Improving Compositional Attribute Binding in Text-to-Image Generative Models via Enhanced Text Embeddings

Anonymous ACL submission

Abstract

Text-to-image diffusion-based generative models have the stunning ability to generate photo-realistic images and achieve state-of-the-art low FID scores on challenging image generation benchmarks. However, one of the primary failure modes of these text-to-image generative models is in composing attributes, objects, and their associated relationships accurately into an image. In our paper, we investigate compositional attribute binding failures, where the model fails to correctly associate descriptive attributes (such as color, shape, or texture) with the corresponding objects in the generated images, and highlight that imperfect text conditioning with CLIP text-encoder is one of the primary reasons behind the inability of these models to generate high-fidelity compositional scenes. In particular, we show that (i) there exists an optimal text-embedding space that can generate highly coherent compositional scenes showing that the output space of the CLIP text-encoder is sub-optimal, and (ii) the final token embeddings in CLIP are erroneous as they often include attention contributions from unrelated tokens in compositional prompts. Our main finding shows that significant compositional improvements can be achieved (without harming the model's FID score) by fine-tuning *only* a simple and parameter-efficient linear projection on CLIP's representation space in Stable-Diffusion variants using a small set of compositional image-text pairs.

1 Introduction

Text-to-image diffusion-based generative models (Rombach et al., 2021; Podell et al., 2023; Ramesh et al., 2021; Saharia et al., 2022) have achieved photo-realistic image generation capabilities on user-defined text prompts. However, recent studies (Huang et al., 2023) reveal that text-to-image models struggle with maintaining high fidelity when handling simple compositional prompts, such as those consisting of attributes, objects, and their associated relations (e.g., "*a red book and a yellow vase*"). This hinders the use of these generative models in various creative scenarios where the end-user wants to generate scenes that accurately reflect the composition and relationships specified in the prompt.

Existing approaches (Chefer et al., 2023; Feng et al., 2023; Agarwal et al., 2023; Wang et al., 2023) explore various strategies to enhance compositionality in text-to-image models. These methods primarily focus on modifying cross-attention maps by utilizing bounding box annotations and performing optimizations in the latent space during inference. Recent advancements, such as fine-tuning the UNet (Huang et al., 2023), have also demonstrated improvements in compositionality. However, the *core reasons* behind compositionality failures remain poorly understood. Gaining insights into these root causes is crucial for developing more effective approaches to augment these models with enhanced compositional capabilities.

In our paper, we investigate the potential causes of compositional attribute binding failures in text-to-image generative models, where the model fails to correctly associate descriptive attributes (such as color, shape, or texture) with the corresponding objects in the generated images. We identify two key sources of error: (i) *Erroneous attention contributions in CLIP output token embeddings*: We observe that output token embeddings in CLIP have significant attention contributions from irrelevant tokens, thereby introducing errors in generation. To explore this, we compare the internal attention contributions in CLIP for compositional prompts with the T5 text encoder, known for its stronger compositionality. Quantitative analysis shows that T5 exhibits fewer erroneous attention contributions than CLIP, indicating a potential reason for its superior compositionality. (ii) *Sub-optimality of CLIP output space for compositional prompts*: We find

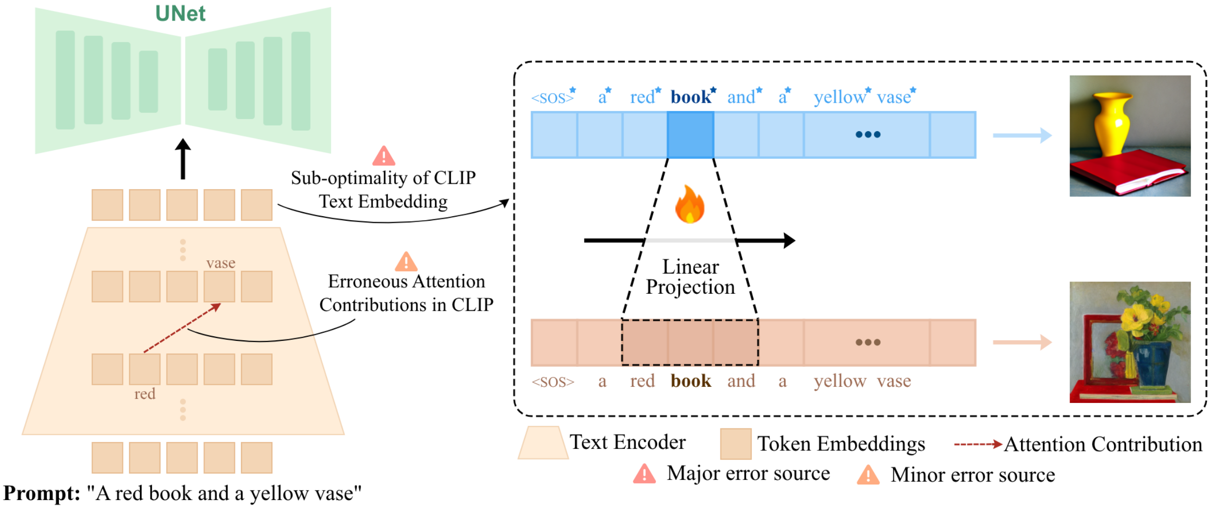

UNet
Sub-optimality of CLIP Text Embedding
Erroneous Attention Contributions in CLIP
vase
red
<sos>* a* red* book* and* a* yellow* vase*
•••
Linear Projection
•••
<sos> a red book and a yellow vase
Text Encoder
Token Embeddings
Attention Contribution
Major error source
Minor error source
Prompt: "A red book and a yellow vase"

Figure 1: Overview of our analysis and proposed methods. The figure identifies two sources of errors in Stable Diffusion's inability to generate compositional prompts: (i) erroneous attention contribution in CLIP (minor) and (ii) sub-optimal CLIP text embedding (major). We propose a window-based linear projection (WiCLP), applying linear projection to a token's surrounding window to enhance embeddings.

out that there exists an alternative text-embedding space capable of generating highly coherent images from compositional prompts. This indicates that the current CLIP output space is inherently sub-optimal. Specifically, we observe that optimizing the text embeddings produced by CLIP, while keeping the Stable Diffusion UNet frozen, converges to a more effective embedding space, enabling better compositional image generation. These findings highlight that refining the output space of the CLIP text encoder could play a critical role in enhancing compositionality.

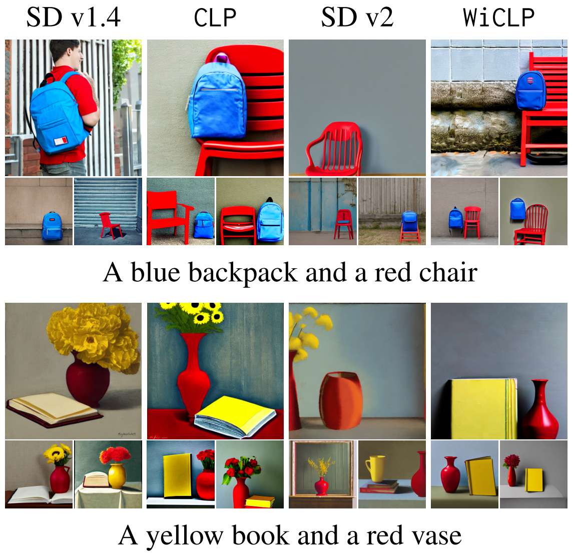
Building on our observations about the deficiencies of CLIP and identifying its text-embedding space as *a core issue* in compositional attribute binding, we explore augmenting diffusion models with a lightweight module to enhance the text-encoder's output and improve compositionality. Remarkably, a simple linear projection achieves significant improvements, comparable or superior to full fine-tuning of CLIP or training more complex networks on top of it. We demonstrate that this linear projection effectively aligns the CLIP text-encoder's output with a more optimal embedding space (see Figure 1), leading to significantly stronger compositional performances.

In particular, we introduce Window-based Compositional Linear Projection (WiCLP), a *lightweight* fine-tuning method that significantly improves the model's performance on compositional prompts (Figure 2), achieving results that are comparable to or surpassing existing methods. Additionally, WiCLP preserves the model's overall performance, maintaining high fidelity on clean prompts as evidenced by a low FID score, while offering a solution that is both *parameter efficient* and *speed efficient*. This ensures robust compositional capabilities without compromising the model's general effectiveness.

In summary, our contributions are as follows:

- We perform an in-depth analysis of the reasons behind compositionality failures in text-to-image generative models, with a particular focus on investigating the attribute binding aspect of compositionality. We highlight two key reasons contributing to these failures.

- Building on our observations, we propose WiCLP as an enhancement for Stable Diffusion (SD) v-1.4, SD v-2, and SDXL. This method significantly improves the models' compositional attribute binding, while preserving their clean accuracy on standard prompts. We observe improvements of $16.18\%, 15.15\%$, and 9.51% on SD v1.4, $14.35\%, 11.14\%$, and 6% on SD v2, and $20.31\%, 13.4\%$, and 5% on SDXL in VQA scores (Huang et al., 2023) across color, texture, and shape datasets, respectively. Our method outperforms or matches existing baselines in VQA scores, while achieving a superior FID score on clean prompts. It requires *fewer parameters for optimization* and enables *faster inference*, making it both efficient and effective.



A blue backpack and a red chair

A yellow book and a red vase

Figure 2: Qualitative comparison between the baseline and our projection methods (CLP and WiCLP). Incorporating CLP and WiCLP significantly improves image alignment with the given prompts.

