This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# INITNO: Boosting Text-to-Image Diffusion Models via Initial Noise Optimization

Xiefan Guo[1,2] Jinlin Liu Miaomiao Cui Jiankai Li[1,2] Hongyu Yang[3,4] Di Huang[1,2*]

[1]State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Institute of Artificial Intelligence, Beihang University, Beijing, China
[4]Shanghai Artificial Intelligence Laboratory, Shanghai, China

{xfguo,lijiankai,hongyuyang,dhuang}@buaa.edu.cn

## Abstract

*Recent strides in the development of diffusion models, exemplified by advancements such as Stable Diffusion, have underscored their remarkable prowess in generating visually compelling images. However, the imperative of achieving a seamless alignment between the generated image and the provided prompt persists as a formidable challenge. This paper traces the root of these difficulties to invalid initial noise, and proposes a solution in the form of Initial Noise Optimization (*INITNO*), a paradigm that refines this noise. Considering text prompts, not all random noises are effective in synthesizing semantically-faithful images. We design the cross-attention response score and the self-attention conflict score to evaluate the initial noise, bifurcating the initial latent space into valid and invalid sectors. A strategically crafted noise optimization pipeline is developed to guide the initial noise towards valid regions. Our method, validated through rigorous experimentation, shows a commendable proficiency in generating images in strict accordance with text prompts. Our code is available at* https://github.com/xiefan-guo/initno.

## 1. Introduction

Text-to-Image Synthesis (T2I) stands at the forefront of cutting-edge research, dedicated to the generation of authentic and visually cohesive images from text prompts. In the domain of generative models, including generative adversarial networks [13, 18–21], variational autoencoders [22], and autoregressive models [7, 10, 36, 37], diffusion models [8, 16] have ascended as a predominant solution. The integration of large-scale vision-language models [2, 11, 14, 17, 30–32, 40, 41] has propelled noteworthy advancements in the text-to-image domain.

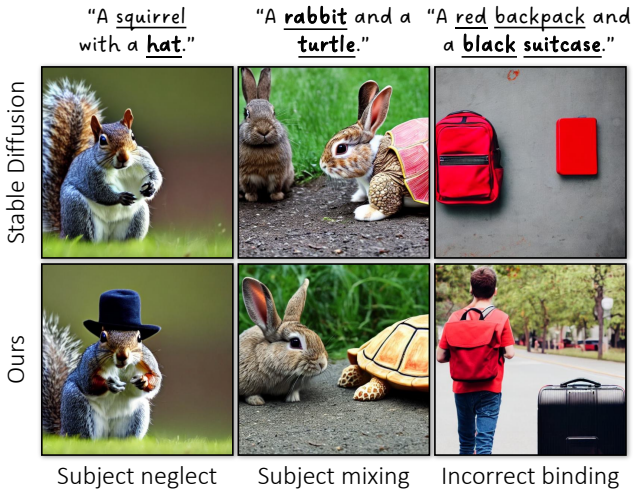


Figure 1. **Example results synthesized by SD and ours.**

Despite the training of state-of-the-art T2I diffusion models on large-scale text-image datasets, synthesizing images precisely aligned with given text prompts remains a considerable challenge. Well-documented issues, *i.e.*, subject neglect, subject mixing, and incorrect attribute binding, as illustrated in Fig. 1, persist. We attribute these challenges to the presence of invalid initial noise.

When different noise inputs are introduced to the T2I diffusion model with identical text prompts, a substantial discrepancy is observed in the alignment between images and the provided text, as depicted in Fig. 2. This observation signifies that not all randomly sampled noise can produce visually-consistent images. Depending on the consistency between the generated image and the target text, the initial latent space can be divided into valid and invalid regions. Noise sourced from valid regions, when input into the T2I diffusion model, results in semantically-reasonable image. Consequently, our aim is to direct any initial noise towards the valid region, thereby facilitating the generation

*Corresponding author.

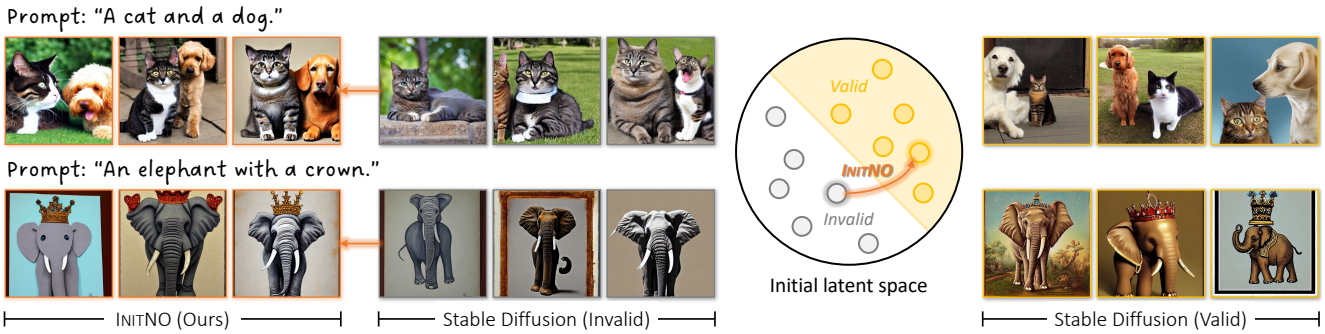


Figure 2. **InitNO.** Our investigation dives into the exploration of various random noise configurations and their subsequent influence on the generated results. Notably, when different noises are input into SD under identical text prompts, there are marked discrepancy in the alignment between the generated image and the given text. Unsuccessful cases are delineated by gray contours, while successful instances are indicated by yellow contours. This observation underscores the pivotal role of initial noise in determining the success of the generation process. Based on this insight, we divide the initial noise space into valid and invalid regions. Introducing Initial Noise Optimization (InitNO), identified as orange arrow, our method is capable of guiding any initial noise into the valid region, thereby synthesizing high-fidelity results (orange contours) that precisely correspond to the given prompt. The same location employs the same random seed.

of visually-coherent image that faithfully aligns with the input prompt.

