CONFORM: Contrast is All You Need For High-Fidelity Text-to-Image Diffusion Models

Tuna Han Salih Meral[1] Enis Simsar[2†] Federico Tombari[3,4] Pinar Yanardag[1]

[1]Virginia Tech [2]ETH Zürich [3]TUM [4]Google

https://conform-diffusion.github.io

Figure 1. Our training-free method combines a contrastive objective with test-time optimization, significantly improving how models such as Imagen and Stable Diffusion generate images with text prompts consisting of multiple concepts or subjects such as *'a **bear** and a **horse**'*.

Abstract

Images produced by text-to-image diffusion models might not always faithfully represent the semantic intent of the provided text prompt, where the model might overlook or entirely fail to produce certain objects. Existing solutions often require customly tailored functions for each of these problems, leading to sub-optimal results, especially for complex prompts. Our work introduces a novel perspective by tackling this challenge in a contrastive context. Our approach intuitively promotes the segregation of objects in attention maps while also maintaining that pairs of related attributes are kept close to each other. We conduct extensive experiments across a wide variety of scenarios, each involving unique combinations of objects, attributes, and scenes. These experiments effectively showcase the versatility, efficiency, and flexibility of our method in working with both latent and pixel-based diffusion models, including Stable Diffusion and Imagen. Moreover, we publicly share our source code to facilitate further research.

1. Introduction

Diffusion text-to-image models [15] have showcased remarkable progress in generating images using textual cues [33, 34, 37]. These models offer a wide set of capabilities, ranging from image editing [2, 3, 9, 13, 29, 42], personalized content creation [36], and inpainting [25]. However, images produced by these models might not always faithfully represent the semantic intent of the given text prompt [6, 39]. Notable semantic discrepancies in models like Stable Diffusion [34] and Imagen [37] include a) *missing objects* where the model might overlook or entirely fail to produce certain objects; b) *attribute binding* where the model might mistakenly link attributes to the wrong subjects [6]; and c) *miscounting* where the model fails to accurately produce the right quantity of objects [22, 48]. Figure 2 illustrates these shortcomings in popular diffusion models, Stable Diffusion [34] and Imagen [37]. For example, the output might neglect certain subjects, as in the 'a bear and an elephant' prompt, where the bear is ignored as depicted in Fig. 2(a). Additionally, the model might mix up attributes, such as mixing the colors in the 'a purple crown and a yellow suitcase' prompt as seen in Fig. 2(b). Another behavior

†Enis Simsar is affiliated with DALAB at ETH Zürich.

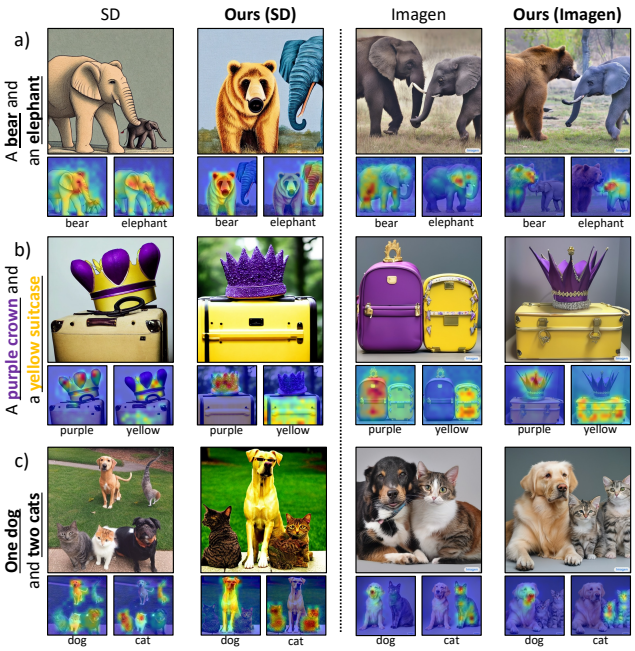

Figure 2. **Failure cases of Stable Diffusion [34] and Imagen [37].** Text-to-image diffusion models may not faithfully adhere to the subjects specified in the text prompt: a) missing objects (*e.g.*, *bear*), b) misaligned attributes (*e.g.*, *the color yellow blends into the crown*), and c) inaccurate object count (*e.g.*, *only one cat is generated instead of two*). Our method steers the diffusion process towards more faithful images in both SD and Imagen.

that is often attributed to the imprecise language comprehension of the CLIP text encoder [28, 30] is the failure to produce the correct quantity of subjects as in Fig. 2(c) where the model either produces an excessive number of cats (SD or failed to include a cat (Imagen) for 'one dog and two cats' prompt.

Recent studies proposed various solutions to these semantic challenges [1, 6, 20, 22, 45]. For example, Chefer et al. [6] optimize cross-attention maps to encourage object presence, while Li et al. [22] use a dual loss function to segregate the attention map into distinct areas of attention and to reinforce attribute association. Kim et al. [20] enhance fidelity by directly adjusting intermediate attention maps according to user-specified layouts. However, a common limitation of these methods is their reliance on tailored objective functions for each issue, leading to sub-optimal performance or challenges when dealing with complex prompts.