2 Background

Compositionality in Text-to-Image Generative Models. Compositionality in text-to-image models refers to the ability of a model to accurately capture the correct compositions of objects, their corresponding attributes, and the relationships between objects described in a given prompt. Huang et al. (2023) introduced a benchmark specifically designed to evaluate compositionality in text-to-image models, highlighting the limitations of models when handling compositional prompts. The benchmark employs disentangled BLIP-Visual Question Answering (VQA) as a key metric for assessing image compositional quality. The VQA score assesses how accurately an image captures the compositional elements described in the prompt by utilizing a vision-language model. This metrics demonstrates a closer correlation with human judgment compared to metrics like CLIP-Score (Hessel et al., 2021). The authors also proposed a fine-tuning baseline to enhance compositionality in these models. Alternatively, compositionality issues can be addressed during inference by modifying cross-attention maps using hand-crafted loss functions and bounding boxes derived from a language model (Chefer et al., 2023; Feng et al., 2023; Agarwal et al., 2023; Wang et al., 2023; Nie et al., 2024; Lian et al., 2023; Liu et al., 2022a). However, Huang et al. (2023) demonstrated that data-driven fine-tuning approaches are more effective for improving compositionality in text-to-image models.

Text-to-image Diffusion Models: Training and Inference. In diffusion models, noise is added to the data following a Markov chain across multiple time-steps $t \in [0, T]$. Starting from an initial random real image $\mathbf{x}_0$ along with its caption c, $(\mathbf{x}_0, c) \sim \mathcal{D}$, the noisy image at time-step t is defined as $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{(1-\alpha_t)}\epsilon$. The denoising network denoted by $\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)$ is pre-trained to denoise the noisy image $\mathbf{x}_t$ to obtain $\mathbf{x}_{t-1}$. For better training efficiency, the noising along with the denoising operation occurs in a latent space defined by $\mathbf{z} = \mathcal{E}(\mathbf{x})$, where $\mathcal{E}$ is an encoder such as VQ-VAE (van den Oord et al., 2017). Usually, the conditional input $\mathbf{c}$ to the denoising network $\epsilon_\theta(.)$ is a text-embedding of the caption c through a text-encoder $\mathbf{c} = v_\gamma(c)$. The pre-training objective for diffusion models can be defined as follows:

$$\mathcal{L}(\theta) = \mathbb{E}_{(\mathbf{x}_0,c)\sim\mathcal{D},\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t)\|_2^2\right],$$

where θ is the set of learnable parameters in the UNet ϵ_θ. During inference, given a text-embedding $\mathbf{c}$, a random Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(0, I)$ is iteratively denoised for a fixed range of time-steps to produce the final image.

3 Sources of Compositionality Failures

This section conducts an in-depth analysis of compositional attribute binding failures in text-to-image models, focusing on the CLIP text-encoder.

3.1 Source (i) : Erroneous Attention Contributions in CLIP

In this section, we leverage attention contributions (Elhage et al., 2021; Dar et al., 2023) to analyze how the final text-embeddings of compositional prompts are obtained by the CLIP text-encoder, a widely adopted component in many text-to-image models. We then compare these attention contribution patterns with those produced by the T5 text-encoder used in DeepFloyd, a model known for its stronger compositional capabilities. Many of the compositional prompts from Huang et al. (2023) have a decomposable template of the form $\mathbf{a}_i\ \mathbf{o}_j + \mathbf{a}_j\ \mathbf{o}_j$, where $\mathbf{a}_i, \mathbf{a}_j$ are attributes (e.g., “black”, “matted”) and $\mathbf{o}_i, \mathbf{o}_j$ represent objects (e.g., “car”, “bag”).

The attention mechanism in layer ℓ of a transformer consists of four weight matrices

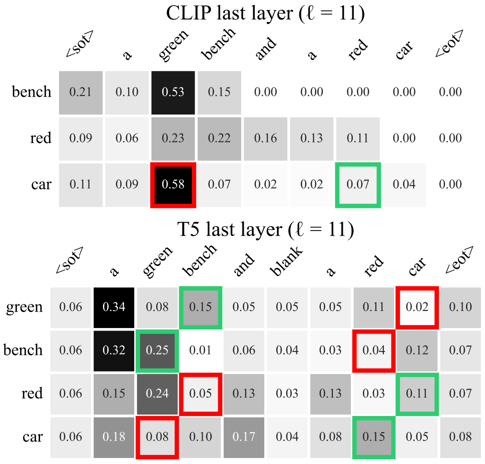


Figure 3: The heatmap illustrates unintended attention contributions in CLIP, while highlighting the more accurate performance of T5.

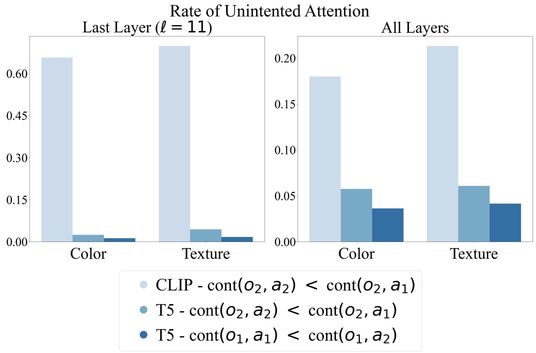


Figure 4: Quantitatively, we find CLIP to have significantly higher erroneous attention contributions averaged across 780 prompts of color dataset and 582 prompts of texture dataset.

W_q, W_v, W_k, and W_o (Vaswani et al., 2017). Each of these matrices is divided into H heads, denoted by $W_\text{q}^h, W_\text{v}^h, W_\text{k}^h \in \mathbb{R}^{d \times d_h}$, $W_\text{o}^h \in \mathbb{R}^{d_h \times d}$, where $h \in [H]$. Here, d_h denotes the dimensionality of the internal token embeddings. For simplicity, we omit ℓ, but each layer has its own attention matrices. These matrices operate on the token embeddings produced by the previous layer $(\ell - 1)$, denoted as $\bar{\mathbf{x}}_j$ for token j. We further denote the projections of $\bar{\mathbf{x}}_j$ onto the query, key, and value matrices of the h-th attention head in layer ℓ as q_j^h, k_j^h, and v_j^h, respectively. More precisely,

$$\text{q}_j^h = \bar{\mathbf{x}}_j W_\text{q}^h, \quad \text{k}_j^h = \bar{\mathbf{x}}_j W_\text{k}^h, \quad \text{v}_j^h = \bar{\mathbf{x}}_j W_\text{v}^h.$$

The *contribution* of token j to token i in layer ℓ, denoted by $\text{cont}_{i,j}$, is computed as follows:

$$\text{cont}_{i,j} = \left\| \sum_{h=1}^{H} \text{attn}_{i,j}^h \ \text{v}_j^h \ W_\text{o}^h \right\|_2$$

where $\text{attn}_{i,j}^h$ is the attention weight of token i to j in the h-th head of layer ℓ. Specifically,

$$\text{attn}_{i,\cdot}^h = \textsc{Softmax}\left(\left\{ \frac{\langle \text{q}_i^h, \text{k}_j^h \rangle}{\sqrt{d_h}} \right\}_{j=1}^{n} \right).$$

Notably, $\text{cont}_{i,j}$ is a significant metric that quantifies the *contribution* of a token j to the norm of a token i at layer ℓ. We employ this metric to identify layers in which important tokens highly attend to *unintended* tokens, or lowly attend to *intended* ones. We refer to Appendix B.1 for more details on attention contribution.

Key Finding: T5 has less erroneous attention contributions than CLIP. We refer to Figure 3 that visualizes attention contribution of both T5 and CLIP text-encoder in the last layer $(\ell = 11)$ for the prompt “a green bench and a red car". Ideally, the attention mechanism should guide the token “car” to focus more on “red” than “green”, but in the last layer of the CLIP text-encoder, “car” significantly attends to “green”. In contrast, T5 shows a more consistent attention pattern, with “red” contributing more to the token “car” and “green” contributing more to the token “bench”.

We further conduct a comprehensive analysis focusing on specific types of compositional prompts from the T2I-CompBench dataset (Huang et al., 2023). This includes 780 prompts from the color category and 582 prompts from the texture category of this dataset, each following the structured format: “$\mathbf{a}_1$ $\mathbf{o}_1$ and $\mathbf{a}_2$ $\mathbf{o}_2$”. For each prompt, we obtain attention contributions in all layers and count the number of layers where *unintended attention contributions* occur. In the CLIP text-encoder, unintended attention occurs when $\mathbf{o}_2$ attends more to $\mathbf{a}_1$ than $\mathbf{a}_2$. For T5, it occurs when $\mathbf{o}_2$ attends more to $\mathbf{a}_1$ than $\mathbf{a}_2$, or $\mathbf{o}_1$ attends more to $\mathbf{a}_2$ than $\mathbf{a}_1$. Figure 4 provides a quantitative comparison of unintended attention across various prompts between the CLIP text-encoder and T5. The T5 model demonstrates superior performance on our metric compared to the CLIP text-encoder, reinforcing the hypothesis that erroneous attention mechanisms in CLIP may contribute to its weaker compositional performance in text-to-image models. Additional details can be found in Appendix B.4. Further experiments with other text-encoders are also reported in Appendix B.3.

To address the attention shortcomings of the CLIP text-encoder, we explored zero-shot reweighting of attention maps in CLIP to reduce unintended

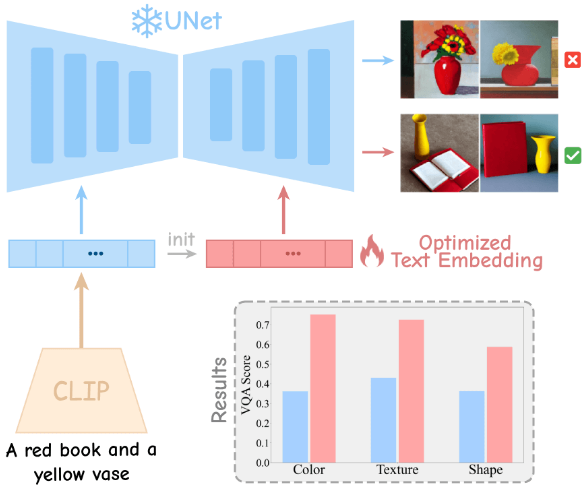


Figure 5: Sub-optimality of CLIP Text-Encoder for Compositional Prompts. Optimizing a learnable vector to represent an improved text embedding, while keeping the UNet frozen, enables the generation of more compositionally accurate images.

attentions while enhancing meaningful ones. While this improved baseline performance, it fell short of our primary method discussed in the following sections. See Appendix B.2 for more details.