The core of this paper lies in the concept of Initial Noise Optimization (InitNO), which encompasses initial latent space partitioning and the noise optimization pipeline, responsible for defining valid regions and directing initial noise navigation, respectively.

Our study thoroughly examines attention layers in the diffusion model, specifically the cross-attention map and the self-attention map. The former captures the correlation between text and image features, while the latter captures the correlation among different spatial positions within the image features. These maps effectively measure subject neglect and subject mixing, respectively (see Fig. 3). With these maps, we design the cross-attention response score and the self-attention conflict score. By setting score thresholds, we achieve further partitioning of the initial latent space. Interestingly, we find that accurate rendering of the subject fosters correct attribute binding.

The primary challenge in initial noise navigation is to balance under-optimization and over-optimization of noise. While under-optimization is still plagued by misalignment, over-optimization risks deviating from the initial distribution, *i.e.*, standard Gaussian distribution, resulting in distorted images (see Fig. 9). To address this challenge, we introduce a carefully-crafted noise optimization pipeline. With the explicit modeling of the initial latent space, a novel distribution alignment loss is incorporated to ensure the optimized noise adheres to the initial distribution.

We establish the superiority of the proposed method over state-of-the-art approaches in generating semantically accurate images. Furthermore, our method is plug-and-play and seamlessly integrates into existing diffusion models, enabling training-free controllable generation and improving alignment with specific conditions, such as layout-to-image generation (see Sec. 5.5).

## 2. Related Work

Text-to-image synthesis strives to generate visually-realistic image that reflects given text. Early research mainly focused on GANs [34, 39, 42–44] and autoregressive models [5, 9, 29, 41]. Recently, diffusion models [8, 16] have taken over the mainstream, achieving impressive results. The incorporation of large-scale vision-language models [14, 17, 31, 32, 41] has fostered significant advancements in the text-to-image domain. Despite synthesizing highquality images, it remains challenging to produce results that properly comply with the given text prompt.

To deal this problem, some works [2, 30, 32, 33, 40] expand the network architecture and introduce large language models [28] to better hint embeddings. However, these works require training new text-to-image models and are not applicable to existing widely used models. Another line of work explores training-free improvement strategies, consistent with our approach.

Specifically, Liu *et al.* [25] propose to use multiple diffusion models for different concepts and then composite the outputs to obtain the final result. This method solves the mentioned problem to some extent, but still suffers from subject mixing . Feng *et al.* [12] propose to split the input text prompt and process the cross attention layers to align the tokens and the outputs. Chefer *et al.* [6] introduce the concept of Generative Semantic Nursing (GSN), and slightly shifts the noisy image at each timestep of the denoising process, where the semantic information from the text prompt is better considered. Agarwal *et al.* [1] and Li *et al.* [24] further improve the optimization objective on this paradigm for better update direction. However, updating the noisy image at each denoising step requires carefully-

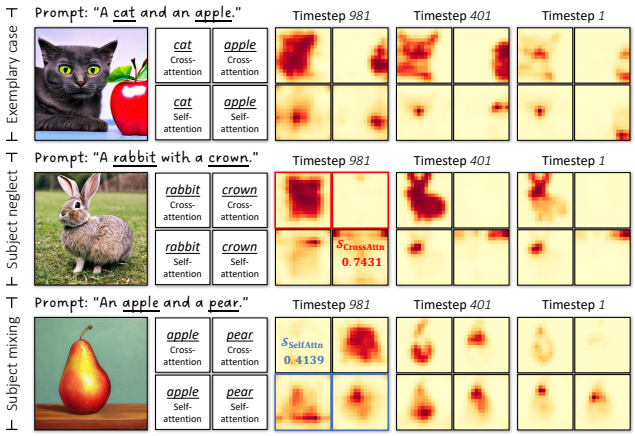


Figure 3. **Visualization of the attention maps.**

designed optimization parameters, posing a delicate balance between under-optimization and over-optimization.

Different from previous methods, we explore a novel path by adjusting the sampled noise in the initial latent space, where full optimization is performed. It cleverly avoids the trade-off between under-optimization and over-optimization, is proven to be efficient and orthogonal to existing methods.

## 3. Preliminaries

**Stable Diffusion Model.** This study focuses on the state-of-the-art Stable Diffusion model (SD) [31], belonging to the category of latent diffusion models (LDMs). Unlike traditional diffusion-based approaches operating directly in the image domain, SD functions within the latent space of an autoencoder. Specifically, an encoder $\mathcal{E}(\cdot)$ learns the mapping from an input image $\mathbf{x} \in \mathcal{X}$ to a latent code $\mathbf{z} = \mathcal{E}(\mathbf{x})$. Subsequently, a decoder $\mathcal{D}(\cdot)$ reconstructs the input image, aiming for $\mathcal{D}(\mathcal{E}(\mathbf{x})) \approx \mathbf{x}$.

