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.

Not All Parameters Matter: Masking Diffusion Models for Enhancing Generation Ability

Lei Wang[1], Senmao Li[1], Fei Yang[1†], Jianye Wang[1], Ziheng Zhang[1]
Yuhan Liu[1], Yaxing Wang[1,2], Jian Yang[1†]
[1]PCA Lab, VCIP, College of Computer Science, Nankai University [2] Shenzhen Futian, NKIARI
{scitop1998, senmaonk, feiyangflyhigher}@gmail.com, {yaxing,csjyang}@nankai.edu.cn

Abstract

The diffusion models, in early stages focus on constructing basic image structures, while the refined details, including local features and textures, are generated in later stages. Thus the same network layers are forced to learn both structural and textural information simultaneously, significantly differing from the traditional deep learning architectures (e.g., ResNet or GANs) which captures or generates the image semantic information at different layers. This difference inspires us to explore the time-wise diffusion models. We initially investigate the key contributions of the U-Net parameters to the denoising process and identify that properly zeroing out certain parameters (including large parameters) contributes to denoising, substantially improving the generation quality on the fly. Capitalizing on this discovery, we propose a simple yet effective method—termed "MaskUNet"— that enhances generation quality with negligible parameter numbers. Our method fully leverages timestep- and sample-dependent effective U-Net parameters. To optimize MaskUNet, we offer two fine-tuning strategies: a training-based approach and a training-free approach, including tailored networks and optimization functions. In zero-shot inference on the COCO dataset, MaskUNet achieves the best FID score and further demonstrates its effectiveness in downstream task evaluations. Project page: https://gudaochangsheng.github.io/MaskUnet-Page/

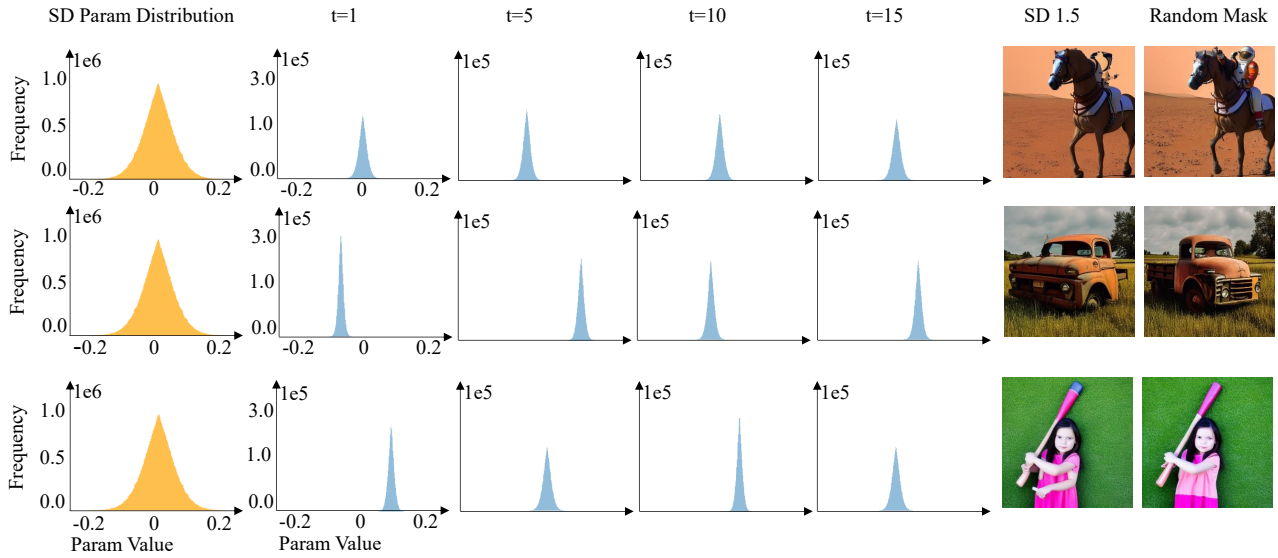
1. Introduction

Diffusion models [20, 54], a class of generative models based on iterative denoising processes, have recently gained significant attention as powerful tools for generating high-quality images, videos, and 3D data representations. Text-to-image models, such as stable diffusion (SD) [47], have successfully applied pre-trained U-Net models to downstream tasks, including personalized text-to-image generation [33, 48], relation inversion [25], semantic binding [2, 12, 24, 46], and controllable generation [7, 41, 64, 68]. The diffusion models, in the early denoising stage, establish spatial information representing semantic structure, and then widen to the regional details of the elements in the later stage [5, 13]. Therefore, at different inference steps, the diffusion models use the same network paramaters (e.g., a U-Net in SD) to forcibly learn different information: *the global structure and characteristics, and edges and textures etc.*.