3.2 Source (ii) : Sub-optimality of CLIP Text-Encoder for Compositional Prompts

In this section, we investigate whether the UNet is capable of generating better compositional scenes if provided with alternative (*improved*) text embeddings, rather than relying on the output of the CLIP text-encoder. For a given input prompt p with a specific composition (e.g., *"a red book and a yellow table"*), we utilize our dataset (described in Section 5) to obtain $\mathcal{D}_p$, a set of high-quality compositional images corresponding to prompt p. Next, we extract the text embedding $\mathbf{c}$ from the CLIP text-encoder for prompt p. Using this embedding as the initialization, we create a learnable vector $\mathbf{c}^*$ of the same dimensionality. Keeping all other components (such as the UNet) frozen, we optimize this learnable vector as follows:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \mathbb{E}_{x_0\sim\mathcal{D}_p,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t,\mathbf{c},t)\|_2^2\right].$$

We then use the optimized text embedding $\mathbf{c}^*$ to generate images with the UNet ϵ_θ. Figure 5 illustrates the complete pipeline.

Key Results. Utilizing Stable Diffusion v1.4, we optimize optimize $\mathbf{c}^*$ for all compositional prompts across the color, texture, and shape categories in the T2I-CompBench dataset. By generating samples with $\mathbf{c}^*$ and comparing them to those generated using $\mathbf{c}$ (the output of the CLIP text-encoder), we observe a significant improvement in the VQA scores. As shown in Figure 5, CLIP text embeddings yield VQA scores of 0.3615 for color, 0.4306 for texture, and 0.3619 for shape. On the other hand, the optimized embeddings achieve scores of 0.7513 for color, 0.7254 for texture, and 0.5873 for shape.

These results indicate that CLIP text-encoder does not output the proper text-embedding suitable for generating compositional scenes. However, the existence of an optimized embedding space demonstrates that the UNet can generate coherent compositional outputs when provided with appropriately improved embeddings. This finding motivates the idea of improving the CLIP output space to mitigate compositionality issues in text-to-image diffusion models. For additional configurations, including results from optimizing a subset of tokens to improve compositionality, refer to Appendix A.

4 Projection Layer for Enhancing Compositionality in the CLIP Text Embedding Space

Building on our previous findings, we focus on improving the text embedding space utilized in text-to-image generative models. Specifically, we propose learning a projection layer over the CLIP output embedding space to transform its sub-optimal representation into an enhanced space better suited for compositionality. In the following sections, we introduce two methods, CLP and WiCLP, which implement linear projections of the CLIP output embedding space to achieve this enhancement.

4.1 CLP: Token-wise Compositional Linear Projection

Given the text-embedding $\mathbf{c} \in \mathbb{R}^{n\times d}$ as the output of the text-encoder for prompt c, i.e., $\mathbf{c} = v_\gamma(c)$, we train a linear projection $\texttt{CLP}_{W,b} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$. This projection includes a matrix $W \in \mathbb{R}^{d\times d}$ and a bias term $b \in \mathbb{R}^d$, which are applied token-wise to the output text-embeddings of the text-encoder. More formally, for $\mathbf{c} \in \mathbb{R}^{n\times d}$ including text-embeddings of n tokens $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n \in \mathbb{R}^d$, $\texttt{CLP}_{W,b}(\mathbf{c})$ is obtained by stacking projected embeddings $\mathbf{c}'_1, \mathbf{c}'_2, \cdots, \mathbf{c}'_n$ where $\mathbf{c}'_i = W^T\mathbf{c}_i + b$.

Finally, we solve the following optimization problem on a dataset $\mathcal{D}$ including image-caption

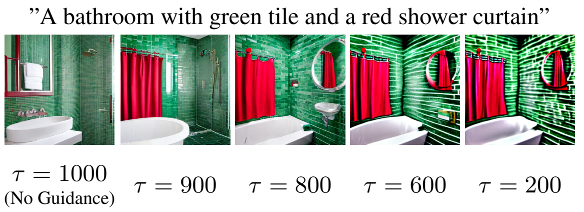


Figure 6: Qualitative results showing the impact of Switch-Off with varying thresholds τ.

pairs of high-quality compositional images:

$$W^*, b^* = \arg\min_{W,b} \mathbb{E}_{(x_0,c)\sim\mathcal{D},\epsilon,t}\left[\Phi_{\texttt{CLP}}\right]$$

$$\Phi_{\texttt{CLP}} = \left\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, \texttt{CLP}_{W,b}\left(\mathbf{c}\right), t\right)\right\|_2^2$$

We then apply $\texttt{CLP}_{W^*,b^*}$ on CLIP text-encoder to obtain improved embeddings.

4.2 `WiCLP`: Window-based Compositional Linear Projection

In this section, we propose a more advanced linear projection scheme where the new embedding of a token is derived by applying a linear projection on that token in conjunction with a set of its adjacent tokens, i.e., tokens within a specified window. This method not only leverages the benefits of `CLP` but also incorporates the contextual information from neighboring tokens, potentially leading to more precise text-embeddings.

More formally, we train a mapping $\texttt{WiCLP}_{W,b} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ including a parameter s (indicating window length), matrix $W \in \mathbb{R}^{(2s+1)d\times d}$, and a bias term $b \in \mathbb{R}^d$. For text-embeddings $\mathbf{c} \in \mathbb{R}^{n\times d}$ consisting of n token embeddings of $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n \in \mathbb{R}^d$, we obtain $\texttt{WiCLP}_{W,b}$ by stacking projected embeddings $\mathbf{c}'_1, \mathbf{c}'_2, \cdots, \mathbf{c}'_n$ where

$$\mathbf{c}'_i = W^T \,\textsc{Concatenation}\left(\left(\mathbf{c}_j\right)_{j=i-s}^{i+s}\right) + b$$

Similarly, we solve the following optimization problem to train the projection:

$$W^*, b^* = \arg\min_{W,b} \mathbb{E}_{(x_0,c)\sim\mathcal{D},\epsilon,t}\left[\Phi_{\texttt{WiCLP}}\right]$$

$$\Phi_{\texttt{WiCLP}} = \left\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, \texttt{WiCLP}_{W,b}\left(\mathbf{c}\right), t\right)\right\|_2^2$$

We observe that `WiCLP` improves over `CLP` (special case of `WiCLP` with $s = 0$) by incorporating adjacent tokens in addition to the actual token. This approach enhances embeddings by reinforcing the contributions of relevant adjacent tokens. For discussion on choosing the window length (s) in `WiCLP`, see Appendix C.6.

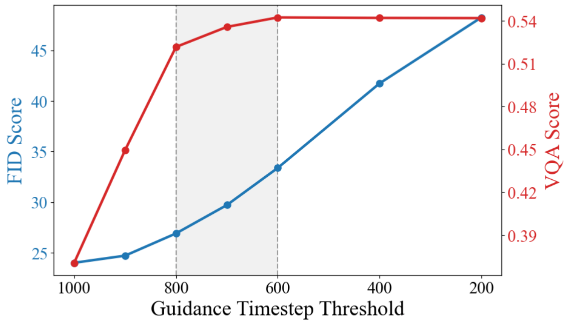


Figure 7: Trade-off between VQA and FID scores with Switch-Off at different thresholds.

4.3 Switch-Off: Trade-off between Compositionality and Clean Accuracy

Fine-tuning models or adding modules to a base model often results in a degradation of image quality and an increase in the Fréchet Inception Distance (FID) score. To balance the trade-off between improved compositionality and the quality of generated images for clean prompts – an important issue in existing work – inspired by Hertz et al. (2022), we adopt Switch-Off, where we apply the linear projection only during the initial steps of inference. Specifically, given a time-step threshold τ, for $t \geq \tau$, we use $\texttt{WiCLP}_{W^*,b^*}(\mathbf{c})$, while for $t < \tau$, we use the unchanged embedding $\mathbf{c}$ as the input to the cross-attention layers.

Figure 7 illustrates the trade-off between VQA score and FID on a randomly sampled subset of MS-COCO (Lin et al., 2014) for different choices of τ. As shown, even a large value of τ suffices for obtaining high-quality compositional scenes as the composition of final generated image is primarily formed at early steps. Thus, choosing a large τ preserves the model's improved compositionality while maintaining its clean accuracy. Setting $\tau = 800$ offers a competitive VQA score compared to the model where projection is applied at all time steps, and achieves a competitive FID similar to that of the clean model. Figure 6 depicts a few images generated using different choices of τ. We refer to Appendix C.5 for more visualizations.

5 Experiments

Existing Baselines. We evaluate the performance of four methods alongside standard models SD v1.4, SD v2, and SDXL (Podell et al., 2023). These include Composable Diffusion (Liu et al., 2022b), which addresses concept conjunction and negation in pretrained diffusion models; Structured Dif-

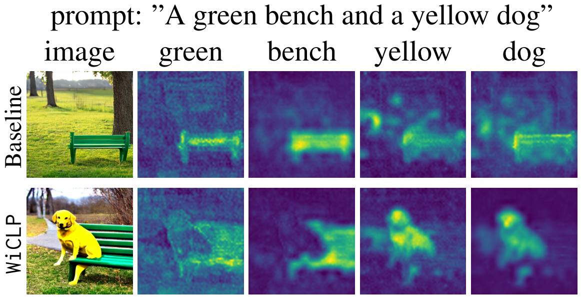


Figure 8: Applying CLP results in more accurate cross-attention maps.