Once the autoencoder is trained, a Denoising Diffusion Probabilistic Model (DDPM) [16] operates within the acquired latent space to generate a denoised version of an input latent $\mathbf{z}_t$ at each timestep $t$. During denoising, the diffusion model can be conditioned on additional inputs. In the case of SD, this condition typically takes the form of a text embedding generated by a pre-trained CLIP text encoder [27]. The conditioning embedding for the given prompt $\mathbf{y}$ is denoted as $\mathbf{c} = f_{\text{CLIP}}(\mathbf{y})$. The DDPM model $\epsilon_\theta(\cdot)$, parametrized by $\theta$, optimizes the following loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}\sim\mathcal{E}(\mathbf{x}),\mathbf{c},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{1}),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t)\|_2^2\right], \quad (1)$$

During inference, a latent variable $\mathbf{z}_T$ is sampled from the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$ and subjected to sequential denoising procedures of DDPM to derive a refined latent representation $\mathbf{z}_0$. This denoised latent $\mathbf{z}_0$ is then fed into the decoder $\mathcal{D}(\cdot)$ to synthesize the corresponding image $\mathcal{D}(\mathbf{z}_0)$.

**Cross-attention layer.** Text-image correspondence in SD is achieved through the cross-attention layer, enabling text condition guidance. Specifically, the pre-trained CLIP text encoder [27] embeds the text prompt $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n\}$ into a sequential embedding, serving as the condition $\mathbf{c} = f_{\text{CLIP}}(\mathbf{y})$. Linear projections are then employed to extract the key $\mathbf{K}$ and value $\mathbf{V}$ from $\mathbf{c}$, while the query $\mathbf{Q}$ is mapped from the intermediate features of UNet. The cross-attention map $\mathbf{A}^c$ is computed via:

$$\mathbf{A}^c = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right), \quad (2)$$

where $d$ is the channel dimension. For ease of representation, we omit the denoising timestep $t$. We denote the attention map that corresponds to the $i$-th text token as $\mathbf{A}^c_{\mathbf{y}_i}$, and $\mathbf{A}^c_{\mathbf{y}_i}[x, y]$ is the probability assigned to text token $\mathbf{y}_i$ for the $(x, y)$-th spatial patch of the intermediate feature map.

**Self-attention layer.** In the self-attention layer, intermediate features serve as the key $\mathbf{K}$ and value $\mathbf{V}$ instead of the text condition $\mathbf{c}$, allowing synthesis of globally coherent structures by correlating image tokens across diverse regions. The self-attention map $\mathbf{A}^s$ is computed using the same formula as Eq. 2, and $\mathbf{A}^s_{x,y}$ denotes the attention map that corresponds to the $(x, y)$-th spatial patch.

## 4. InitNO

The core of our methed is the Initial Noise Optimization (InitNO), which comprises the initial latent space partitioning and the noise optimization pipeline, responsible for defining valid regions and steering noise navigation, respectively. This section provides detailed explanations of both components in Sec. 4.1 and 4.2.

### 4.1. Initial latent space partitioning

Prevailing studies [4, 6, 15, 26, 35] demonstrate the informativeness of cross-attention maps, elucidating their capability to partially quantify the alignment between generated image and provided text. In this work, we extend these findings by thoroughly investigating the self-attention maps, which capture inter-spatial correlations within image features. We argue that cross-attention maps and self-attention maps can inherently gauge subject neglect and subject mixing, respectively. Leveraging these maps, we formulate the cross-attention response score and the self-attention conflict score. By establishing the score threshold, we achieve the partitioning of the initial latent space.

**Cross-attention response score.** We follow [6] and derive the final cross-attention map by averaging the cross-attention maps in all layers and heads with a resolution of 16×16 pixels. The aggregated map $\mathbf{A}^c \in \mathbb{R}^{16\times16\times n}$ contains $n$ spatial attention maps, each corresponding to a token of the text prompt. We also re-weight the attention values

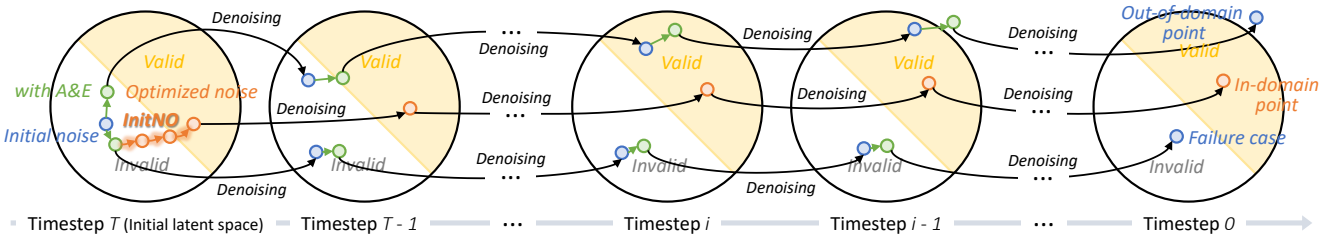


Figure 4. **Comparison of Attend-and-Excite with our method.** Attend-and-Excite suffers from the trade-off between under-optimization and over-optimization. Under-optimization (*lower path*) remains confined to invalid regions, while over-optimization (*upper path*) carries the risk of deviating from the distribution of diffusion model. Our approach (*middle path*) skillfully addresses this challenge by prioritizing noise optimization in the initial latent space, ensuring sufficient and appropriate optimization.