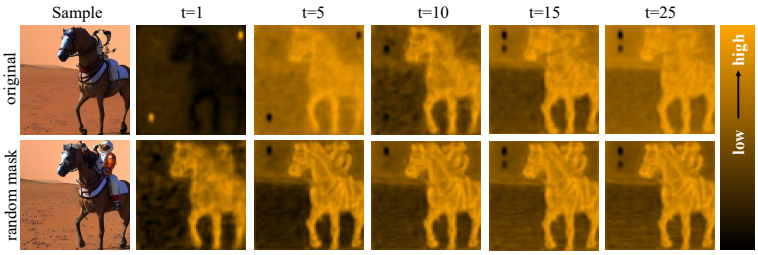
However, the traditional classification models [17, 23, 53, 58], such as ResNet [17], they capture the image information (e.g., the structure and semantic features) at different layers. Typically, the shallow layers focus on extracting the structure information, while the deeper layers capture higher-level semantic information [30, 51, 67]. Similarly, in traditional generative models [27, 28], the first few layers of the generator control the synthesis of structural information, while the deeper layers represent texture and edge details. Both classical classification and generative tasks leverage distinct model parts to represent the internal properties of sample, reducing the difficulty of network optimization and enhancing its representational capacity. Distinct from above two classes, the diffusion models use the same parameters to forcibly learn different information when generating a sample. However, to our best knowledge this difference of the diffusion U-Net remains largely underexplored.

Beyond the application of diffusion models, in this paper, we are interested in investigating the effectiveness of the pretrained U-Net parameters for the denoising process. To better understand the denoising process, we first present a empirical analysis using a random mask at inference time to examine the generation process of diffusion models, an area that has received limited prior investigation. As illustrated in Figure 1 (c), we multiply the pre-trained U-Net weights by a random binary UNet-like mask at inference time, ensuring that we have different networks at every time step. This aims to keep the consistency with the traditional network design that the vary semantic features are modeled at different lay-

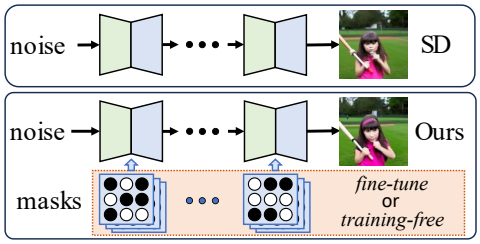
†Corresponding authors.



(a) Analysis of parameter distributions and denoising effects across different time steps for Stable Diffusion (SD) 1.5 with and without random masking. The first column shows the parameter distribution of SD 1.5, while the second to fifth columns display the distributions of parameters removed by the random mask. The last two columns compare the generated samples from SD 1.5 and the random mask.



(b) Comparison of original and random mask results in the denoising process.



(c) Illustration of our method.

Figure 1. The motivation of our method.

ers. As shown in Figure 1 (a) (the second and last columns), using certain random masks enhances the denoising capability of the U-Net architecture, thereby contributing to a superior output in terms of both fidelity and detail preservation. Further, we also visualize the corresponding features at different timesteps (see Figure 1 (b)). Compared to the original SD features, the masked backbone features obtain more details and structure information, improving the denoising capability. This results indicate that the generated samples benefit from distinct U-Net weight configurations.

Based on the above findings, we are interested to select desire parameters of the diffusion models which hold the potential to improve sample quality. To achieve this goal, we need to learn a desire binary mask, which zeros out the useless parameters, and retains the desire ones. Naively using a random mask fails to guarantee a good generation result, since the desire mask is related to the denoised sample. As illustrated in Figure 1 (from third to sixth columns), a vertical examination of each sample reveals that the desire weights differs across samples, indicating that we need a tailored mask to synthesize a high-quality sample. This insight motivates us to introduce sample dependency in mask generation, allowing the model to better adapt to each prompt's specific needs.

In this paper, we propel forward with the introduction of a novel strategy, called *MaskUNet*, which improves the inherent capability of text-to-image generation without updating any parameters of the pre-trained U-Net. Specifically, as shown in Figure 1 (c), we introduce a strategy that uses a learnable binary mask to sample parameters from the pre-trained U-Net, thereby obtaining a timestep-dependent and sample-dependent U-Net that emphasizes the importance of parameters sensitive to generation. To efficiently learn the mask, we design two fine-tuning strategies: a training-based approach and a training-free approach. In the training-based approach, a parameter sampler produces timestep-dependent and sample-dependent masks, supervised by diffusion loss. The parameter sampler is implemented with an MLP, whose parameter count is negligible compared to the pre-trained U-Net. The training-free approach, on the other hand, generates masks directly under the supervision of a reward model [61, 62], eliminating the need for a mask generator compared to the training-based approach.

Compared with existing fine-tuning methods, MaskUNet

aims to tap into the inherent potential of the model, achieving improvements in zero-shot inference accuracy on the COCO 2014 [34] and COCO 2017 [34] datasets. We further applied MaskUNet to downstream tasks, including image customization [10, 48], video generation [29], relation inversion [25], and semantic binding [2, 46], to verify its effectiveness. The main contributions of this paper can be summarized as follows:

- We conduct an in-depth study of the relationship between parameters in the pre-trained U-Net, samples, and timesteps, revealing the effectiveness of parameter independence, which provides a new perspective for efficient utilization of U-Net parameters.
- We propose a novel fine-tuning framework for text-to-image pre-trained diffusion models, called MaskUNet. In this framework, the training-based method optimizes masks through diffusion loss, while the training-free method uses a reward model to optimize masks. The learnable masks enhance U-Net's capabilities while preserving model generalization.
- We evaluate MaskUNet on the COCO dataset and various downstream tasks. Experimental results demonstrate significant improvements in sample quality and substantial performance gains in key metrics.