	Model	Color	Texture	Shape
SD v1.4	Baseline	0.3765	0.4156	0.3576
	CLP	0.4837	0.5312	0.4307
	WiCLP	**0.5383**	**0.5671**	**0.4527**
SD v2	Baseline	0.5065	0.4922	0.4221
	Composable	0.4063	0.3645	0.3299
	Structured	0.4990	0.4900	0.4218
	Attn-Exct	0.6400	0.5963	0.4517
	GORS	0.6414	0.6025	0.4546
	CLP	0.6075	0.5707	0.4567
	WiCLP	**0.6500**	**0.6036**	**0.4821**
SDXL	Baseline	0.5770	0.5217	0.4666
	WiCLP	0.6930	0.6007	0.4758
	WiCLP*	**0.7801**	**0.6557**	**0.5166**

Table 1: Quantitative comparison with state-of-the-art and baseline methods across different categories of the T2I-CompBench dataset

fusion (Feng et al., 2022), which focuses on attribute binding; Attn-Exct (Chefer et al., 2023), which ensures correct attention to all subjects in the prompt; and GORS (Huang et al., 2023), which fine-tunes Stable Diffusion v2 using a reward function. GORS optimizes more parameters but underperforms slightly compared to our method, while Attn-Exct requires iterative optimizations during inference, making it slower than our method, which adds only a linear projection layer.

Training Setup. All of the models are trained using the objective function of diffusion models on color, texture, and shape datasets described in Section 5. During training, we keep all major components frozen, including the U-Net, CLIP text-encoder, and VAE encoder and decoder, and only the linear projections are trained. We refer to Appendix C.1 for details on the training procedure.

Dataset Collection. We utilize the T2I-CompBench dataset (Huang et al., 2023), focusing on three key categories: color, texture, and shape, with a total of 1,000 prompts across both training and evaluation sets. T2I-CompBench is a well-established and widely recognized dataset (Esser et al., 2024). This dataset provides distinct training and evaluation splits for each category, enabling a structured approach to assessing performance. To generate high-quality images, we use three generative models: SD 1.4 (Rombach et al., 2021), DeepFloyd, and SynGen (Rassin et al., 2024), creating 100 samples per prompt with SD 1.4, 60 with DeepFloyd, and 50 with SynGen. This ensures a wide variety of generated images, leveraging each model's strengths. For each prompt, we combined all 210 samples from the three models and selected the top 30 with the highest VQA scores, ensuring the final dataset consisted of images that most accurately reflected the prompts.

Furthermore, for SDXL, we explored training WiCLP (WiCLP* in Table 1) on a higher-quality dataset generated by more recent text-to-image models, such as SDXL itself and SD3. Importantly, leveraging an appropriately curated dataset results in a substantial improvement in VQA scores, highlighting the importance of high-quality training data for compositional understanding.

5.1 Qualitative and Quantitative Evaluation

Qualitative Evaluation. Figure 2 presents images generated when applying CLP and WiCLP. When generating compositional prompts with a baseline model, objects are often missing or attributes are incorrectly applied. However, with CLP and WiCLP, objects and their corresponding attributes are more accurately generated. We refer to Appendix C.3 for more visualizations.

Figure 8 illustrates cross-attention maps for a sample prompt. In the base model, attention maps are flawed, with some tokens incorrectly attending to the wrong pixels. However, with both CLP and WiCLP, objects and attributes more accurately attend to their respective pixels. For more visualizations, see Appendix C.4.

Quantitative Evaluation. Table 1 presents the VQA scores for our methods, CLP and WiCLP, alongside the baselines discussed. VQA scores of our method and other discussed baselines are provided in Table 1. As shown, both CLP and WiCLP significantly improve upon the baselines. Both methods demonstrate substantial improvements over the baselines, with WiCLP achieving the highest VQA scores among state-of-the-art approaches that utilize the same baseline model (e.g., Stable Diffusion v2), despite its simplicity.

Notably, our methods maintain the model's general utility, introducing only a slight increase in the FID score; for example, experiments on MS-COCO prompts show that while our methods slightly increase FID compared to base models, this increase is smaller than that of other baselines—for instance, WiCLP achieves an FID score of 27.40, outperforming GORS at 30.54. Additional details on FID performance can be found in Appendix C.2.

Human Experiments. We conducted a human evaluation where participants compared images generated by SD v1.4 and SD v1.4 + WiCLP, selecting the image that best matched the given prompt. The results showed that in 34.625% of cases, evaluators chose the base model's image; in 51.875%, they preferred the WiCLP images; and in 13.50%, they rated both equally. Further details can be found in Appendix C.2.

5.2 Impact of WiCLP on Subsets of Tokens

To better understand the impact of WiCLP on token embeddings, we conducted experiments applying the trained WiCLP to specific subsets of tokens from a sample of sentences in the color category of the T2I-CompBench dataset. The results, shown in Fig. 9, compare the following token groups: nouns only; nouns and adjectives; nouns, adjectives, and the EOS (End of Sentence token) token; all sentence tokens; and all tokens outputted by CLIP (sentence tokens plus padding tokens). As can be seen, applying WiCLP only to a small number of tokens is sufficient for improving compositionality. Interestingly, applying WiCLP to the group of nouns, adjectives, and EOS achieves even higher VQA scores than applying WiCLP to all tokens. Despite these findings, we applied WiCLP to all tokens in our main work, leaving this targeted approach as an avenue for future research.

5.3 Alternatives to WiCLP

We explored various fine-tuning strategies for improving CLIP, including fine-tuning the entire CLIP, fine-tuning only the last layers of CLIP combined with WiCLP, and using WiCLP alone. Our results show that the original baseline model (SD v1.4) achieves a VQA score of 0.3765 on the color category of the dataset. Fine-tuning the entire CLIP without WiCLP improves the score to 0.5173, fine-tuning the last layers of CLIP combined with WiCLP achieves 0.5497, and WiCLP alone achieves 0.5383.

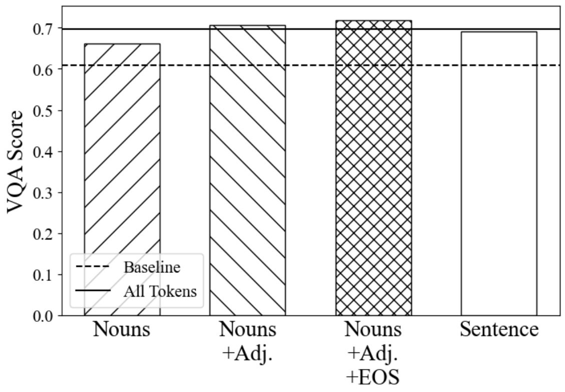


Figure 9: Effect of applying WiCLP to specific tokens. Applying WiCLP to a subset of tokens is sufficient to enhance compositionality, achieving comparable or superior performance to applying it across all tokens.

These findings highlight the effectiveness of WiCLP, which outperforms full fine-tuning of CLIP while being significantly more parameter-efficient. While fine-tuning the last layers of CLIP combined with WiCLP achieves slightly better performance than using WiCLP alone, it requires optimizing a much larger number of parameters. Given this trade-off, we prioritize WiCLP alone to minimize the number of parameters while achieving substantial compositional performance improvements. Additionally, keeping the original CLIP unchanged makes our approach more suitable for SWITCH-OFF functionality, allowing the module to be easily enabled or disabled as needed.

6 Conclusion

Our paper examines potential error sources in text-to-image models for generating images from compositional prompts. We identify two error sources: (i) A minor error source, where the token embeddings in the CLIP text-encoder have erroneous attention contributions and (ii) A major error source, where we find the output space of the CLIP text-encoder to be sub-optimally aligned to the UNet for compositional prompts. Leveraging our observations, we propose a simple and strong baseline WiCLP which involves fine-tuning a linear projection on CLIP's representation space. WiCLP though inherently simple and parameter efficient, outperforms existing methods on compositional image generation benchmarks and maintains a low FID score on a broader range of clean prompts.

Limitations

In this paper, we have conducted a comprehensive analysis of one of the primary reasons why Stable Diffusion struggles with generating compositional attribute binding prompts and proposed a lightweight, efficient method to address this challenge. While our approach demonstrates promising results, there remains substantial room for improvement in this area. Our method primarily targets the attribute binding aspect of compositionality, leaving other critical categories, such as spatial relationships (e.g., "a book to the left of a pen"), numeracy (e.g., "four books"), and others, less explored. Investigating the underlying causes of these issues is crucial for advancing the field further.

Moreover, the reliance on CLIP—particularly the CLIP score—as a metric for recognizing and evaluating compositionality poses its own limitations. CLIP, in its current form, does not perform optimally for such tasks. A promising direction for future research would be to first improve CLIP's ability to handle compositionality effectively and then adapt this enhanced version of CLIP for Stable Diffusion. This could pave the way for more robust and accurate text-to-image generation models.

References

Aishwarya Agarwal, Srikrishna Karanam, K J Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2023. A-star: Test-time attention segregation and retention for text-to-image synthesis. *Preprint*, arXiv:2306.14544.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *Preprint*, arXiv:2301.13826.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. *Preprint*, arXiv:2209.02535.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Preprint*, arXiv:2403.03206.

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

Weixi Feng, Wanrong Zhu, Tsu jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Layoutgpt: Compositional visual planning and generation with large language models. *Preprint*, arXiv:2305.15393.

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Preprint*, arXiv:2307.06350.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *ArXiv*, abs/2305.13655.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. 2022a. Compositional visual generation with composable diffusion models. *ArXiv*, abs/2206.01714.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022b. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer.

Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. 2024. Compositional text-to-image generation with dense blob representations. *Preprint*, arXiv:2405.08246.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Preprint*, arXiv:2307.01952.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2024. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Preprint*, arXiv:2306.08877.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494. Curran Associates, Inc.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *CoRR*, abs/1711.00937.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. 2023. Compositional text-to-image synthesis with attention map control of diffusion models. *Preprint*, arXiv:2305.13921.

A Optimizing the Text-embeddings of a Subset of Tokens

Given $\mathbf{c} \in \mathbb{R}^{n \times d}$, where n refers to the number of tokens and d refers to the dimensionality of the text-embedding, for the second configuration we only optimize a subset of tokens $n' \in n$. We refer to this subset of tokens as $\mathbf{c}'$. These tokens correspond to relevant parts of the prompt which govern compositionality (e.g., “red book” and “yellow table” in “A red book and an yellow table”).

$$\mathbf{c}'^{*} = \arg\min_{\mathbf{c}'} \mathbb{E}_{\epsilon,t}||\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}', t)||_2^2,$$

Figure 10 shows the results for the sample prompt "a red book and a yellow vase". We considered different subsets of tokens n': adjectives ("red" and "yellow"), nouns ("book" and "vase"), both nouns and adjectives, and all tokens in the sentence. The results indicate that optimizing even a few tokens significantly improves the VQA score. However, optimizing all tokens in the sentence yields the highest score.

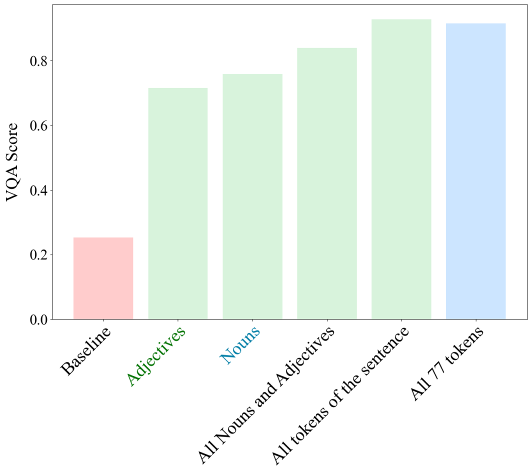


Figure 10: Comparison of VQA scores when optimizing different subsets of tokens for the sample prompt: ”A red book and a yellow vase”

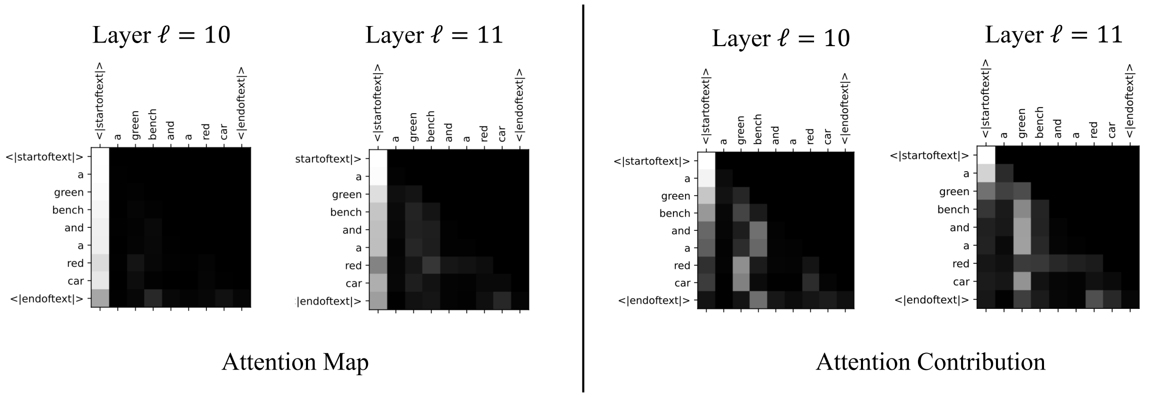
B Source (i) : Erroneous Attention Contributions

B.1 Attention Contribution

In this Section, we provide more details on our analysis to quantitatively measure tokens’ contribution to each other in a layer of attention mechanism. One natural way of doing this analysis is to utilize attention maps $\text{attn}_{i,j}^{h}$ and aggregate them over heads, however, we observe that this map couldn’t effectively show the contribution. Attention map does not consider norm of tokens in the previous layer, thus, does not provide informative knowledge on how each token is formed in the attention mechanism. In fact, as seen in Figure 11, we cannot obtain much information by looking at these maps while attention contribution clearly shows amount of norm that comes from each of the attended tokens.

B.2 Zero-shot Attention Reweighting

Inspired by attention mechanism shortcomings of CLIP text-encoder, we aim to improve compositionality of CLIP-based diffusion models by zero-shot reweighting of the attention maps. Specifically, we apply a hand-crafted zero-shot manipulation of the attention maps in certain layers of the CLIP text-encoder to effectively reduce unintended attentions while enhancing meaningful ones. This



Attention Map

Attention Contribution

Figure 11: Visualization of attention map and attention contribution for prompt “a green bench and a red car" over different layers of CLIP. Contribution provides better insight on the attention mechanism.

zero-shot reweighting is applied to the logits before the SOFTMAX layer in the last three layers of the text-encoder. More precisely, we compute a matrix $M \in \mathbb{R}^{n\times n}$ and add it to the attention logits. For each head h, the new attention values are computed and then propagated through the subsequent layers of the text encoder:

$$\text{attn}'^{h}_{i,\cdot} = \textsc{Softmax}\left(\left\{\frac{\langle \mathrm{q}_i^h, \mathrm{k}_j^h\rangle}{\sqrt{d_h}} + M_{i,j}\right\}_{j=1}^{n}\right).$$

We set the values in M by considering the ideal case where no incorrect attentions occur in the mechanism. For example, for prompt “a green bench and a red car", we ensure that the token "car" does not attend to the token "green" by assigning a sufficiently large negative value to the corresponding entry in matrix M.

To fix unintended attentions, we aim to compute a matrix M to be applied across various heads in the last few layers of CLIP, reducing the effect of wrong attention, leading to more accurate text-embeddings that are capable of generating high-quality compositional scenes. To avoid unintended attention for prompts of the form “$\mathbf{a}_1\mathbf{o}_1 + \mathbf{a}_2\mathbf{o}_2$", we add large negative values to entries $M_{\mathbf{o}_2,\mathbf{a}_1}$, $M_{\mathbf{a}_2,\mathbf{a}_1}$, and some positive value to $M_{\mathbf{o}_2,\mathbf{a}_2}$ and $M_{\mathbf{o}_1,\mathbf{a}_1}$, and small negative value to $M_{\mathbf{o}_2,\mathbf{o}_1}$. To find what values to assign to those entries, we consider a small set of prompts in color dataset (5 prompts in total) and obtain parameters for that matrix to maximize VQA score. Figure 12 shows few examples of zero-shot modification.

Applying zero-shot attention reweighting with matrix M on 780 compositional prompts of the color category of T2I-CompBench dataset, we achieved a 2.93% improvement in VQA scores.

		LLaMa3	CLIP
color	last layer	0.015	0.657
	all layers	0.081	0.187
texture	last layer	0.033	0.696
	all layers	0.066	0.213

[t]

Table 2: Unintended attention rate in LLaMa3 8B vs CLIP. LLaMa3 shows significant less unintended attentions.

	FID Score
SD v1.4	24.33
SD v1.4 + WiCLP	25.40
SD v2	23.27
SD v2 + WiCLP	27.40
GORS	30.54

Table 3: Comparison of FID scores between the baseline models and WiCLP using SWITCH-OFF with $\tau = 800$, as well as the GORS approach.

B.3 Experiments with LLaMa3 8B

We explored the analysis of attention contributions to identify unintended attention in LLaMa3 8B, which utilizes a more advanced text encoder specifically designed for language modeling and pretrained on large-scale text corpora. Table 2 reports the rate of unintended attention across prompts in the color and texture datasets. The results demonstrate that unintended attention occurs less frequently in more advanced text encoders, further emphasizing the limitations of the CLIP text encoder.

B.4 Models with T5 text-encoder

We conducted experiments to measure the VQA score on the color dataset for models that use T5 as their text encoder. DeepFloyd achieved a score of 0.604, which is significantly higher than that of SD-v1.4. Additionally, DeepFloyd-I-M, which employs a smaller first-stage UNet compared to DeepFloyd, obtained a score of 0.436, also surpassing the SD-v1.4 score.

C Experiments

C.1 Training setup

In this section, we present the details of the experiments conducted to evaluate our proposed methods. The training is performed for 25,000 steps with a batch size of 4. An RTX A5000 GPU is used for training models based on Stable Diffusion 1.4, while an RTX A6000 GPU is used for models based on Stable Diffusion 2. We employed the Adam optimizer with a learning rate of 1×10^{-5} and utilized a Multi-Step learning rate scheduler with decays ($\alpha = 0.1$) at 10,000 and 16,000 steps. For the `WiCLP`, a window size of 5 was used. All network parameters were initialized to zero, leveraging the skip connection to ensure that the initial output matched the CLIP text embeddings. Our implementation is based on the Diffusers[1] library, utilizing their modules, models, and checkpoints to build and train our models. This comprehensive setup ensured that our method was rigorously tested under controlled conditions, providing a robust evaluation of its performance.

C.2 Extended Evaluation

Human Evaluation We conducted a human evaluation in which participants compared images generated by SD v1.4 and SD v1.4 + `WiCLP`, selecting the image that best matched the given prompt (Figure 19). Five evaluators were presented with 200 randomly selected image pairs, evaluating a total of 1000 image-caption pairs.

TIFA Metric. To provide a more comprehensive evaluation, in addition to the disentangled BLIP-VQA score proposed by

Using TIFA, we observed that SD v1.4 and SD v2 achieved scores of 0.6598 and 0.7735, respectively. Notably, the scores for `WiCLP` applied on top of SD v1.4 and SD v2 improved to 0.7462 and 0.8133, respectively, demonstrating the enhanced performance of our approach.