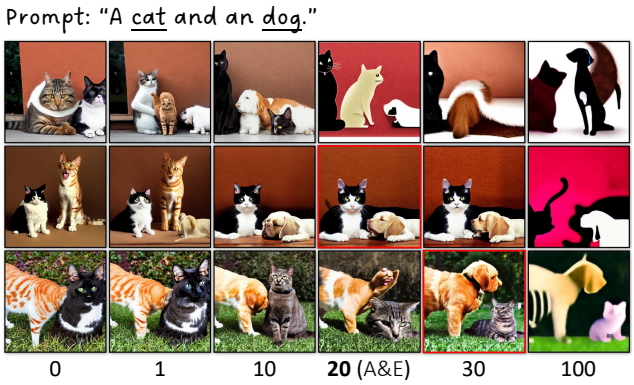


Figure 5. **Effect of the scale factor of Attend-and-Excite.** Given the same text prompt, we adjust the scale factor of Attend-and-Excite [6], 20 is used in the original work. Images of varying quality are synthesized, and the red box indicates the highest quality image. Images on the same row share an identical random seed.

by ignoring the attention of the specialized token sot and performing a softmax operation on the remaining tokens. Fig. 3 shows the cross-attention map for each subject token, reflecting the spatial layout information corresponding to the generated image. Notably, a proficient denoising process can assign sufficiently high cross-attention responses to each subject during the early stages.

Conversely, a low attention response can lead to the absence of the target subject, *i.e.*, subject neglect. As depicted in the red box in Fig. 3, the token *crown* registers a significantly lower response than *rabbit* in the initial denoising step, and *crown* is missing from the generated image. Building upon the insights from [6], we define the cross-attention response score as:

$$\mathcal{S}_{\text{CrossAttn}} = 1 - \min_{\mathbf{y}_i \in \mathcal{Y}} \max \left( \mathbf{A}^c_{\mathbf{y}_i} \right), \tag{3}$$

where $\mathcal{Y}$ denotes the set of target tokens. Specifically, in this work, $\mathcal{S}_{\text{CrossAttn}}$ is calculated over the initial noise and the first denoising step. We further define the cross-attention response score threshold $\tau_c$, empirically set to 0.2 in our experiments, where scores below this threshold are classified as valid, and those above are deemed invalid.

**Self-attention conflict score.** Consistent with the cross-attention map extraction strategy, we extract the final self-attention map $\mathbf{A}^s \in \mathbb{R}^{16\times16\times(16\times16)}$. For the subject token $\mathbf{y}_i$ of interest, we first query the spatial coordinates $(x_i, y_i)$ corresponding to its maximum cross-attention value:

$$x_i, y_i = \arg\max_{x,y} \ \mathbf{A}^c_{\mathbf{y}_i}[x, y]. \tag{4}$$

Subsequently, we obtain its corresponding self-attention map $\mathbf{A}^s_{x_i,y_i}$, as shown in Fig. 3. For an exemplary denoising process, the self-attention maps of different subject tokens can be distinctly partitioned in the spatial domain. However, existing diffusion models, *e.g.*, SD, suffer from self-attention map overlap, leading to a failure case of subject mixing. As indicated in the blue box in Fig. 3, the self-attention maps of *apple* and *pear* significantly overlap, resulting in a combination of both. To quantify this phenomenon, we introduce the self-attention conflict score:

$$\begin{aligned} \mathcal{S}_{\text{SelfAttn}} &= \sum_{\mathbf{y}_i, \mathbf{y}_j \in \mathcal{Y}, \forall i<j} \frac{\mathrm{f}(\mathbf{y}_i, \mathbf{y}_j)}{N}, \\ \mathrm{f}(\mathbf{y}_i, \mathbf{y}_j) &= \frac{\sum_{x,y} \min\left( \mathbf{A}^s_{x_i,y_i}[x, y], \mathbf{A}^s_{x_j,y_j}[x, y] \right)}{\sum_{x,y} \left( \mathbf{A}^s_{x_i,y_i}[x, y] + \mathbf{A}^s_{x_j,y_j}[x, y] \right)}, \end{aligned} \tag{5}$$

where $N$ is the number of pairs of $\mathbf{y}_i$ and $\mathbf{y}_j$. $\mathcal{S}_{\text{SelfAttn}}$ is also calculated based on the initial noise and the first denoising step. We further define the self-attention conflict score threshold $\tau_s$, set to 0.3 in our experiments. Scores below $\tau_s$ are classified as valid, and those above as invalid.

Finally, we define the initial noise that passes both the cross-attention response score test and the self-attention conflict test as valid noise, and vice versa as invalid.

### 4.2. Noise optimization pipeline

As illustrated in Fig. 4, existing methods most relevant to ours, represented by Attend-and-Excite [6], modify the noisy image at each denoising step, and encourage it to better encapsulate the semantic information from text. However, they necessitate meticulously calibrated optimization

parameters. As shown in Fig. 5, different optimization parameter settings directly affect the final generated image. Under-optimization leads to subpar quality, while over-optimization skews the noisy image away from the distribution of SD, resulting in an out-of-domain image. Moreover, even with the same text prompt, optimal parameters vary with noise, making case-specific tuning a laborious task.

We bypass this challenge by prioritizing noise optimization in the initial latent space, ensuring sufficient and appropriate adjustment. Specifically, INITNO is exclusively dedicated to navigating within the initial latent space, moving randomly sampled initial noise into the valid region. Unlike denoising processes that implicitly model the noisy image distribution, the initial latent space typically adheres to a standard Gaussian distribution, providing an avenue to restrict the optimized noise to the domain.