Attention maps, which depict the relationship between the input text and the generated pixels, offer a valuable lens for understanding these challenges, as emphasized by prior research [1, 6, 13]. For example, for the 'a bear and an elephant' prompt, a significant overlap is observed in the cross-attention maps dedicated to each subject (refer to Fig. 2(a)). This overlap makes it difficult to differentiate the two subjects and leads the model to produce only *elephants*. Similarly, when prompted to produce a *purple crown* and a *yellow suitcase*, the attentions designated for *purple* and *yellow* are misaligned, causing the model to mistakenly mix colors of both (see the Fig. 2(b)). Regarding counting, the attention maps usually concentrate solely on one region (refer to Fig. 2(c) Imagen), resulting in the generation of an incorrect number of objects, such as *one* cat instead of *two*. Additionally, during the backward process, the attention maps corresponding to various attributes tend to scatter (see Fig. 3). Therefore, to effectively reduce the scattering and ensure more focused and coherent attention allocation, we incorporated attention maps from the previous iteration. This enhances the model's ability to maintain consistency across the generation process, as seen in Fig. 3.

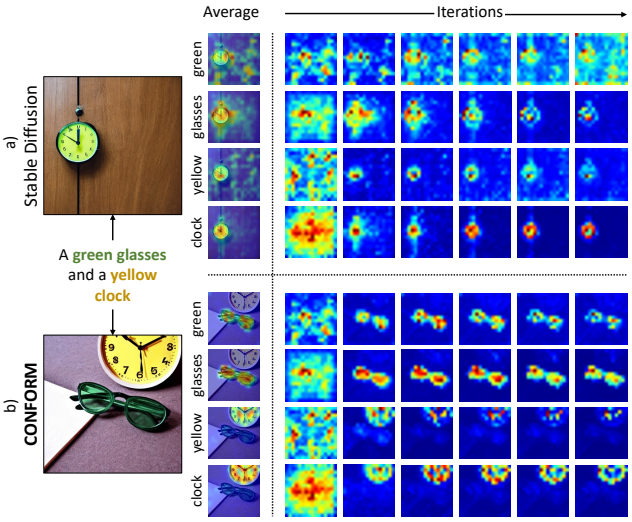

Figure 3. **Attention scattering in backward process.** In Stable Diffusion, the attention to attributes like *green* and *yellow* dissolves over backward timesteps (a). Our method effectively preserves these attention maps (b).

In this work, we tackle the challenge of high-fidelity generation in text-to-image models within a contrastive framework. This framework considers the attributes of a specific object as positive pairs while contrasting them against attributes and objects outside their pairing. For example, in the prompt *'a green dog and a white clock'* (see Fig. 1), *green* and *dog* are treated as mutual positives, while *white* and *clock* become their contrastive counterparts, and vice versa. This approach separates distinct objects within the attention map, addressing the overlapping attention, and encourages distinct high-response areas for each object and attribute. As a result, objects are distinctly separated from one another while being closely associated with their specific attributes, ensuring that the attention map represents both concepts effectively (see Fig. 1 and 2).

The key contributions of our work are as follows:

- We propose a training-free method utilizing a contrastive objective combined with test-time optimization to en-

hance the fidelity of pre-trained text-to-image diffusion models.
- Our approach is model-agnostic, applicable to popular text-to-image diffusion models like Stable Diffusion and Imagen.
- Our comprehensive experiments demonstrate the superiority of our method over baselines and competing approaches, evidenced by its performance on various benchmark datasets and user studies.

2. Related work

Text-to-image diffusion models. Before diffusion-based large-scale conditional image generation models, generative adversarial networks [18, 40, 46, 47, 50, 51], variational autoencoders [17], and autoregressive models [32, 49] were the main focus for both conditional and unconditional image synthesis. However, with the advent of diffusion-based image generation models [15, 26, 38], and their evolution into large-scale text-to-image models [4, 33, 37], they became the state-of-the-art for the text-to-image generation. Although the quality of generated samples increased significantly, it is still a challenge to create images that are faithful to the input prompt. Classifier-free guidance [14] is introduced to enhance text reliance but there is still a need for prompt engineering [24, 43, 44] to produce input prompts so that the generated samples satisfy the intended properties specified in the input prompts.