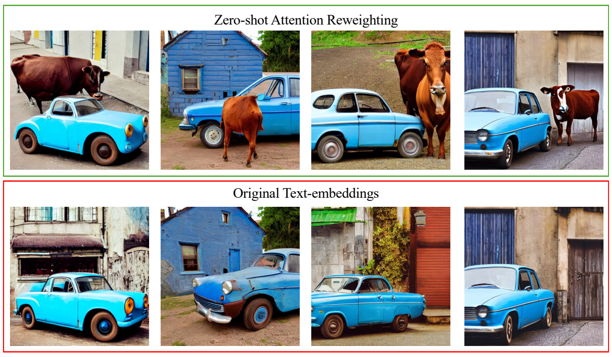


Figure 12: Visualization of some images generated with same set of seeds using original text-embeddings of prompt “a blue car and a brown cow" and text-embeddings that are obtained as the result of zero-shot reweighting of attention matrix.

FID Score Comparison Our method results in a modest increase in FID score on MS-COCO prompts compared to the base models, as shown in Table 3. However, this increase is less pronounced than in other baselines—for example, SD v2 + `WiCLP` scores 27.40, whereas GORS reaches 30.54.

C.3 `CLP` and `WiCLP` Visualization

In this section, we provide additional visualizations comparing `CLP`, `WiCLP`, and baseline models in Figures 15, 16.

C.4 Visualization of Cross-Attentions

In this section, we provide additional cross-attention map visualizations in Figures 15 and 16.

C.5 Visualization of SWITCH-OFF

In this section, we present more qualitative samples illustrating the effect of SWITCH-OFF at different timestep thresholds for various prompts in Figures 17 and 18.

C.6 Choice of Window Length in `WiCLP`

One might suggest that instead of using token-wise linear projection (`CLP`) or a window-based linear projection with a limited window (`WiCLP`), employing a linear projection that considers all tokens when finding a better embedding for each token might yield better results. However, our thorough quantitative study and experiments tested various window sizes for `WiCLP`. We found that using a window size of 5 ($s = 2$) achieves the highest performance.

[1] https://github.com/huggingface/diffusers

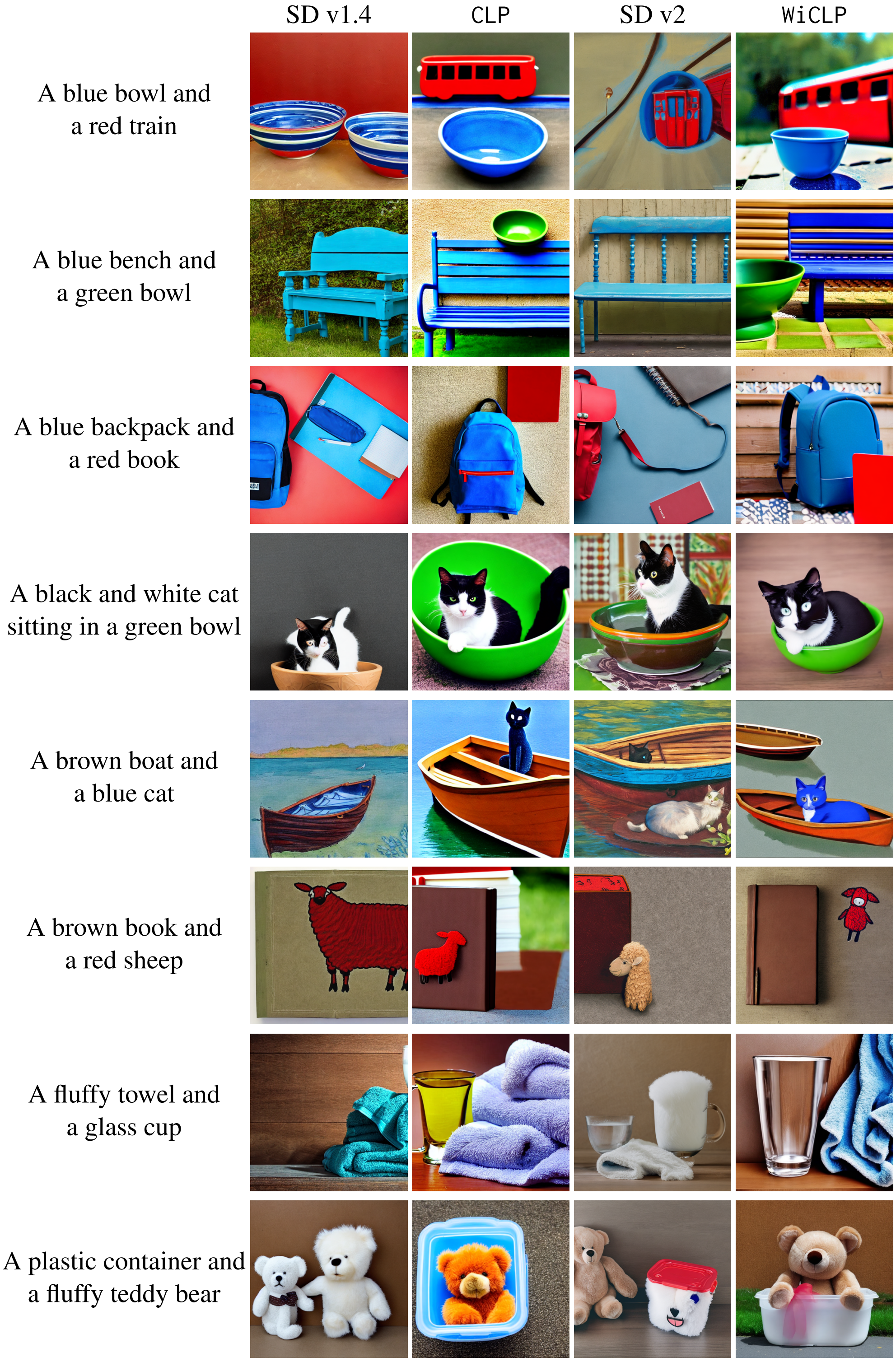

SD v1.4
CLP
SD v2
WiCLP
A blue bowl and a red train
A blue bench and a green bowl
A blue backpack and a red book
A black and white cat sitting in a green bowl
A brown boat and a blue cat
A brown book and a red sheep
A fluffy towel and a glass cup
A plastic container and a fluffy teddy bear

Figure 13: Qualitative comparison between the baseline and our projection methods (CLP and WiCLP).

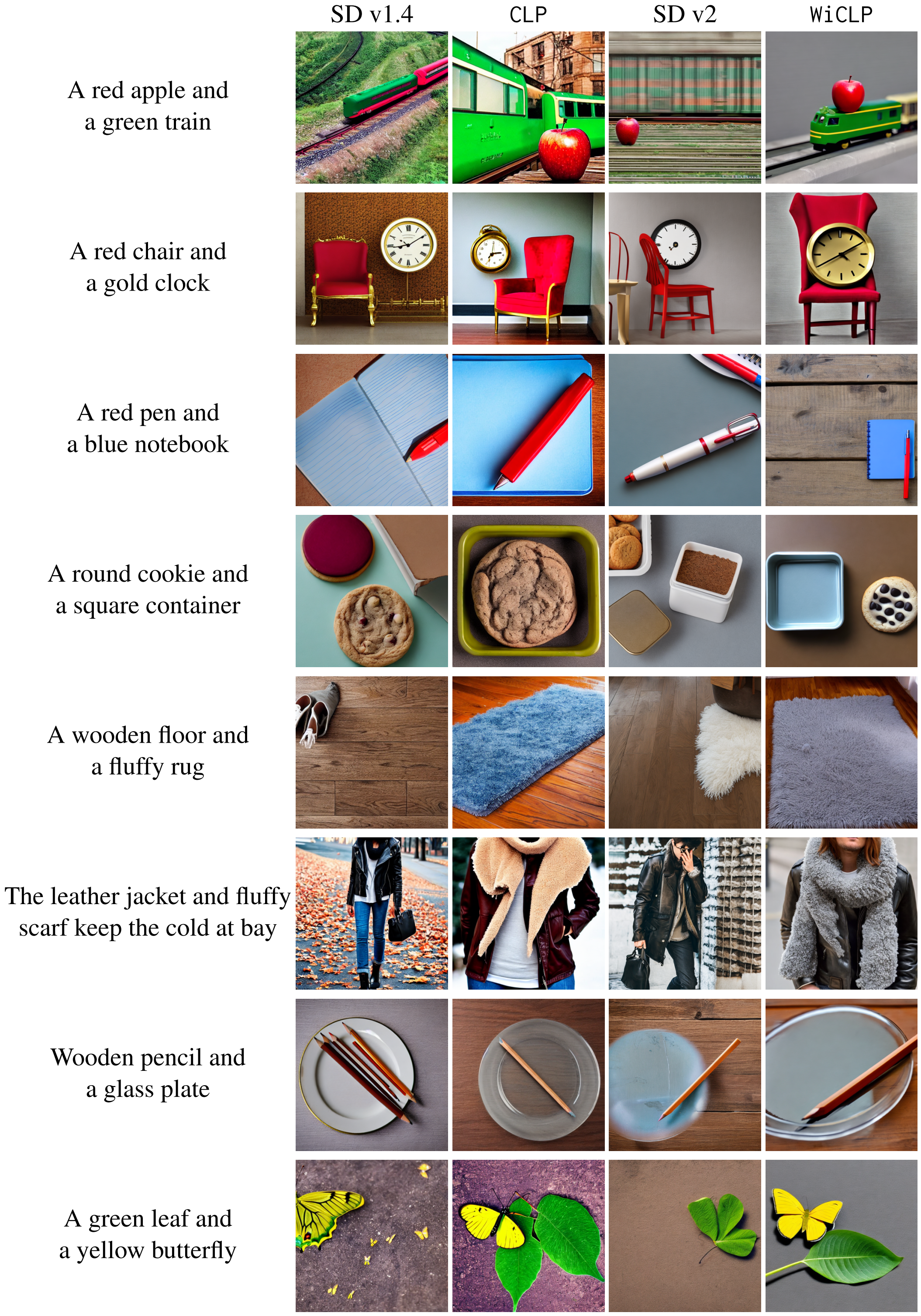

SD v1.4
CLP
SD v2
WiCLP
A red apple and a green train
A red chair and a gold clock
A red pen and a blue notebook
A round cookie and a square container
A wooden floor and a fluffy rug
The leather jacket and fluffy scarf keep the cold at bay
Wooden pencil and a glass plate
A green leaf and a yellow butterfly

Figure 14: Qualitative comparison between the baseline and our projection methods (`CLP` and `WiCLP`).