**Noise update strategy.** Given a noise $\epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{1})$, as opposed to the commonly used incremental update: $\epsilon' \leftarrow \epsilon + \Delta\epsilon$, we adopt a distribution optimization strategy to update the noise:

$$\mu' \leftarrow \mu + \Delta\mu, \qquad \sigma' \leftarrow \sigma + \Delta\sigma, \tag{6}$$

where $\mu$ and $\sigma$ are the mean and standard deviation of the Gaussian distribution, initialized to $\mathbf{0}$ and $\mathbf{1}$, respectively. Thus, we have: $\epsilon' \leftarrow \mu' + \sigma'\epsilon$, and $\epsilon' \sim \mathcal{N}\left(\mu', \sigma'^2\right)$.

**Loss functions.** We employ a joint loss for optimizing initial noise, comprising cross-attention response loss, self-attention conflict loss, and distribution alignment loss. Intuitively, the first two are formulated as:

$$\mathcal{L}_{\text{CrossAttn}} = \mathcal{S}_{\text{CrossAttn}}, \qquad \mathcal{L}_{\text{SelfAttn}} = \mathcal{S}_{\text{SelfAttn}}. \tag{7}$$

The distribution alignment loss utilizes Kullback-Leibler divergence [22] to constrain the consistency between the target Gaussian distribution $\mathcal{N}\left(\mu, \sigma^2\right)$ and the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$:

$$\mathcal{L}_{\text{KL}} = \mathsf{KL}\left(\mathcal{N}\left(\mu, \sigma^2\right) \| \mathcal{N}(\mathbf{0}, \mathbf{1})\right). \tag{8}$$

In summary, the joint loss is written as:

$$\mathcal{L}_{\text{joint}} = \lambda_1 \mathcal{L}_{\text{CrossAttn}} + \lambda_2 \mathcal{L}_{\text{SelfAttn}} + \lambda_3 \mathcal{L}_{\text{KL}}, \tag{9}$$

where we empirically set $\lambda 1 = 1$, $\lambda_2 = 1$, and $\lambda_3 = 500$.

**Noise optimization procedure.** The outlined noise optimization procedure is encapsulated in Algorithm 1. Sampling the initial noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, we initialize the learnable parameters $\mu$ to $\mathbf{0}$ and $\sigma$ to $\mathbf{1}$. We optimize these parameters using the joint loss, ensuring that the optimized noise $\mu + \sigma\mathbf{z}^T$ resides within the valid initial latent space. The optimization employs the Adam optimizer with a learning rate of $1 \times 10^{-2}$.

To mitigate computational inefficiency due to excessively challenging samples, we impose a maximum limit on optimization iterations, denoted as $\tau_{\text{MaxStep}}$. If we fail to reach the valid region after $\tau_{\text{MaxStep}}$ updates, we resample $\mathbf{z}_T$, reset $\mu$ and $\sigma$, and commence a new round of the optimization process. We also define a maximum limit on optimization rounds, denoted as $\tau_{\text{MaxRound}}$. In our experiments, we set $\tau_{\text{MaxStep}}$ and $\tau_{\text{MaxRound}}$ as 50 and 5, respectively. For a minority of intricate texts where effective noise cannot be found within affordable optimization rounds, we opt for the noise that minimizes the sum of $\mathcal{S}_{\text{CrossAttn}}$ and $\mathcal{S}_{\text{SelfAttn}}$ from the existing noise pool $\mathcal{P}$.

**Algorithm 1: INITNO**

**Input:** A pre-trained T2I diffusion model SD$(\cdot)$, a text prompt $\mathbf{y}$, a noise $\mathbf{z}_T$, threshold: $\tau_c$, $\tau_s$, $\tau_{\text{MaxStep}}$, and $\tau_{\text{MaxRound}}$.
**Output:** Generated image $\mathbf{x}$.

Noise pool $\mathcal{P} \leftarrow \{\}$
**for** $i = 1$ to $\tau_{\text{MaxRound}}$ **do**
  Initialize $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, $\mu \leftarrow \mathbf{0}$, $\sigma \leftarrow \mathbf{1}$
  **for** $j = 1$ to $\tau_{\text{MaxStep}}$ **do**
    $\_, \mathbf{A}^c, \mathbf{A}^s \leftarrow \text{SD}\left(\mu + \sigma\mathbf{z}^T, \mathbf{y}\right)$
    Calculate $\mathcal{S}_{\text{CrossAttn}}$ (Eq. 3), $\mathcal{S}_{\text{SelfAttn}}$ (Eq. 5)
    **if** $\mathcal{S}_{\text{CrossAttn}} < \tau_c$ and $\mathcal{S}_{\text{SelfAttn}} < \tau_s$ **then**
      $\hat{\mathbf{z}}^T \leftarrow \mu + \sigma\mathbf{z}^T$
      Go to step 15 ▷ valid noise
    **else**
      Calculate $\mathcal{L}_{\text{joint}}$ (Eq. 9)
      $\mu, \sigma \leftarrow \text{Adam}(\mu, \sigma, \mathcal{L}_{\text{joint}})$
  Add $\mu + \sigma\mathbf{z}^T$ to $\mathcal{P}$
$\hat{\mathbf{z}}^T \leftarrow \arg\min_{\mathbf{z}^T \in \mathcal{P}} \mathcal{S}_{\text{CrossAttn}} + \mathcal{S}_{\text{SelfAttn}}$
$\mathbf{x}, \_, \_ \leftarrow \text{SD}\left(\hat{\mathbf{z}}^T, \mathbf{y}\right)$ or $\text{AE}\left(\hat{\mathbf{z}}^T, \mathbf{y}\right)$
**return** $\mathbf{x}$

## 5. Experiments

### 5.1. Experimental settings

**Implementation Details.** Following [6], we utilize the official Stable Diffusion v1.4 text-to-image model. A fixed guidance level of 7.5 is employed. We apply a Gaussian filter with a kernel size of 3 and a standard deviation of 0.5 to smooth the cross-attention map $\mathbf{A}^c$ and the self-attention map $\mathbf{A}^s$. For the denoising process, $T$ is set to 50. The target token identification process can be performed either manually or automatically with the assistance of GPT [3].