Improving the fidelity of text-to-image diffusion models. The challenge of aligning text-to-image model outputs with input prompts has been discussed in [39]. They identified that adjectival modifiers and co-hyponyms result in entangled features in cross-attention maps. To address this, Liu et al. [23] introduced ComposableDiffusion allowing users to apply conjunction and negation operators in prompts to guide concept composition. Similarly, StructureDiffusion [10] segments the prompts into noun phrases for more precise attention distribution. Wu et al. [45] developed an algorithm with a layout predictor for spatial layout generation, addressing the cross-attention map control. Agarwal et al. [1] proposed A-star to minimize concept overlap and change in attention maps through iterations. Kim et al. [20] proposed DenseDiffusion, for region-specific textual feature accumulation. Chefer et al. [6] focused on enhancing attention to neglected tokens, and Li et al. [22] proposed two separate tailored objective functions to address the missing objects and wrong attribute binding problems separately. Although these methods are taking steps forward to resolve the mentioned issues, they fail in several cases (see Fig. 5). The Attend and Excite method addresses solely the issue of neglected objects, but it falls short in effectively resolving the problem when the areas of maximum attention are close. On the other hand, Divide and Bind provides an approach to tackle the issue of incorrect attribute binding. However, its use in situations where the tokenizer of text embedding divides single object words into multiple tokens is unclear.

3. Methodology

In this section, we begin by outlining the basics of diffusion models and contrastive learning, followed by a detailed discussion of our methodology. An overview of our method is shown in Fig. 4.

3.1. Diffusion models

We applied our novel approach to two leading text-to-image models: Stable Diffusion (SD) and Imagen. Stable Diffusion operates in the latent space of an autoencoder, where an encoder $\mathcal{E}$ converts the input image x into a lower-dimensional latent code $z = \mathcal{E}(x)$. The decoder $\mathcal{D}$ then reconstructs this latent back into the image space, achieving $\mathcal{D}(z) \approx x$. On the other hand, Imagen operates within pixel space, extending its output via two consecutive image-to-image diffusion models for super-resolution.

Upon having a trained autoencoder, Stable Diffusion employs a diffusion model [15] that is trained within the latent space of the autoencoder. The training process involves gradually adding noise to the original latent code z_0 over time, leading to the generation of z_t at timestep t. This latent code z_0 is in pixel space in Imagen and latent space in Stable Diffusion. A UNet [35] denoiser, denoted as ϵ_θ, is trained to predict the noise added to z_0. The training objective is formally expressed as:

$$\mathcal{L} = \mathbb{E}_{z_t, \epsilon \sim \mathrm{N}(0,\mathrm{I}), c(\mathcal{P}), t} \left[\|\epsilon - \epsilon_\theta(z_t, c(\mathcal{P}), t)\|^2 \right] \quad (1)$$

where $c(\mathcal{P})$ represents the conditional information and $\mathcal{P}$ is the text prompt fed to the text embedding model.

In Stable Diffusion, the sequential embedding of CLIP [30] model c is supplied to a UNet network through a cross-attention mechanism, serving as conditioning to generate images that adhere to the provided text prompt $\mathcal{P}$. In Imagen, a pre-trained T5 XL language model [31] is used as a text-encoder instead. The cross-attention layers perform a linear projection of c into queries (Q) and values (V), and they map intermediate representations from UNet to keys (K). Then, the attention at time t is calculated as $A_t = \mathrm{Softmax}(QK^\intercal/\sqrt{d})$. Notably the attention map at timestep t, A_t, can be reshaped into $\mathbb{R}^{h \times w \times l}$, where h, w represents the resolution of the feature map, which can take values from $\{16 \times 16, 32 \times 32, 64 \times 64\}$, and l corresponds to the sequence length of the text embedding. In our work, we primarily focus on the $\{16 \times 16\}$ attention maps, as they have been identified by Hertz et al. [13] as the most semantically meaningful attention maps.

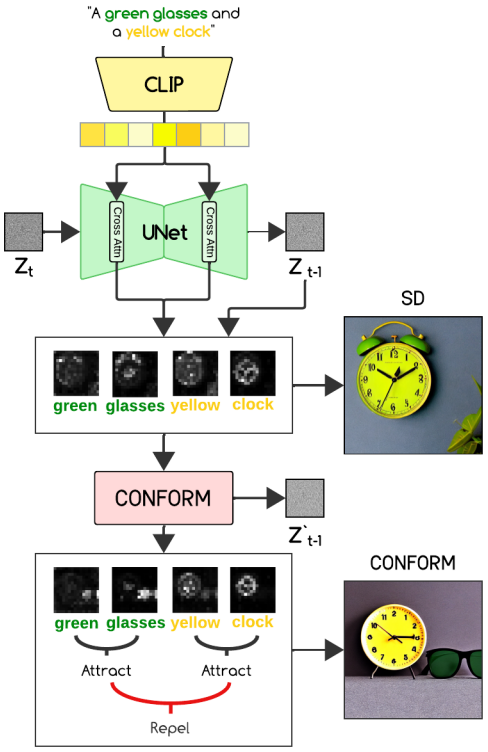

Figure 4. **An overview of CONFORM**. Given a prompt (*e.g.*, *'A green glasses and a yellow clock'*), we extract the subject tokens *green*, *glasses*, *yellow*, and *clock* and their corresponding attention maps ($A^{\texttt{green}}$, $A^{\texttt{glasses}}$, $A^{\texttt{yellow}}$, $A^{\texttt{clock}}$) from timesteps t and $t+1$. We employ our contrastive objective at each time step to repel negative pairs and attract positive pairs.