2. Related Work

2.1. Diffusion Models

Diffusion models [20, 54, 56, 57] have achieved remarkable success in the field of image generation, but direct computation in pixel space is inefficient. To address this, Latent Diffusion Model (LDM) [47] introduces Variational Autoencoders (VAE) to compress images into latent space. Additionally, to tackle iterative denoising during inference, some works have proposed samplers that require fewer steps [37, 55, 65], while others have utilized knowledge distillation to reduce sampling steps [4, 6, 38, 42]. Furthermore, some methods employ structured pruning to accelerate inference [32, 40]. With the emergence of large-scale image-text datasets [49, 50] and visual language models [26, 44], text-to-image generation networks represented by stable diffusion (SD) have found widespread applications, supporting various tasks such as controllable image generation [41, 64], controllable video generation [7, 68], and image customization [31, 33, 48].

2.2. Training-based Models

Training-based models enhance the U-Net by updating model parameters, typically using the following strategies: introducing trainable modules at specific layers to adapt pretrained weights to new tasks [15, 41, 45, 63], selectively fine-tuning a subset of existing parameters [14, 22], or directly updating all parameters. However, these approaches carry a risk of overfitting. Recently, methods like LoRA [21] and DoRA [35] have been proposed, which inject low-rank matrices into pretrained weights to increase model flexibility and mitigate overfitting. However, these methods still adjust the original parameter space, potentially affecting the generalization of the pretrained model. In contrast, our proposed MaskUNet preserves the generalization capacity of the pretrained model by avoiding any updates to the U-Net parameters.

2.3. Training-free Models

Training-free models designed to enhance the generative capability of U-Net can be broadly categorized into three main approaches. The first approach focuses on adjusting feature scales [16, 39, 52]. For instance, FreeU [52] introduces sample-dependent scaling factors for U-Net features and suppresses skip connection features to redistribute feature weights, thereby improving generation quality. The second approach emphasizes optimizing latent codes by leveraging various supervisory methods, such as attention maps [1, 2, 46, 66], noise inversion [43], or reward models [8], to strengthen U-Net's generative performance. The third approach centers on optimizing text embeddings [3, 9, 59]. For example, Chen *et al*. [3] employ balanced text embedding loss to eliminate potential issues within key token embeddings, thus improving generation quality. Unlike these methods, MaskUNet uses a reward model for mask supervision to dynamically select effective U-Net parameters, enhancing its performance.

3. Proposed Method

The diffusion models use the same parameters to forcibly learn different information when synthesizing a sample, limiting its generation adaptability. In this paper, we aim to learn a timestep-dependent and sample-dependent mask generation model, which further select the target parameters from the pretrained U-Net, enhancing the pretrained U-Net in the diffusion model. This section first provides an overview of the diffusion model as our foundation (Sec. 3.1), followed by methods for enhancing U-Net through fine-tuning with training (Sec. 3.2) and training-free (Sec. 3.3) approaches.

3.1. Preliminary

Diffusion models add noise to data, creating a Markov chain that approximates a simple prior (usually Gaussian) [20]. A neural network is trained to reverse this process, starting from noise and progressively denoising to recover the original data, learning to extract useful information at each step.

In the Latent Diffusion Model (LDM) [47], the diffusion process occurs in a lower-dimensional latent space instead of pixel space, offering significantly improved computational

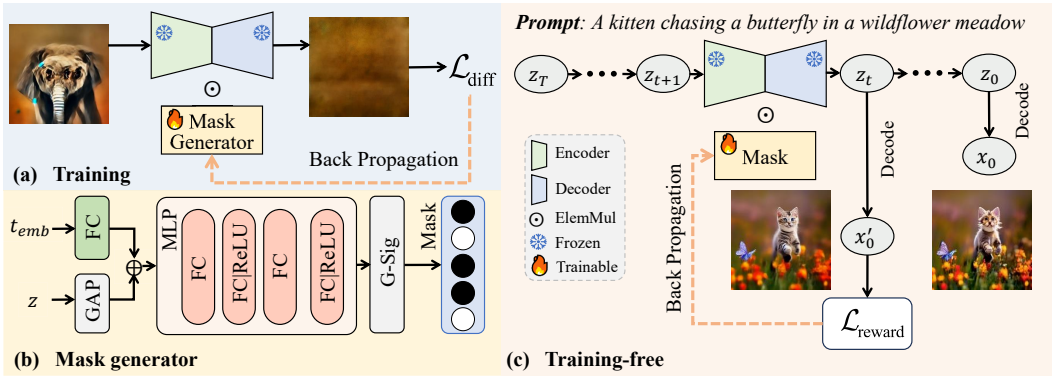


Figure 2. The pipeline of the MaskUnet. G-Sig represents the Gumbel-Sigmoid activate function. GAP is global average pooling.

efficiency. The training objective of LDM can be formulated as minimizing the following loss function:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\varepsilon,z,t}\left[\text{MSE}\left(\varepsilon_\theta\left(z_t,t\right),\varepsilon\right)\right], \tag{1}$$

where $\text{MSE}(\cdot)$ denotes the mean squared error, $\varepsilon \sim \mathcal{N}(0,I)$ is noise from a standard Gaussian distribution, z_t is the noisy latent variable at time step t, and $\varepsilon_\theta(z_t,t)$ is the noise predicted by the denoising network, parameterized by θ.