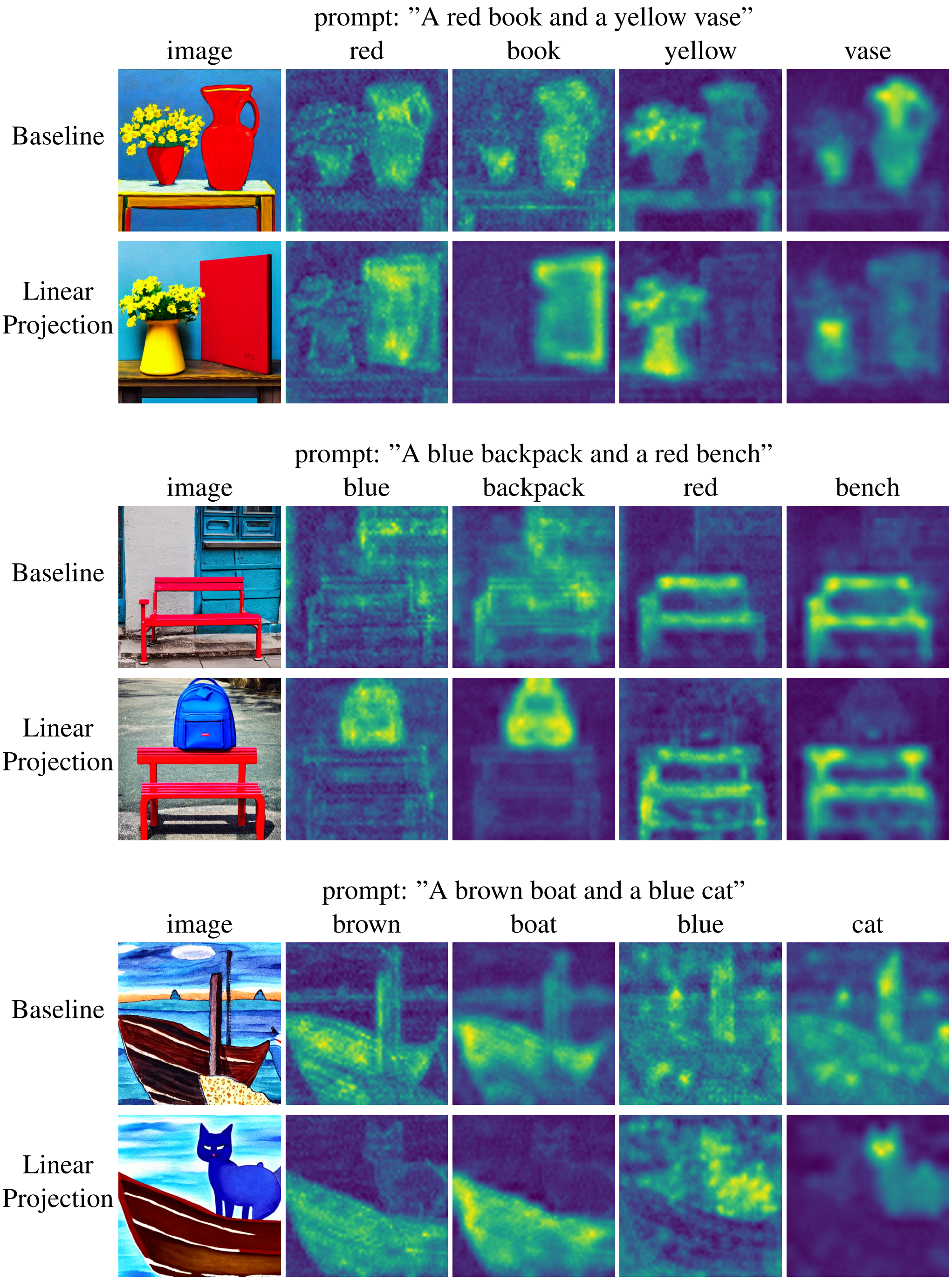


Figure 15: Comparison of cross-attention maps of the U-Net with and without the `CLP`

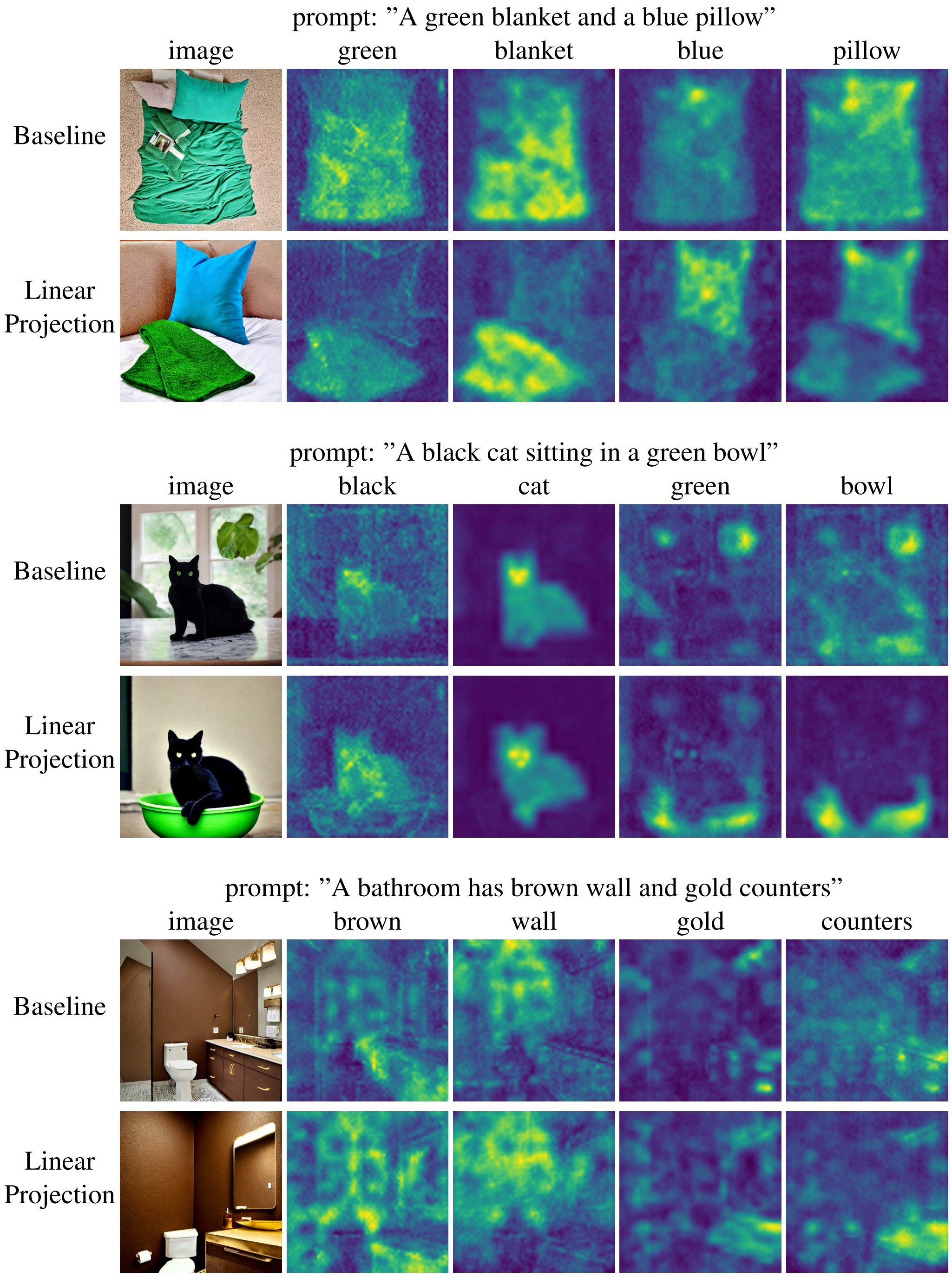


Figure 16: Comparison of cross-attention maps of the U-Net with and without the `CLP`