3.2. Contrastive learning

Contrastive learning has recently gained substantial popularity, delivering state-of-the-art results across multiple unsupervised representation learning tasks [7, 8, 11, 27, 41]. The core objective of contrastive learning is to develop representations that bring similar data points closer while pushing dissimilar data points apart. Let $x \in \mathcal{X}$ represent an input data point. We can define x^+ as a positive pair, where both data points, x and x^+, share the same label, and x^- as a negative pair, in which the data points have different labels. The kernel $f : \mathcal{X} \to \mathbb{R}^N$, takes an input x and generates an embedding vector. InfoNCE, also known as NT-Xent, [7, 12, 27] is one of the popular contrastive learning objectives defined as follows:

$$\mathcal{L} = -\log \frac{\exp(f(x) \cdot f(x^+)/\tau)}{\sum_{i=0}^{M} \exp(f(x) \cdot f(x_i)/\tau)} \tag{2}$$

In this equation, τ is the temperature parameter, regulating the penalties. The summation is performed over one positive sample, denoted as x^+, and M negative samples. Essentially, this loss can be interpreted as the log loss of a softmax-based classifier aiming to classify the data point x as the positive sample x^+. We utilized InfoNCE loss since we will operate on very limited data and need an objective function supporting fast convergence.

3.3. CONFORM

In our approach, we utilize attention maps of object and attribute tokens as features. For a given prompt, such as 'a red **backpack** and a green **suitcase**' we group the objects and their corresponding attributes. For instance, the attention maps for red and **backpack** are grouped together, while green and **suitcase** are put into another group. Consequently, pairs (red, **backpack**) and (green, **suitcase**) are treated as positive, while pairs (red, green), (red, **suitcase**), (**backpack**, green), and (**backpack**, **suitcase**) form negative pairs. Moreover, to maintain the consistency of attention maps through successive steps in the backward diffusion process (see Fig. 3), we also incorporate the attention maps from the timestep $t+1$ into the loss calculation, effectively doubling the token count used to calculate the loss function, creating pairs based on attention maps from the same timesteps, as well as cross-timesteps. This entails, for the color 'red', pairs (red$_t$, red$_{t+1}$), and (red$_t$, **backpack**$_{t+1}$) serving as positive pairs, in addition to those formed from attention maps within the same timestep. Likewise, for the color 'red', we introduce negative pairs like (red$_t$, green$_{t+1}$), and (red$_t$, **suitcase**$_{t+1}$) to the loss calculation. For the contrastive objective, we employ InfoNCE loss, known for its fast convergence compared to previous methods. The InfoNCE loss operates on pairs of cross-attention maps, involving both object and attribute tokens from timestep t and $t+1$. The loss function can be expressed for a given attention map A^j as follows for a single positive pair:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(A^j, A^{j^+})/\tau)}{\sum_{n \in \{j^+, j_1^-, \cdots j_N^-\}} \exp(\text{sim}(A^j, A^n)/\tau)} \tag{3}$$

where sim function represents cosine similarity:

$$\text{sim}(u, v) = \frac{u^T \cdot v}{\|u\|\|v\|} \tag{4}$$

In this equation, τ is the temperature parameter, and the summation in the denominator contains one positive pair and all negative pairs for A^j. We compute the average InfoNCE loss across all positive pairs.

Optimization. In our approach, the loss function consists of a single term, detailed in Section 3.3. We then direct the latent representation in the desired direction as measured by the loss function. Similar to [1, 6], latent representation is updated at each step as follows:

$$z_t' = z_t - \alpha_t \nabla_{z_t} \mathcal{L} \tag{5}$$

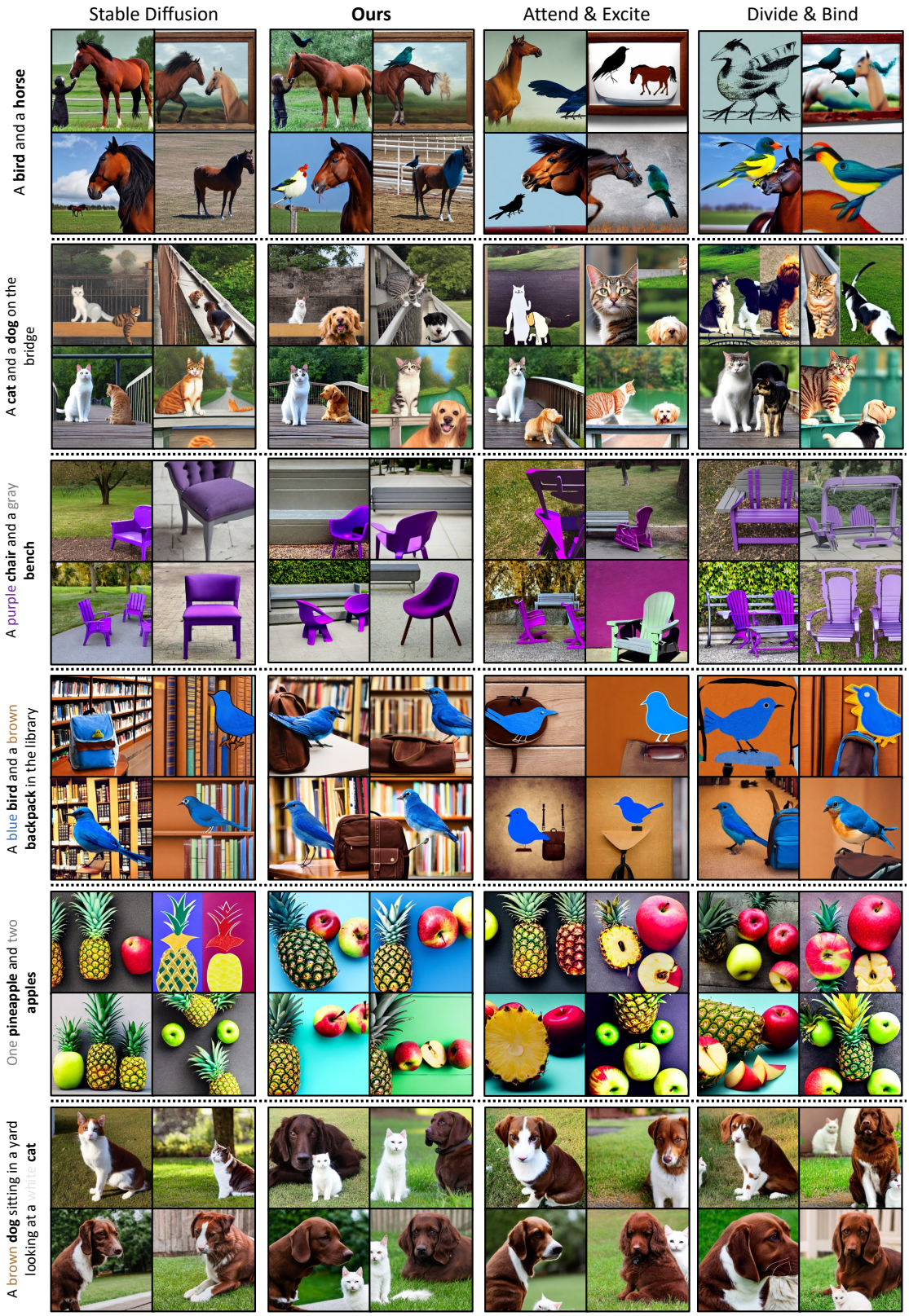

Figure 5. **Qualitative comparison of CONFORM on Stable Diffusion with other state-of-the-art methods**. Our method generates more faithful images for the input text prompt on both simple and complex prompts.