**Datasets.** We assess our method using the Animal-Animal, Animal-Object, and Object-Object datasets [6], which comprise two subjects and selectively assign a color to each subject. It is noteworthy that our method is not constrained to this scenario, and it can be applied to complex prompts containing any number or type of subjects and attributes, as

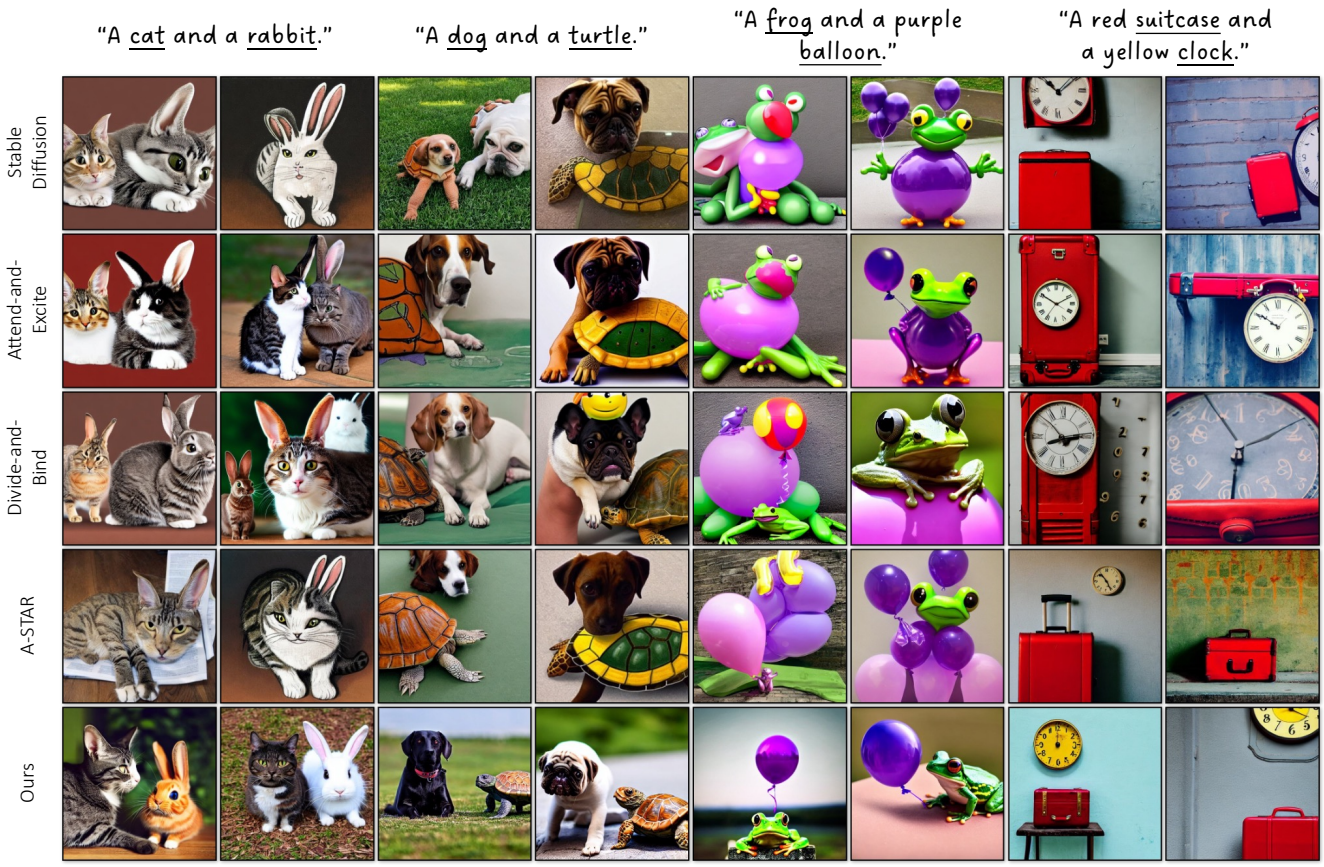


Figure 6. **Qualitative comparison.** Each image is generated with the same text prompt and random seed for all methods. The subject tokens are highlighted in underline. Our method shows excellent alignment with text prompts while maintaining a high level of realism.

discussed in Sec. 5.6 and supplementary material.

### 5.2. Qualitative comparison

Fig. 6 presents a comparative analysis of our results against state-of-the-art counterparts, under the same text prompts and random seeds. It can be seen that existing methods commonly suffer from serious problems of subject mixing. For instance, in the case of *a cat and a rabbit*, conventional methods frequently yield erroneous outputs characterized by the fusion of cat faces and rabbit ears. Leveraging the self-attention conflict mechanism, our method can effectively decouple the two concepts of *cat* and *rabbit* and render them realistically and separately. Concurrently, Stable Diffusion [31], Composable Diffusion [25], and Structure Diffusion [12] suffer from the issue of subject neglect. Attend-and-Excite [6], Divide-and-Bind [24], and A-STAR [1] adhere to the pattern of adjusting noisy images at each denoising step. Given the same text prompt, using the same adjustment strategy, the quality of results generated by sampling different noises vary significantly. For instance, for the prompt *a dog and a turtle*, only the first example produced by Divide-and-Bind [24] is successful. In contrast, our method exhibits better generalization and adapts to diverse prompts. Furthermore, our approach also demonstrates more faithful binding of properties, *e.g.*, *yellow clock*.

