUCIBILITY STATEMENT

We release our core codebase containing the methodology implementation, settings, benchmarks containing compositional prompts under supplementary materials.

ETHICAL STATEMENT

As for the data collection and verification, we use the Amazon Mechanical Turk platform and form the comparison task as batches of HITs. We select workers from English-speaking countries, including the US, CA, UK, AU, and NZ, since the task require understanding the English input prompt. Each HIT takes around 15-30 seconds on average to accomplish, and we pay each submitted HIT with 0.15 US dollars, resulting in an hourly payment of 18 US dollars.

REFERENCES

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W.Taylor. Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction. In *ICCV*, 2019.

Tsu-Jui Fu, Xin Eric Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. SSCR: Iterative Language-Based Image Editing via Self-Supervised Counterfactual Reasoning. In *EMNLP*, 2020.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. *Findings of Empirical Methods in Natural Language Processing*, 2020.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022a.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022b.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020a.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020b.

Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7986–7994, 2018.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1219–1228, 2018.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2021a.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022.

Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021b.

Chao Lou, Wenjuan Han, Yuhuan Lin, and Zilong Zheng. Unsupervised vision-language parsing: Seamlessly bridging visual scene graphs with language structures via dependency relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15607–15616, June 2022.

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. *ICML*, 2021.

Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

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*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pp. 70–80, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16515–16525, 2022.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Bo Wan, Wenjuan Han, Zilong Zheng, and Tinne Tuytelaars. Unsupervised vision-language grammar induction with shared structure modeling. In *International Conference on Learning Representations*, 2021.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6609–6618, 2019.

Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7764–7773, 2022.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021.

Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *European Conference on Computer Vision*, pp. 211–229. Springer, 2020.

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17907–17917, 2022.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5802–5810, 2019.

A Related Work

Text-to-Image Synthesis There are mainly three types of models for text-to-image synthesis: GAN-based (Tao et al., 2022; Zhu et al., 2019; Li et al., 2019; Fu et al., 2020; El-Nouby et al., 2019), autoregressive (Gu et al., 2022b; Lee et al., 2022; Ding et al., 2022) and diffusion models (Liu et al., 2021b; Nichol et al., 2021; Ruiz et al., 2022). Zhang et al. (2021) proposes XMC-GAN, a one-stage GAN that employs multiple contrastive losses between image-image, image-text, and region-token pairs. More recently, LAFITE (Zhou et al., 2022) enables language-free training by constructing pseudo image-text feature pairs using CLIP (Radford et al., 2021). As for autoregressive models, DALL-E adopts VQ-VAE to quantize image patches into tokens and then uses a transformer to generate discrete tokens sequentially (Ramesh et al., 2021). Parti (Yu et al., 2022) and Make-A-Scene (Gafni et al., 2022) both leverage classifier-free guidance to improve controllability. As for diffusion models, Gu et al. (2022a) concatenates VQ-VAE with the diffusion model and shows that the diffusion process can operate in discrete latent space. DALL-E 2 adopts the CLIP text encoder so that the diffusion process inverts the textual features into images (Ramesh et al., 2022).

Structured Representations for Vision and Language Inferring shared structures across language and vision has been a long-term pursuit in unifying these modalities (Schuster et al., 2015; Johnson et al., 2018; Zhong et al., 2020; Lou et al., 2022). Wu et al. (2019) utilizes the structure from semantic parsing in a visual-semantic embedding framework to facilitate embedding learning. Wan et al. (2021) proposes a new task in which the goal is to learn a joint structure between semantic parsing and image regions. To the best of our knowledge, our work is the first attempt in T2I to incorporate language structures into the image synthesizing process.

Diffusion Guidance To convert an unconditional diffusion model into a class-conditional one, Dhariwal & Nichol (2021) input the noisy image from each step into a classifier and calculate the classification loss. The loss can be back-propagated to the image space to provide a gradient that marginalizes the score estimation from the log of conditional probability. Similarly, in the T2I subdomain, Liu et al. (2021b) and Nichol et al. (2021) apply a noisy CLIP model to measure the cosine similarity between text prompts and noisy images.