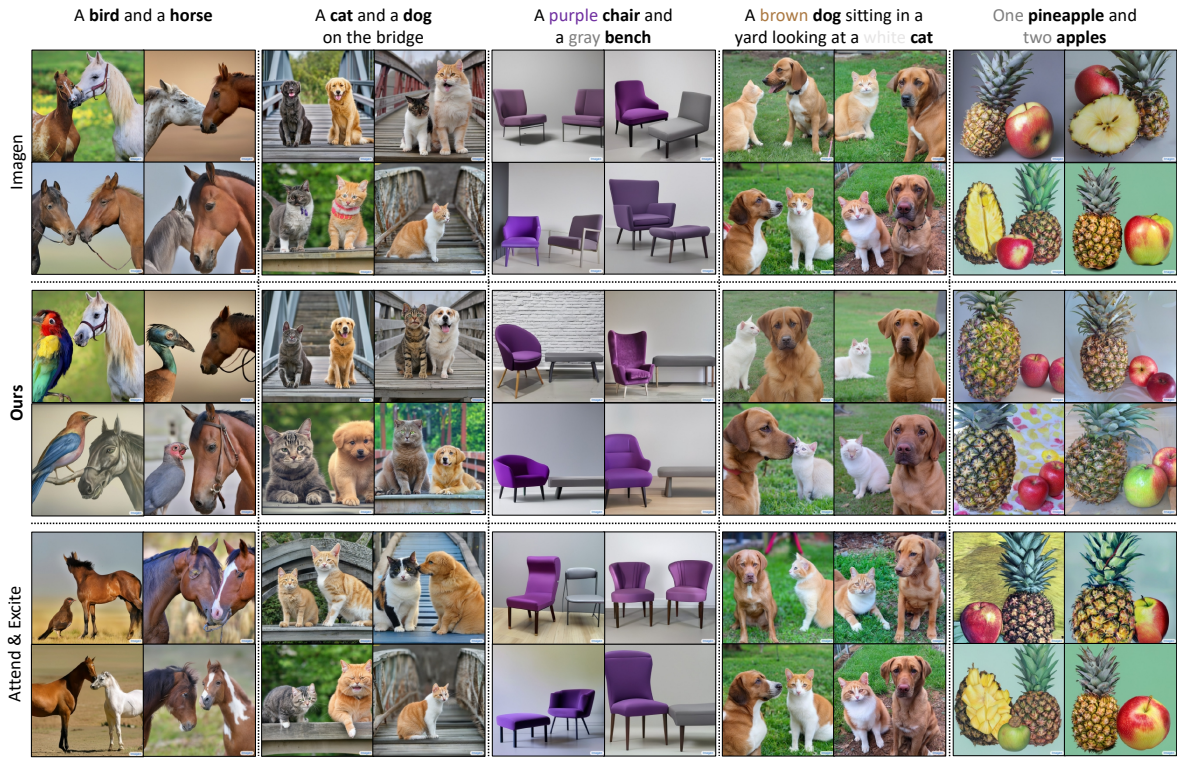

Figure 6. **Qualitative comparison of CONFORM on Imagen.** Our approach consistently produces images that more accurately reflect the input text prompts, effectively handling both simple and complex scenarios in the Imagen model.

Please see the detailed algorithm in *Supplementary Material*.

4. Experiments

Experimental setup. Due to the absence of standardized benchmarks for the evaluation of text-to-image generation models, we adopt a comprehensive evaluation strategy that combines commonly used prompts for qualitative analysis and protocols established in prior works [6, 22] for quantitative assessment. The benchmark protocol we follow comprises structured prompts 'a [animalA] and a [animalB]', 'a [animal] and a [color][object]', 'a [colorA][objectA] and a [colorB][objectB]' [6], and multi-instance prompts from [22]. Details of the benchmark sets and the number of prompts for each benchmark set are detailed in *Supplementary Material*. For each prompt, we use 64 different seeds per prompt, utilizing 50 iterations. Using Stable Diffusion [34] v1.5, the process takes approximately 20 seconds on an NVIDIA L4 GPU. The scale factor α is set to 20 (Eq. 5), and the temperature τ to 0.5 (Eq. 3). To enhance the effectiveness of our updates, we perform optimization multiple times before initiating a backward step at iterations $i \in \{0, 10, 20\}$. After $i = 25$, we also stop any further optimization to prevent unwanted artifacts in the output. Details for the ablation study to determine these parameters are detailed in *Supplementary Material*.

Baselines. We compare our results with several state-of-the-art methods, including Attend & Excite (A&E) [6], Divide & Bind (D&B) [22], ComposableDiffusion (ComposableD.) [23], and StructureDiffusion (StructureD.) [10]. Note that while A-Star [1] is one of our competitors, we are not able to include a comparison since their code is not available.

4.1. Qualitative experiments

Stable Diffusion. Figure 5 presents a side-by-side comparison between CONFORM and other state-of-the-art methods using the Stable Diffusion model. Each method is evaluated using identical input seeds for consistency. CONFORM successfully addresses the issue of missing objects, as demonstrated with the *'A bird and a horse'* text prompt. In scenarios where the Stable Diffusion (SD) model misses the 'bird' in the image, the CONFORM method successfully integrates it, maintaining the image's original semantic integrity. Conversely, competing methods either fail to add the missing object or produce an image significantly different from the original semantic. Our method successfully incorporates missing objects into images featuring scenes, such as the prompt *'A cat and a dog on the bridge'*. Our approach effectively inserts the absent object, like a *dog*, into the image. In cases where the Stable Diffusion (SD)

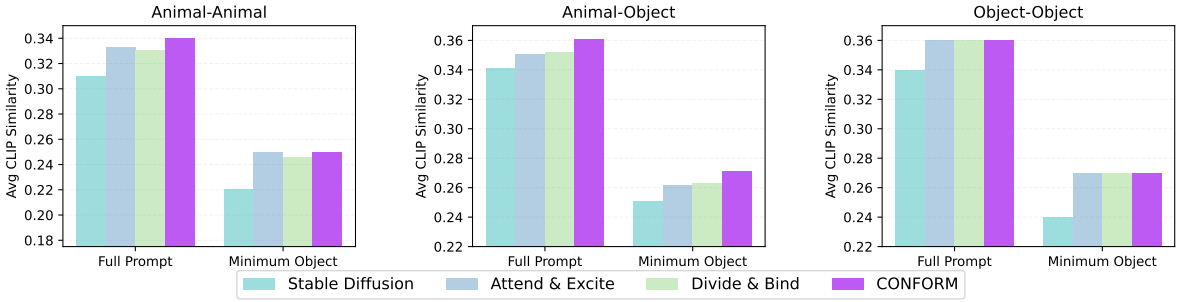

Figure 7. **CLIP similarity scores.** Average CLIP image-text similarities between the text prompts and the images generated by each Stable Diffusion-based method.