In text-to-image diffusion, an additional prompt c guides image generation for more controllable outputs, so the training objective is:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\varepsilon,z,t,c}\left[\text{MSE}\left(\varepsilon_\theta\left(z_t,t,c\right),\varepsilon\right)\right]. \tag{2}$$

3.2. Training with Learnable Masks

To exploit the full potential of the model's parameters, we introduce a learnable mask to sample weights from the pretrained U-Net. We propose a training-based fine-tuning approach to optimize the mask, as shown in Figure 2(a). The mask is trained using the diffusion loss defined in Equ. (2). And a mask is generated by a mask generator, as shown in Figure 2(b). Let's define the flattened input feature map $h \in \mathbb{R}^{B\times N\times C_{in}}$, where B represents the batch size, N is the number of patches and C_{in} represents the number of input channels. The mask generator takes as input both the timestep embedding $t_{\text{emb}} \in \mathbb{R}^{B\times C_1}$ and the latent codes $z \in \mathbb{R}^{B\times C\times H\times W}$, where H and W represent the height and width of z.

We first merge t_{emb} and z as follows:

$$z' = \text{FC}(t_{\text{emb}}) + \text{GAP}(z), \tag{3}$$

where $z' \in \mathbb{R}^{B\times C}$ is the merged output, $\text{FC}(\cdot)$ is the fully connected layer, and $\text{GAP}(\cdot)$ is global average pooling. We then apply a 4-layer MLP with 2 ReLU activations to introduce non-linearity:

$$\hat{z} = \text{MLP}(z'), \tag{4}$$

where $\hat{z} \in \mathbb{R}^{B\times C_2}$ is the MLP output.

To sample the weights, we treat $\hat{z}$ as a binary mask:

$$m = \sigma\left(\hat{S};\tau,\delta\right), \tag{5}$$

where $m \in \mathbb{R}^{B\times C_2}$ is the output of the Gumbel-Sigmoid [11] activation function $\sigma(\cdot;\tau,\delta)$. The temperature coefficient $\tau \in (0,\infty)$ controls the discreteness of m: as $\tau \to 0$, m tends to a binary distribution; as $\tau \to \infty$, m tends to a uniform distribution. The threshold δ is used to discretize the probability distribution.

Next, we apply the reshaped $m' \in \mathbb{R}^{B\times C_{\text{out}}\times C_{\text{in}}}$ to the U-Net's linear layer weight $w \in \mathbb{R}^{C_{\text{out}}\times C_{\text{in}}}$ to obtain the masked weight:

$$\hat{w} = m' \odot w, \tag{6}$$

where $\hat{w} \in \mathbb{R}^{B\times C_{\text{out}}\times C_{\text{in}}}$ is the masked weight, and $\odot$ denotes element-wise multiplication.

Finally, the input h and weight $\hat{w}$ are calculated to obtain the output features,

$$o = \text{BMM}\left(z,\hat{w}\right), \tag{7}$$

where $o \in \mathbb{R}^{B\times N\times C_{out}}$, $\text{BMM}(\cdot,\cdot)$ represents the batch matrix-matrix multiplication.

By introducing a mask generator, we expect the pretrained U-Net weights to dynamically adapt to different sample features and timestep embeddings. Notably, this design does not update the U-Net's parameters; instead, it leverages sample- and timestep-dependent adjustments, allowing the model to selectively activate specific U-Net weights tailored to each input. This approach enhances the flexibility of the pretrained U-Net while preserving the stability of the pretrained structure.

3.3. Training-Free with Learnable Masks

To further demonstrate the effectiveness of the mask, inspired by ReNO [8], we propose a training-free algorithm to guide the optimization of the mask.