### 5.3. Quantitative comparison

**Objective Evaluation.** Following to the protocol in [6], we quantitatively evaluate the proposed method using CLIP Image-Text Similarity and CLIP Text-Text Similarity. For each prompt, 64 images are generated, and the average CLIP cosine similarity is computed, maintaining consistent random seeds across all methods. Specifically, for CLIP Image-Text Similarity, we report full prompt similarity (cosine similarity between the full prompt and generated image) and minimum object similarity (minimum of the similarities between the generated image and each of the two subject prompts). For CLIP Text-Text Similarity, text-text similarities are computed by captioning the generated images with BLIP [23] and comparing them with the input prompt. As shown in Fig. 7, the proposed method outperforms other approaches.

**User study.** We also conduct a subjective user study involving 12 volunteers with expertise in image processing. Participants were asked to select the most visually-appealing

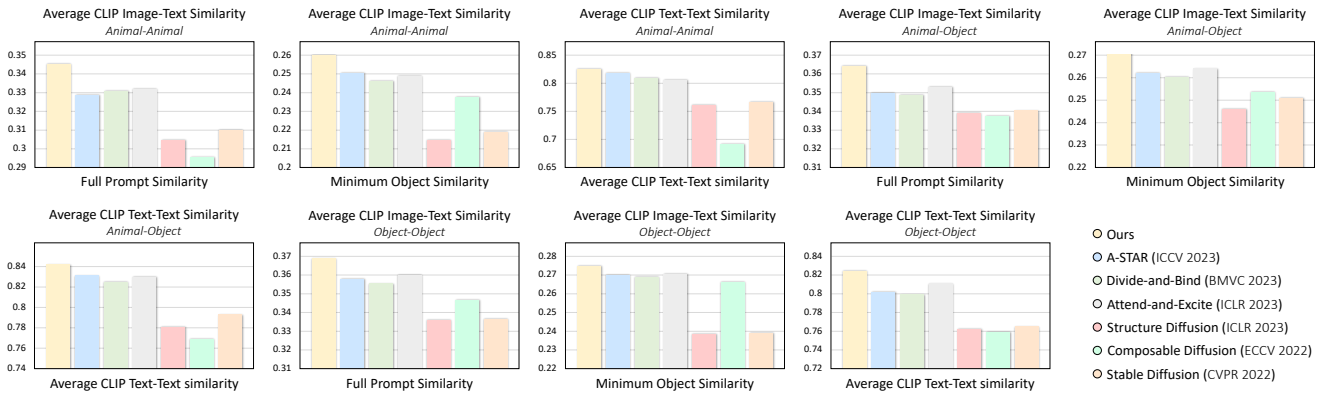


Figure 7. **Objective evaluation.** Average CLIP Image-Text Similarity, including Full Prompt Similarity and Minimum Object Similarity, and Average CLIP Text-Text Similarity are reported for the quantitative measurement. *Higher* is better.

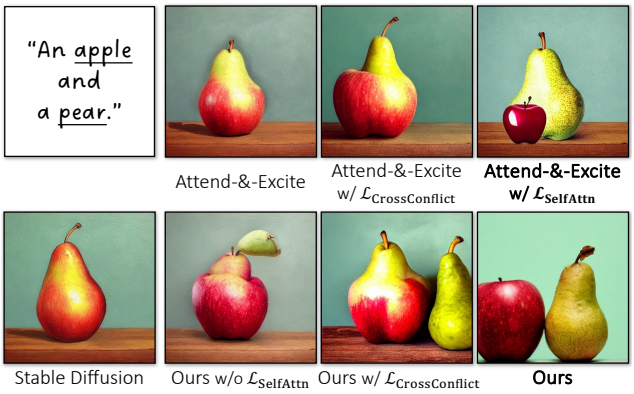


Figure 8. **Visualization of effects of $\mathcal{L}_{\text{SelfAttn}}$.**

| Method | User study |
|---|---|
| Stable Diffusion [31] | 4.17% |
| Composable Diffusion [25] | 2.50% |
| Structure Diffusion [12] | 3.33% |
| Attend-and-Excite [6] | 14.17% |
| Divide-and-Bind [24] | 6.67% |
| A-STAR [1] | 5.83% |
| INITNO (Ours) | **63.33%** |

Table 1. **User study.** INITNO performs over other counterparts.

and semantically-faithful image from those generated by our proposed method and the state-of-the-art approaches. Specifically, each participant has 20 questions. We tally the votes and show the statistics in Table 1. Our method performs favorably against the other methods.

**Inference time.** Evaluated on a single Tesla V100 (32GB), 100 images with a resolution of 512×512 pixels are randomly generated, SD synthesizes an image in an average of 8.34 seconds, while our method takes 18.93 seconds.

## 5.4. Ablation study

**On self-attention conflict loss.** The self-attention conflict loss is designed to address the subject mixing issue arising from self-attention map overlap, and it can be effortlessly incorporated into existing methods. As depicted in Fig. 8, the self-attention conflict loss effectively addresses the mixing issue between *apple* and *pear*. We also experiment with a cross-attention conflict loss [1], which replaces the self-attention map with cross-attention ones, but it did not produce satisfactory results. We surmise this is due to the inaccurate text embedding extracted by CLIP. Conversely, the self-attention map offers a more intuitive solution.