model outputs an image with *two cats*, our method can transform one of the *cats* into a *dog*, while preserving the original semantics of the image. In contrast, other methods either fail to respect critical scene components (*e.g.*, the *bridge*) or struggle to generate the correct object. Additionally, our method can handle text prompts where objects are described with specific colors, like in *'A purple chair and a gray bench'*. In such cases, the Stable Diffusion (SD) model often struggles, either failing to generate both objects simultaneously (for instance, omitting the *bench*) or incorrectly assigning colors to objects. Conversely, our technique consistently produces images with both the *chair* and *bench*, accurately applying the designated colors (*e.g.*, *purple* for the *chair* and *gray* for the *bench*). In contrast, other methods tend to merge the colors, resulting in a bench colored in both purple and gray, or they fail to generate the *bench* altogether. Our method handles attribute binding in more complex prompts involving scenes such as *'A blue bird and a brown backpack in the library'* or *'A brown dog sitting in a yard looking at a white cat'*. Unlike Stable Diffusion (SD) and other methods, which often struggle to produce the objects or accurately color them, our method consistently generates both the correct colors and objects. Lastly, our method manages scenarios with specific item quantities. For instance, in the text prompt *'One pineapple and two apples'*, our method accurately produces the correct number of items, whereas Stable Diffusion (SD) and other methods frequently generate an excessive amount of apples.

Imagen. Additionally, the effectiveness of our method with Imagen is demonstrated in Fig. 6. Our primary comparison is against Attend & Excite, adapted to maximize the presence of multiple tokens constituting the target word. Given Imagen's use of the T5 model and its tendency to split words into tokens (*e.g.*, *zebra* becomes *ze, bra*), it is not clear how Divide & Bind approach, specifically the attribute binding regularization, can be applied to Imagen. Notably, CONFORM naturally handles such situations by treating the attention maps of tokens like *ze* and *bra* as positive pairs. Our findings reveal that our method successfully addresses the issue of missing objects, as seen in the '*A bird and a horse*' prompt. Where Imagen originally failed to generate a bird and produced two horses instead, our method effectively substitutes a horse for a bird while maintaining the original semantics of the image. In contrast, Attend & Excite often either fails to generate the image or significantly alters the scene. Similarly, our method successfully handles prompts like *'A cat and a dog on the bridge'*, where Imagen or Attend & Excite typically result in images of two cats; our method replaces one of the cats with a dog. For text prompts involving specific colors, like *'A purple chair and a gray bench'* and *'A brown dog sitting in a yard looking at a white cat'*, our method accurately assigns the colors to the appropriate objects. In contrast, Imagen and Attend & Excite struggle with these tasks, often failing to produce a bench or incorrectly coloring the objects. Lastly, our method successfully generates the accurate number of objects for *'One pineapple and two apples'* prompt, while other methods fail to generate the correct number of apples.

Table 1. Average CLIP text-text similarities between the text prompts and captions generated by BLIP for Stable Diffusion-based methods.

Method	Animal-Animal	Animal-Object	Object-Object
SD	0.76	0.78	0.77
ComposableD.	0.69	0.77	0.76
StructureD.	0.76	0.78	0.76
A&E	0.80	0.83	0.81
D&B	0.81	0.83	0.81
CONFORM	**0.82**	**0.85**	**0.82**

4.2. Quantitative experiments

To quantitatively assess the efficacy of our approach, we employ multiple metrics, including image-text similarity, text-text similarity, and the recently introduced TIFA score [16]. We assess image-text similarity using the CLIP similarity metric, comparing the generated image with the input prompt. We calculate both the full-prompt similarity (CLIP-full), representing the likeness between the entire prompt and the generated image, and the minimum object similarity (CLIP-min), which is the minimum of the similarities

Table 2. Average TIFA scores for SD and Imagen.

Method	Animal-Animal	Animal-Obj	Obj-Obj	Multi-Obj
SD	0.68	0.80	0.65	0.59
A&E	0.92	0.91	0.82	0.72
D&B	0.93	0.91	0.83	0.73
CONFORM	**0.95**	**0.94**	**0.88**	**0.74**
Imagen	**0.84**	0.93	0.88	0.73
A&E	**0.84**	0.93	0.88	0.73
CONFORM	**0.84**	**0.94**	**0.91**	**0.76**

between the generated image and each of the two subject prompts. It is noteworthy that while our model achieved comparable or higher results compared to the reference methods, these metrics should be interpreted with caution, as the models used for comparison are already conditioned on CLIP embeddings. In direct comparison with Stable Diffusion, our method outperformed in both CLIP-full and CLIP-min similarity scores across most of the benchmark sets while performing similarly at others (see Fig. 7).

For text-text similarity, we leverage BLIP [21] to generate captions for the generated image. Then, we evaluate the similarity between the input prompt and these captions. This assessment aims to capture subjects and attributes present in the original prompt, providing insights into the coherence and relevance of the textual descriptions. In comparative analysis with Stable Diffusion and other competitors, our method consistently demonstrated superior performance in text-text similarity scores across all benchmark sets (see Tab. 1).