As shown in Figure 2(c), given the intermediate state z_t of the denoising process, which is obtained by denoising the

Algorithm 1 Training-free based Fine-tuning

Require prompt c, a pretrained unet ε_θ, reward models $\sum_{i=1}^{n} \Psi_i$, balance factor of reward models ω_i, optimize the number of iterations λ, mask logits l, temperature factor τ, threshold δ, maximum time step T
Initialize m=1.0, τ=1.0, δ=0.5, $x_T \sim \mathcal{N}(0, \mathbf{I})$
for $t = T$ **to** 0 **do**
  for $k = 0$ **to** λ **do**
    Get binary mask ${m'}_t^k \leftarrow \sigma\left(l_t^k; \tau, \delta\right)$
    Apply to pre-training unet $g_{\theta'} : \theta' \leftarrow \theta \odot {m'}_t^k$
    Predict noisy latent $z_{t-1} \leftarrow \varepsilon_{\theta'}\left(z_t, t, c\right)$
    Predict the original latent $z_0 \leftarrow z_{t-1}$
    Decode to image space $x_t^k \leftarrow z_0$
    Reward loss $\mathcal{L}_{\text{reward}} \leftarrow \sum_{i=1}^{n} \omega_i \Psi_i\left(x_t^k, c\right)$
    Update mask logits $l_t^{k+1} \leftarrow l_t^k$
  end for
end for
return x_0^λ

representation z_{t+1} in the previous step and guided by the prompt c, it can be expressed as:

$$z_t = \varepsilon_\theta\left(z_{t+1}, t+1, c\right), \tag{8}$$

where $\varepsilon_\theta(\cdot,\cdot,\cdot)$ is the pretrained U-Net, and θ represents its parameters. Similar to the training-based approach, we introduce the mask m to apply to parameter θ, *i.e.*, $\theta' \leftarrow \theta \odot m$. The key difference is that m does not rely on the generator. Therefore, Equ. (8) can be rewritten as:

$$z_t = \varepsilon_{\theta'}\left(z_{t+1}, t+1, c\right). \tag{9}$$

Next, z_t is decoded into pixel space through the VAE to obtain ${x_0}'$. Using ${x_0}'$ and the prompt c, it is fed into the reward model to calculate the loss. The reward loss is then backpropagated to update the mask parameters, improving the consistency between the generated image and the prompt. The reward loss can be formulated as:

$$\mathcal{L}_{\text{reward}} = \sum_{i=1}^{n} \omega_i \Psi_i\left({x_0}', c\right), \tag{10}$$

where $\Psi_i(\cdot,\cdot)$ denotes the pre-trained reward model, and ω_i is the balancing factor. In this work, we set $n = 2$, n is the number of reward models. We use ImageReward [62] and HPSv2 [62] as the reward models. Please check the full details in Algorithm 1.

4. Experiments

4.1. Experiment Setting

Datasets and Metrics. (1) *Training-based approach.* For zero-shot text-to-image generation, we fine-tune the MaskUNet on a subset of the Laion-art (a subset of Laion-5B [50]), which contains 20.1k pairs of image and text. To verify the effectiveness of our method, we generated 30k images for COCO 2014 [34], 5k images for COCO 2017 [34] respectively. We evaluate the image quality using Fréchet Inception Distance (FID) [18] and the alignment of the image text using CLIP score [44]. (2) *Training-free approach.* We evaluated the effectiveness of MaskUNet on two semantic binding datasets T2I-CompBench [24] and GenEval [12]. We use BLIP-VQA score [24] for the evaluation of attribute correspondences and GENEVAL score for the image correctness.

Baselines. (1) *Training-based approach.* We choose SD 1.5 [47], Full Fine-tune and LoRA [21] as baselines to compare with MaskUnet. We also apply MaskUNet to downstream tasks such as image customization, relation inversion, and text-to-video generation. For these tasks, we use Dreambooth [48], Textual Inversion [10], ReVersion [25] and Text2Video-zero [29] as baselines. (2) *Training-free approach.* We select SD 1.5 [47], SD 2.0 [47], SynGen [46] and Attend-and-excite [2] as baselines for comparison with MaskUNet.

Implementation Details. (1) *Training-based approach.* In our implementation, the learning rate (LR) is set to $1e$-5, and AdamW [36] with a weight decay of $1e$-2 is used as the optimizer. The training process consists of 12 epochs, with 50 inference steps. The classifier-free guidance (CFG) [19] is set to 7.5, and DDIM [65] is employed as the sampler. (2) *Training-free approach.* The number of iterations λ is set to 15. The optimizer used is AdamW [36], with an LR of $1e$-2. We utilize ImageReward [62] and HPSV2 [61] as reward models, with equilibrium coefficients set to 1.0 and 5.0, respectively. The number of inference steps is set to 15, with the CFG [19] set to 7.5. The sampler uses DPM-Solver [37].

Table 1. Quantitative results of zero-shot generation on the COCO 2014 and COCO 2017 datasets, with the best results in **bold**.

Method	COCO 2014		COCO 2017	
	FID-30k (↓)	CLIP (↑)	FID-5k (↓)	CLIP (↑)
SD 1.5 [47]	12.85	0.32	23.39	0.33
Full Fine-tune	14.06	0.32	24.45	0.33
LoRA [21]	12.82	0.32	23.18	0.33
MaskUnet	**11.72**	0.32	**21.88**	0.33

4.2. Training-based Text-to-image Generation

4.2.1. Zero-shot Text-to-image Generation

Table 1 presents the zero-shot generation performance of our method and baselines on the COCO 2014 and COCO 2017 datasets. For COCO 2014, MaskUNet improves the FID by 1.13 compared to SD 1.5 [47] and by 1.10 over LoRA [21]. In contrast, Full Fine-tune shows an increased FID value by 1.21 compared to SD v1.5, indicating a risk of overfitting. A similar trend is observed on the COCO 2017 dataset. In



Figure 3. Quality results compared to other methods.

summary, by leveraging the dynamic masking mechanism, MaskUNet effectively enhances the generative performance of the SD [47] model.