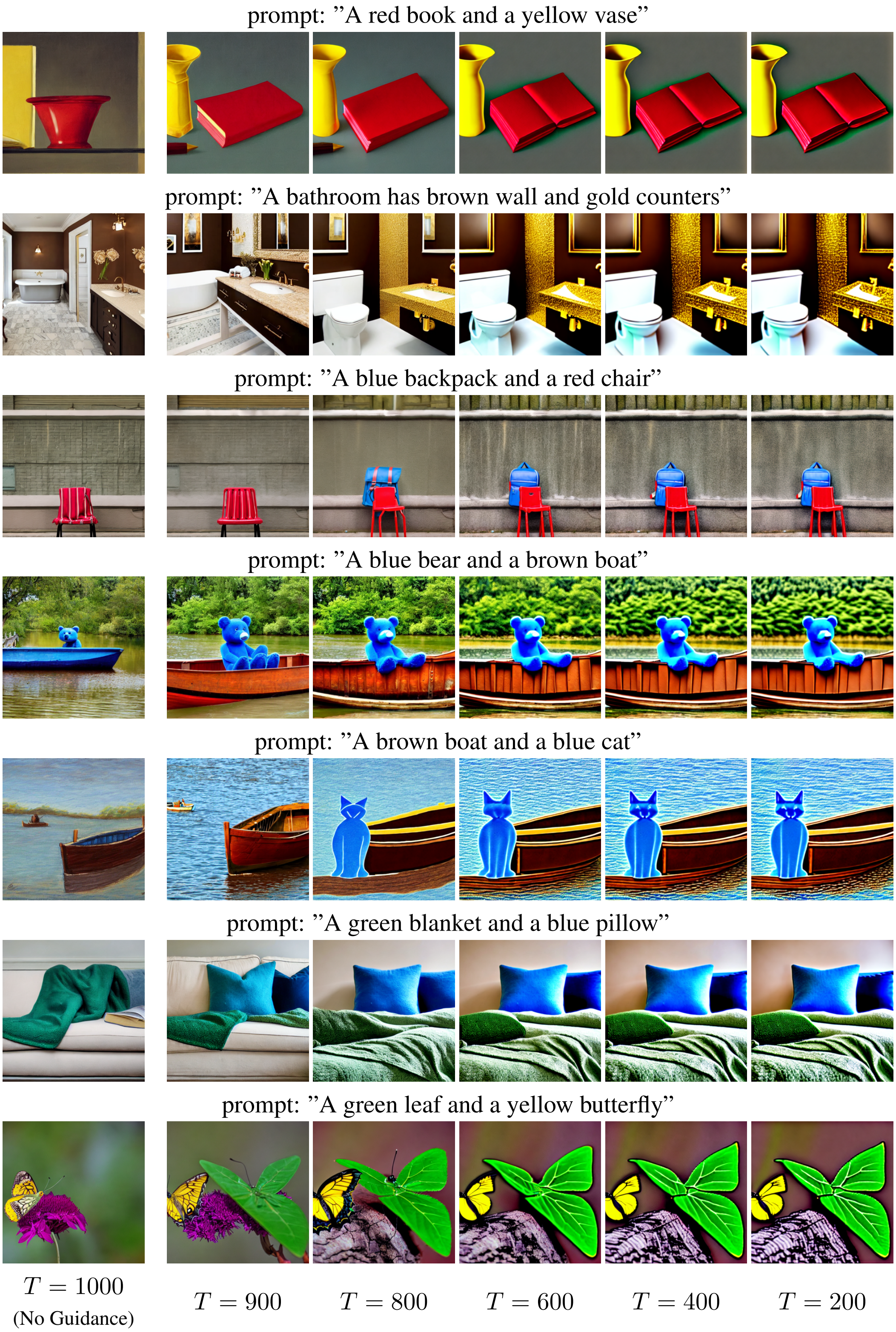

prompt: ”A red book and a yellow vase”
prompt: ”A bathroom has brown wall and gold counters”
prompt: ”A blue backpack and a red chair”
prompt: ”A blue bear and a brown boat”
prompt: ”A brown boat and a blue cat”
prompt: ”A green blanket and a blue pillow”
prompt: ”A green leaf and a yellow butterfly”
$T = 1000$
(No Guidance)
$T = 900$
$T = 800$
$T = 600$
$T = 400$
$T = 200$

Figure 17: Qualitative results showing the impact of SWITCH-OFF with varying thresholds T

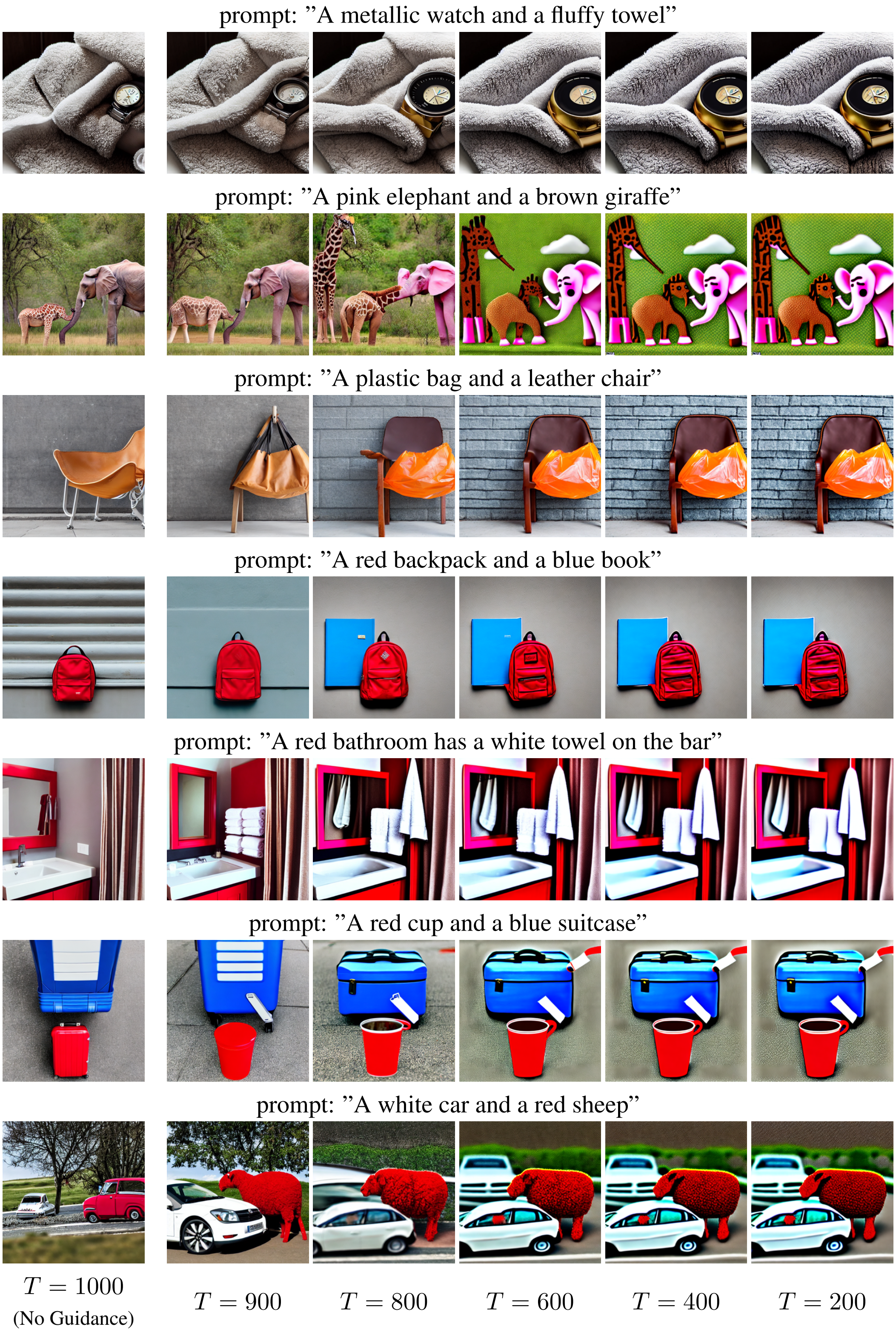
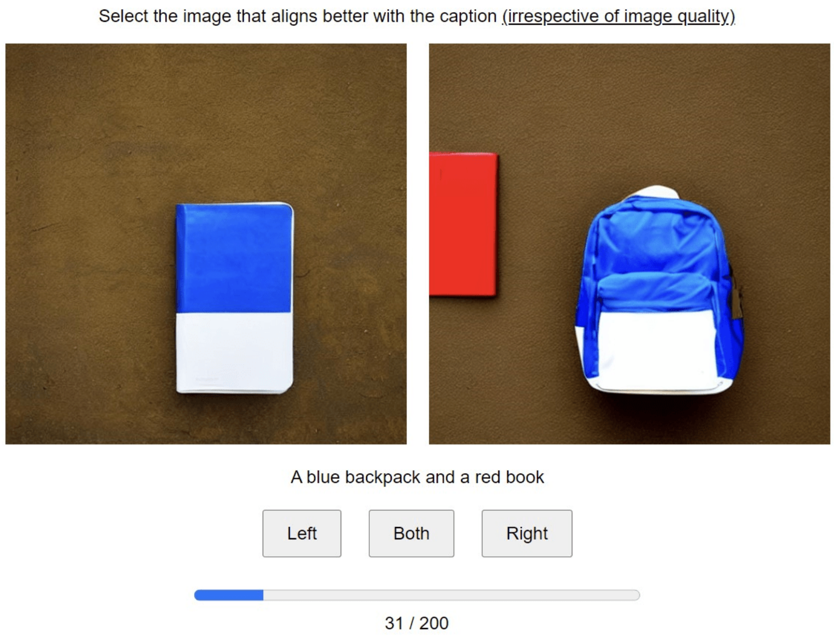

prompt: ”A metallic watch and a fluffy towel”
prompt: ”A pink elephant and a brown giraffe”
prompt: ”A plastic bag and a leather chair”
prompt: ”A red backpack and a blue book”
prompt: ”A red bathroom has a white towel on the bar”
prompt: ”A red cup and a blue suitcase”
prompt: ”A white car and a red sheep”
$T = 1000$
(No Guidance)
$T = 900$
$T = 800$
$T = 600$
$T = 400$
$T = 200$

Figure 18: Qualitative results showing the impact of SWITCH-OFF with varying thresholds T



Select the image that aligns better with the caption (irrespective of image quality)

A blue backpack and a red book

Figure 19: A sample from the human evaluation study, where participants were presented with a pair of images and a caption. They were asked to select the image that best represented the caption or choose ’both’ if the images equally captured the caption’s meaning.