The TIFA score [16] provides an evaluation by assessing the faithfulness of the generated image to the input prompt. To compute the TIFA score, we automatically generate a set of question-answer pairs using the GPT-3.5 [5] language model. Image faithfulness is then determined by evaluating the proportion of correct answers using the visual question answering model UnifiedQA-v2 [19]. This metric offers a comprehensive evaluation by considering both textual and visual aspects of the generated content. Our method consistently outperforms across all benchmark sets in the TIFA metric in SD. In addition, CONFORM outperforms Attend & Excite and Imagen in most of the benchmarks while performing similarly at the 'Animal-animal' benchmark (see Tab. 2).

User study. To evaluate the fidelity of images generated by our model, we conducted a user study involving 25 participants. We selected 10 random prompts and generated four images for each using different seeds. This process was repeated separately for both Stable Diffusion-based and Imagen-based models. Participants were asked to choose the image reflecting the text prompt best for each combination of prompt and seed. Results, detailed in Tab. 3, overwhelmingly favored CONFORM. For Stable Diffusion,

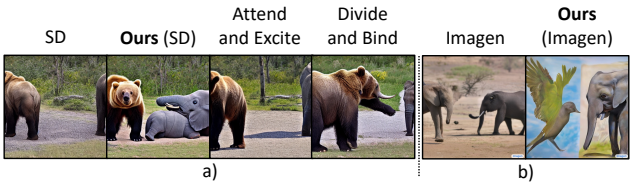

Figure 8. **Limitations.** (a) Shows challenges when key objects are missing from SD image for the 'A bear and an elephant' prompt. (b) Displays object separation in Imagen, despite enhanced prompt accuracy for the 'A bird and an elephant' prompt.

Table 3. User study with 25 respondents for SD and Imagen.

Method	Animal-Animal	Animal-Obj	Obj-Obj	Multi-Obj
SD	5%	3%	0%	5%
A&E	14.75%	2%	7.5%	2%
D&B	8.25%	1%	4.5%	2%
CONFORM	**72%**	**94%**	**88%**	**91%**
Imagen	3.25%	4%	2%	3.5%
CONFORM	**96.75%**	**96%**	**98%**	**96.5%**

CONFORM led in all categories, achieving 72% to 94% of the votes across different benchmark sets. For Imagen study, it similarly dominated, receiving 96% to 98% of the votes. These results highlight CONFORM's effectiveness in closely aligning generated images with the text prompts.

5. Limitations

A limitation of our method based on SD: when the initial map significantly excludes objects, ours may struggle to generate successful images, although it is still able to place the desired objects into the generated image (Fig. 8(a)). This issue does not apply to our method; others also encounter difficulties when starting with challenging attention maps. In Imagen, our refinement process might sometimes lead to the separation of objects, yet it still enhances the accuracy of the text prompt in the final image (Fig. 8(b)).

6. Conclusion

In our study, we introduced a novel framework centered on a contrastive objective, designed to enhance the fidelity of text-to-image diffusion models. Our approach is model-agnostic and applied to popular text-to-image generators like Stable Diffusion and Imagen. Through comprehensive experiments on multiple benchmark datasets, we assessed our method using text-image similarity, text-text similarity, and TIFA scores, comparing it with several leading techniques. Our findings reveal that our method consistently produces images that are more faithful to the original text prompts than the baseline methods in both Stable Diffusion and Imagen models.

References

[1] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. *arXiv preprint arXiv:2306.14544*, 2023. 2, 3, 4, 6

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 1

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 8

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1, 2, 3, 4, 6

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[8] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 539–546. IEEE, 2005. 4

[9] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 1

[10] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*, 2023. 3, 6

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 1735–1742. IEEE, 2006. 4

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4

[13] 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. 1, 2, 3

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 3

[16] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 7, 8

[17] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 783–791, 2017. 3

[18] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023. 3

[19] Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*, 2022. 8

[20] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023. 2, 3

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 8

[22] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. *arXiv preprint arXiv:2307.10864*, 2023. 1, 2, 3, 6

[23] 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. 3, 6

[24] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2022. 3

[25] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 1

[26] 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. *arXiv preprint arXiv:2112.10741*, 2021. 3

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[28] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*, 2023. 2

[29] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 1

[30] 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*, pages 8748–8763. PMLR, 2021. 2, 3

[31] 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. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[32] 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*, pages 8821–8831. PMLR, 2021. 3

[33] 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. 1, 3

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 6

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[36] 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. 1

[37] 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. 1, 2, 3

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[39] Raphael Tang, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022. 1, 3

[40] 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*, pages 16515–16525, 2022. 3

[41] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 4

[42] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 1

[43] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 3

[44] Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462*, 2022. 3

[45] Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7766–7776, 2023. 2, 3

[46] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 3

[47] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021. 3

[48] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022. 1

[49] 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. 3

[50] 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*, pages 833–842, 2021. 3

[51] 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*, pages 5802–5810, 2019. 3