Figure 3 (left) presents the generative results of different methods for various prompts. In the first row, MaskUNet generates realistic and well-aligned images, while other methods either introduce artifacts (e.g., LoRA [21]) or display unnecessary background elements due to overfitting (e.g., Full Fine-tune). In the third row, MaskUNet accurately captures the mouse's attire and pose, a level of consistency that other methods struggle to achieve. Overall, MaskUNet effectively balances image quality and fidelity to prompts, capturing prompt-specific details while maintaining visual coherence across diverse scenes.

4.2.2. Downstrean tasks

MaskUnet also has the potential to enhance image quality in a variety of downstream tasks, with evaluations ranging from image customization, relation inversion, and text-to-video generation tasks.

Image Customization. DreamBooth [48] is a pioneering method for image customization, which it requires full fine-tuning of the U-Net. We compared the performance of full fine-tuning (DreamBooth), LoRA [21], and MaskUNet. As shown in Figure 4, MaskUNet excels in maintaining subject consistency and background diversity, producing high-quality images across diverse prompts, while DreamBooth and LoRA exhibit overfitting. For example, with a rare prompt combination like "on the moon", DreamBooth fails to generate coherent images, and LoRA retains unwanted



Figure 4. Quality results compared to other methods.

elements from the training set, such as background details. Notably, MaskUNet achieves effective personalization without updating U-Net parameters, demonstrating the untapped potential of the pretrained U-Net.

Textual Inversion [10] learns text embeddings to capture new concepts and is further enhanced with the introduction of MaskUNet. As shown in Figure 5, adding mask significantly improves the generation quality of Textual Inversion. For instance, the results in the first and second columns show enhanced sensitivity to quantity, while the third column better preserves subject characteristics, result-



Figure 5. Quality results by Textual Inversion [10] with or without mask.

ing in more accurate outputs.

Relation Inversion. ReVersion [25], a relationship-guided image synthesis method based on SD, can be enhanced by integrating MaskUNet. As shown in Figure 6, adding mask improves sensitivity to relational embeddings and enhances image fidelity. For instance, with the prompt "inside," ReVersion might place the rabbit on the surface of the cup or outside it, but with MaskUnet, sensitivity to the "inside" embedding is increased, resulting in images with the correct relational context. Additionally, for prompts like "cat," adding mask significantly enhances image quality.

Text-to-video Generation. Tex2Video-Zero [29] is a training-free diffusion model for text-to-video generation. By integrating our MaskUnet into Tex2Video-Zero, we can enhance the continuity and consistency of generated videos, as illustrated in Figure 7. For instance, in response to the prompt "A panda is playing guitar on Times Square," the addition of the mask enables the generation of a complete guitar. This indicates that the mask is orthogonal to Tex2Video-Zero, thereby facilitating the production of high-quality content.

4.3. Training-free Text-to-image Generation

Semantic Binding. Table 2 presents the quantitative results of MaskUNet on the T2I-Compbench benchmark. We observe that, compared to SD v1.5, MaskUNet achieves over 7% improvement across color, shape, and texture categories. When compared to SD 2.0 [47], MaskUNet slightly underperforms in color but surpasses it in the other two categories. To further verify the generalizability of MaskUNet, we applied it to SynGen [46], resulting in over 4% improvement in all three categories. Similar findings are shown in Table 3 on the GenEval benchmark, where the color attribution score in SynGen increased by 21% after applying MaskUNet. In summary, our MaskUNet demonstrates robust generalization capabilities in semantic binding tasks.

Figure 3 (right) compares samples generated by different methods to evaluate the effectiveness of MaskUnet in semantic binding tasks. In the first row, adding MaskUnet to SD 1.5 highlights the semantic information of the "book", while its addition to SynGen enhances the vase's texture from blurry to detailed. In the second row, MaskUnet improves sensitivity to quantity and shape. In the third row, it



Figure 6. Quality results by ReVersion [25] with or without mask.

enhances texture generation. In the last row, MaskUnet enables more accurate adherence to the specified color and object combination. Overall, MaskUnet significantly improves generative quality in semantic binding tasks, demonstrating higher fidelity to prompt specifications.

Table 2. Semantic binding evaluation for T2I-CompBench, with the best results in **bold**.

Method	NFE	BLIP-VQA		
		Color (↑)	Texture (↑)	Shape (↑)
SD 1.5 [47]	15	0.3750	0.4159	0.3742
SD 2.0 [47]	50	0.5056	0.4922	0.4221
SynGen [46]	15	0.6288	0.5796	0.3881
Atten-Exct [2]	50	0.6400	0.5963	0.4517
MaskUNet	15	0.4958	0.4938	0.4529
SynGen+MaskUNet	15	0.**6989**	**0.6209**	**0.4644**

Table 3. Semantic binding evaluation for GeneVal, with the best results in **bold**.