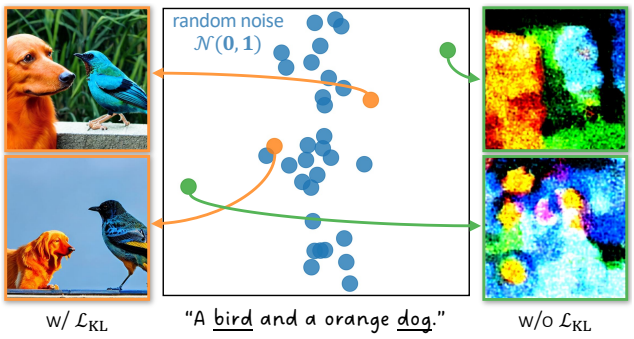


Figure 9. **Visualization of effects of $\mathcal{L}_{\text{KL}}$.**

**On distribution alignment loss.** The distribution alignment loss is introduced to ensure that the optimized noise $\hat{\mathbf{z}}_T$ adheres to the standard Gaussian distribution. As visualized in the t-SNE plot of initial noise in Fig. 9, without constraints, the optimized noise may deviate from the standard Gaussian distribution, resulting in distorted results.

## 5.5. Grounded Text-to-Image

INITNO is a plug-and-play approach that can be effortlessly integrated with existing diffusion models to enable training-free controllable generation, such as layout-to-image, mask-to-image generation, etc. As demonstrated in Fig. 11, we incorporate INITNO into BoxDiff [38], where modifications to the intermediate noisy image are transferred to the initial latent space, resulting in the synthesis


Figure 10. **Qualitative comparison with complex text prompts.** Each image is generated with the same text prompt and random seed for all methods. The subject tokens are highlighted in underline.


Figure 11. **Grounded Text-to-Image.** Due to sufficient adjustments, INITNO generates more accurately location-aware *cat*.

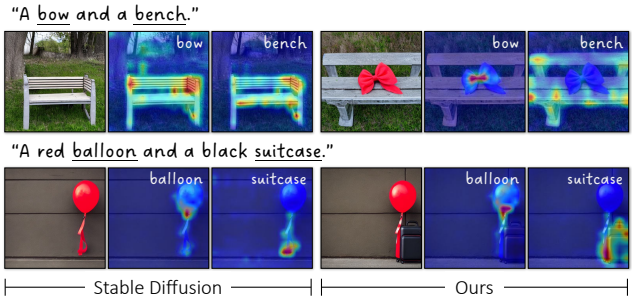


Figure 12. **Visualization of the final cross-attention maps.**

of faithful and location-aware images.

### 5.6. More results

**Complex prompts.** Fig. 10 presents examples of complex prompts, inclusive of prompts with three or more subjects and intricate attributes. The synthesis of high-fidelity images is demonstrated, underscoring the capacity of our method to manage complex semantic scenes.

**Visualization of the attention maps.** Following [6], we also provide a visualization of the final cross-attention map for each subject token after denoising process in Fig. 12. Our method achieves reasonable attention allocation, leading to semantically-consistent results.

## 6. Conclusion

The core of our methed is the Initial Noise Optimization (INITNO), which consists of the initial latent space partitioning and the noise optimization pipeline, responsible for defining valid regions and steering noise navigation, respectively. Specifically, we investigate the utility of attention maps in the diffusion model, harnessing them to formulate the cross-attention response score and the self-attention conflict score, both instrumental to the partitioning of the initial latent space. Moreover, we bypass the complex balance between under-optimization and over-optimization of noise via a meticulously crafted pipeline that introduces a novel distribution alignment loss. We further demonstrate the superiority of our approach over existing state-of-the-art methods in generating semantically-faithful images, providing a versatile plug-and-play solution that effectively integrates into existing diffusion models for training-free controllable generation

## Acknowledgment

This work is partly supported by the National Key R&D Program of China (2022ZD0161902), the National Natural Science Foundation of China (62022011, 62202031), the Research Program of State Key Laboratory of Software Development Environment (SKLSDE-2023ZX-14), and the Fundamental Research Funds for the Central Universities. We also give specical thanks to Alibaba Group for their contribution to this paper.

## 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. In *ICCV*, 2023. 2, 6, 7

[2] 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. 1, 2

[3] 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. In *NeurIPS*, 2020. 5

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 3

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[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. In *SIGGRAPH*, 2023. 2, 3, 4, 5, 6, 7, 8

[7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 1

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 1, 2

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021. 2

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 1

[11] Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, et al. Dreamoving: A human dance video generation framework based on diffusion models. *arXiv preprint arXiv:2312.05107*, 2023. 1

[12] 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. 2, 6, 7

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[14] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Josh Susskind, and Navdeep Jaitly. Matryoshka diffusion models. *arXiv preprint arXiv:2310.15111*, 2023. 1, 2

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2, 3

[17] 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. 1, 2

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[19] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020.

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.

[21] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 1

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1, 5

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 6

[24] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. In *BMVC*, 2023. 2, 6, 7

[25] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 2, 6, 7

[26] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *SIGGRAPH*, 2023. 3

[27] 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 *ICML*, 2021. 3

[28] 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. 2

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[30] 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, 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 6, 7

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1, 2

[33] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*, 2023. 2

[34] 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 *CVPR*, 2022. 2

[35] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 3

[36] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *NeurIPS*, 2016. 1

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[38] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, 2023. 7

[39] 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 *CVPR*, 2018. 2

[40] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. In *NeurIPS*, 2023. 1, 2

[41] 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. *TMLR*, 2022. 1, 2

[42] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 2

[43] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, 2021.

[44] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, 2019. 2