B Ablation Study

B.1 A case study of attribute binding

Here, we present a case study to show evidence of two root causes of incorrect attribute binding. The first one is the contextualized token embeddings due to causal attention masks. As is shown on the left side of Fig. 8, we first encode two different prompts with a shared component, e.g. "a red apple" as the naive one and "a green bag and a red apple". Using the encoding sequence of the naive prompt, we are able to get an image of red apple only. It is reasonable to assume that the yellow green regions are natural results of learning from authentic apple images. Then, we replace the tokens of the naive prompt with embeddings of the same token from the more complicated prompt. We use the same gaussian noise as initialization and generate an unnatural image with a solid green region (in the yellow bounding box). This result proves that the token "red" is contaminated with the semantics of "green" before it and explains some images with color leakage problems (e.g., Fig. 1).

The second reason attributes to inaccurate attention maps. On the right side of Fig. 8, we visualize five cross attention maps (averaged across attention heads) from both downsample and upsampling blocks. The attention maps show the salient regions corresponding to the token "bird". These maps

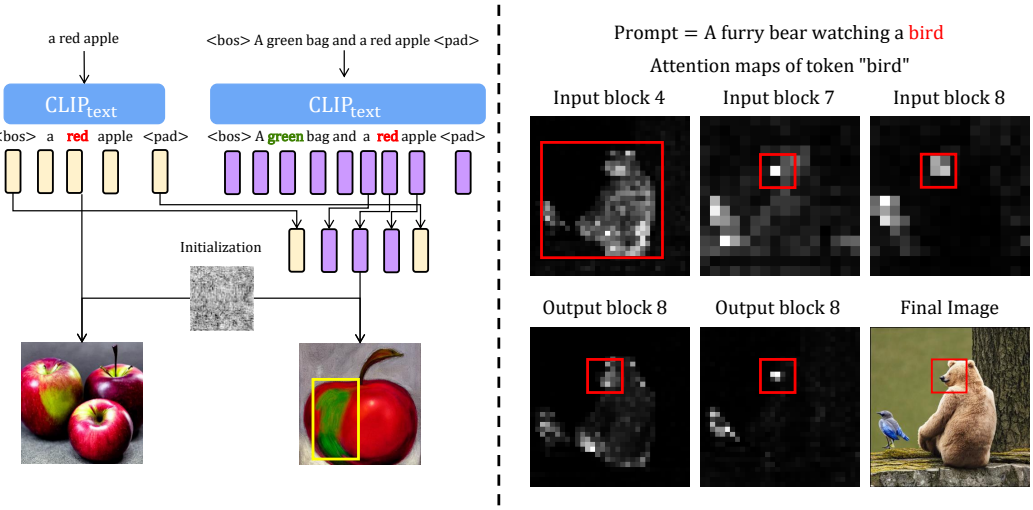

Figure 8: Examples showing the potential root causes of incorrect attribute binding. **Left**: The large green regions in the second image prove that the hidden state's output of token "red" is contextualized with token "green" before it. **Right**: Visualization of attention maps showing that the semantics from the token "bird" is mistakenly attended to the mouth region of the bear. The final image shows the unnatural beak-like shape of the bear.

demonstrate highlighted regions in the bottom left corner where the bird is located in the final image. Despite the interpretable structures, the maps also show saliency around the mouth region of the bear across all five layers. Thus, the inaccurate attention maps lead to a beak-like mouth of the bear in the final image.

C Limitations & Future Work

There are several limitations of our work. First of all, our method depends on an external parsing function that may not be perfect. We adopt the commonly used Stanza Library Qi et al. (2020) for constituency parsing. The parsing function can be replaced with a more advanced learning-based method for improvement. Secondly, our method mainly focus on compositional T2I neglecting any style descriptions. The parsing mechanism may categorize a style descripton, e.g. "in Van Gogh style" as a separate noun phrase that cannot be grounded in the image space. Future work may focus on devising explicit methods to associate attributes to objects using spatial information as input. For example, how to make text-to-image synthesis model interprets coordinate information with limited fine-tuning or prompt tuning steps would be an appealing direction.

D Additional Results

Figure 9: Qualitative results on CC-500