Model	SD 1.5 [47]	SynGen [46]	MaskUNet	SynGen+ MaskUNet
Overrall (↑)	0.39	0.43	0.46	**0.50**
Single (↑)	0.98	0.94	0.98	**0.10**
Two (↑)	0.26	0.39	0.42	**0.43**
Counting (↑)	0.28	0.31	0.38	**0.39**
Colors (↑)	0.74	0.80	0.82	**0.88**
Position (↑)	0.02	0.06	0.06	**0.08**
Color Attri (↑)	0.05	0.05	0.08	**0.26**

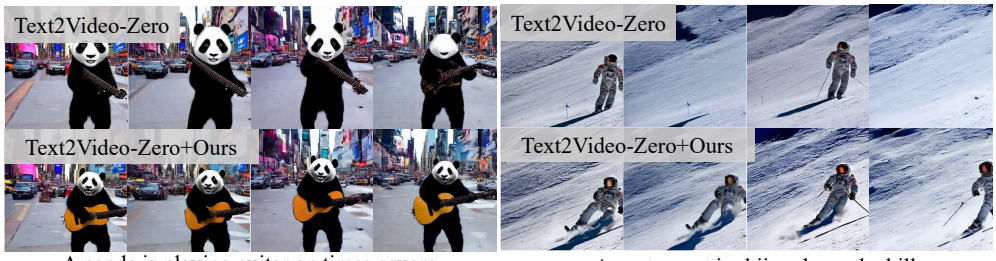


Figure 7. Quality results by Text2Video-Zero [29] with or without mask.

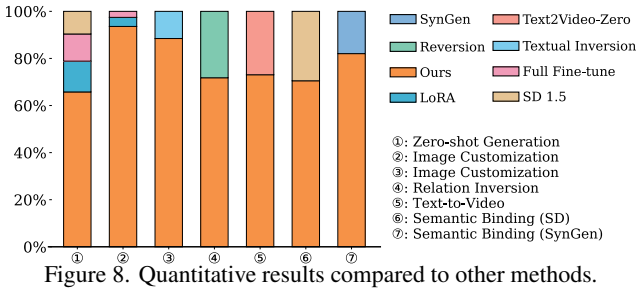


Figure 8. Quantitative results compared to other methods.

4.4. User Study

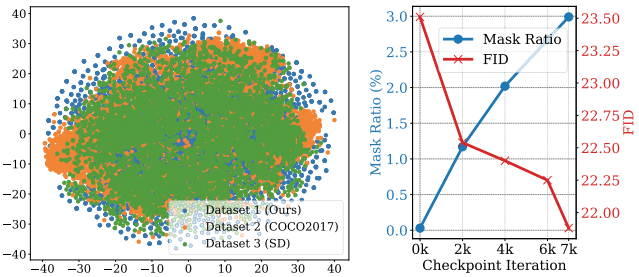
We conducted a study with 26 participants to evaluate image quality and text-image alignment, covering zero-shot generation and downstream tasks. Figure 8 presents the voting results, where the majority of votes favored our method, indicating that our approach effectively enhances the generative capability of SD.

4.5. Analysis of UNet Weight Masks

As shown in Figure 9 (a), we visualized the distributions of images generated by MaskUNet, images generated by SD, and real images from COCO 2017 using the t-SNE [60] dimensionality reduction method. It can be observed that the distribution of images generated by MaskUNet is closer to the real image distribution. Therefore, this reveals the reason why MaskUNet enhances the generalization ability of SD. Then, as shown in Figure 9 (b), we observe that as the number of iterations increases, the mask ratio shows an upward trend, while the FID gradually decreases, indicating that the mask is continuously enhancing the generative capability. It is worth noting that although the overall mask ratio remains constant, the mask locations change dynamically, resulting in a varying distribution of masked parameters (see the *supplementary material*).

4.6. Ablation Studies

For the training-based approach, Table 4 shows an ablation study on the effectiveness of different inputs to the mask generator. The MaskUNet, with both timestep embeddings and sample inputs, achieves the lowest FID score. Removing either the timestep embeddings or sample inputs results in higher FID scores, with the SD 1.5 (no mask) performing the worst. All experiments have similar CLIP scores,



(a) Image distribution (b) Mask ratio v.s. FID

Figure 9. (a) Visualization of image distributions for different methods using t-SNE. (b) Relationship between mask ratio and FID across checkpoint iterations.

indicating that the mask primarily improves image quality without significantly affecting semantic alignment.

Table 4. Ablation study on the impact of different inputs to the mask generator on COCO 2017.

Model	FID (↓)	CLIP (↑)
MaskUNet	**21.88**	**0.33**
w/o temb	22.30	0.32
w/o sample	22.14	0.32
SD 1.5	23.39	0.33

5. Conclusion

This paper proposes MaskUNet, an enhanced method for U-Net parameters in diffusion models. By utilizing learnable binary masks, MaskUNet generates time-step and sample-dependent U-Net parameters during inference. Experimental results demonstrate that MaskUNet significantly enhances the generative capability of U-Net, with improved sample quality observed in the COCO zero-shot task. Additionally, our method outperforms existing approaches in downstream tasks such as image customization, relation inversion, and text-to-video generation. To optimize computational efficiency, we also introduce a mask learning approach that requires no training, and we validate its effectiveness on two semantic binding benchmarks.

Limitations. While dynamic masking enhances model generalization, it does not enable learning of new knowledge. Future work will explore combining this approach with LoRA and extending it to other base models.

Acknowledgement. This work was supported by the National Science Fund of China under Grant Nos, 62361166670 and U24A20330, the "Science and Technology Yongjiang 20" key technology breakthrough plan project (2024Z120), the Shenzhen Science and Technology Program (JCYJ20240813114237048), and the Supercomputing Center of Nankai University (NKSC).

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023. 3

[2] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1, 3, 5, 7

[3] Chieh-Yun Chen, Chiang Tseng, Li-Wu Tsao, and Hong-Han Shuai. A cat is a cat (not a dog!): Unraveling information mix-ups in text-to-image encoders through causal analysis and embedding optimization. *Advances in Neural Information Processing Systems*, 2024. 3

[4] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-{\delta}: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. 3

[5] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 1

[6] Trung Dao, Thuan Hoang Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, and Anh Tran. Swiftbrush v2: Make your one-step diffusion model better than its teacher. In *European Conference on Computer Vision*, pages 176–192. Springer, 2025. 3

[7] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[8] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *arXiv preprint arXiv:2406.04312*, 2024. 3, 4

[9] 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 *The Eleventh International Conference on Learning Representations*, 2023. 3

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 5, 6, 7

[11] Xinwei Geng, Longyue Wang, Xing Wang, Bing Qin, Ting Liu, and Zhaopeng Tu. How does selective mechanism improve self-attention networks? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2986–2995, 2020. 4

[12] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 5

[13] Hyojun Go, Yunsung Lee, Jin-Young Kim, Seunghyun Lee, Myeongho Jeong, Hyun Seung Lee, and Seungtaek Choi. Towards practical plug-and-play diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1962–1971, 2023. 1

[14] Demi Guo, Alexander M Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*, 2020. 3

[15] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[16] Feihong He, Gang Li, Mengyuan Zhang, Leilei Yan, Lingyu Si, Fanzhang Li, and Li Shen. Freestyle: Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636*, 2024. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[21] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3, 5, 6

[22] Teng Hu, Jiangning Zhang, Ran Yi, Hongrui Huang, Yabiao Wang, and Lizhuang Ma. Sara: High-efficient diffusion model fine-tuning with progressive sparse low-rank adaptation. *arXiv preprint arXiv:2409.06633*, 2024. 3

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[24] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 1, 5

[25] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 1, 3, 5, 7

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1

[28] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 1

[29] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 3, 5, 7, 8

[30] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 1

[31] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[32] Senmao Li, Taihang Hu, Fahad Shahbaz Khan, Linxuan Li, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, and Jian Yang. Faster diffusion: Rethinking the role of unet encoder in diffusion models. *arXiv e-prints*, pages arXiv–2312, 2023. 3

[33] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 1, 3

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3, 5

[35] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 3

[36] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[37] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 3, 5

[38] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 3

[39] Jiajun Ma, Shuchen Xue, Tianyang Hu, Wenjia Wang, Zhaoqiang Liu, Zhenguo Li, Zhi-Ming Ma, and Kenji Kawaguchi. The surprising effectiveness of skip-tuning in diffusion sampling. *arXiv preprint arXiv:2402.15170*, 2024. 3

[40] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15762–15772, 2024. 3

[41] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 1, 3

[42] Thuan Hoang Nguyen and Anh Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7807–7816, 2024. 3

[43] Zipeng Qi, Lichen Bai, Haoyi Xiong, et al. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024. 3

[44] 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. 3, 5

[45] Lingmin Ran, Xiaodong Cun, Jia-Wei Liu, Rui Zhao, Song Zijie, Xintao Wang, Jussi Keppo, and Mike Zheng Shou. X-adapter: Adding universal compatibility of plugins for upgraded diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8775–8784, 2024. 3

[46] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 5, 7

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 5, 6, 7

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven gen-

eration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 3, 5, 6

[49] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3

[50] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3, 5

[51] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[52] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743, 2024. 3

[53] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[54] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1, 3

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3

[56] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3

[57] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[58] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[59] Hazarapet Tunanyan, Dejia Xu, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Multi-concept t2i-zero: Tweaking only the text embeddings and nothing else. *arXiv preprint arXiv:2310.07419*, 2023. 3

[60] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 8

[61] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2, 5

[62] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5

[63] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 3

[65] Qinsheng Zhang, Molei Tao, and Yongxin Chen. gddim: Generalized denoising diffusion implicit models. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 5

[66] Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[67] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1

[68] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 1, 3