FreeU: Free Lunch in Diffusion U-Net

Chenyang Si   Ziqi Huang   Yuming Jiang   Ziwei Liu✉

S-Lab, Nanyang Technological University

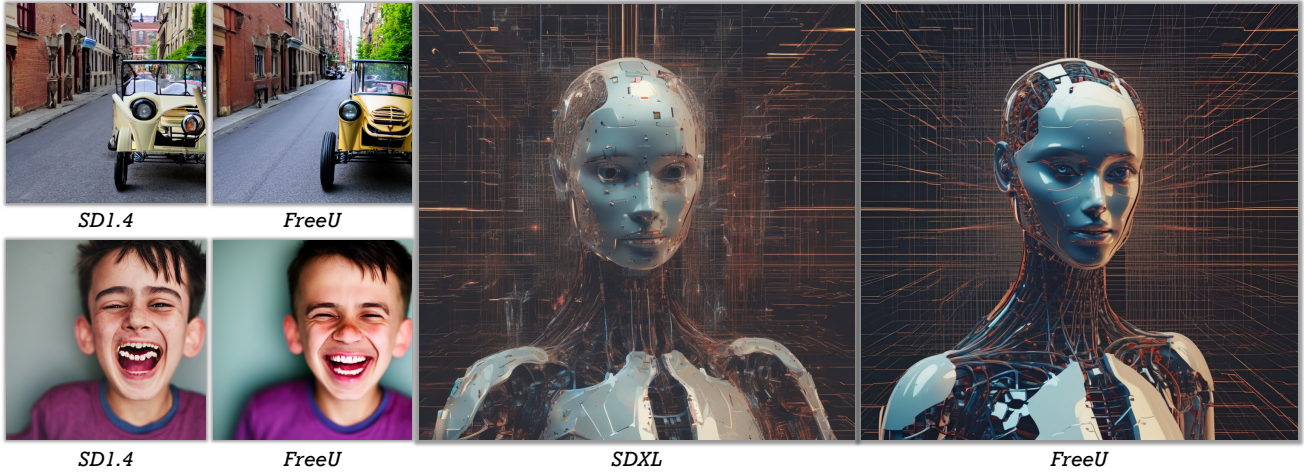{chenyang.si, ziqi002, yuming002, ziwei.liu}@ntu.edu.sg

Figure 1. We propose *FreeU*, a method that substantially improves diffusion model sample quality at no costs: no training, no additional parameter introduced, and no increase in memory or sampling time.

Abstract

*In this paper, we uncover the untapped potential of diffusion U-Net, which serves as a "free lunch" that substantially improves the generation quality on the fly. We initially investigate the key contributions of the U-Net architecture to the denoising process and identify that its main backbone primarily contributes to denoising, whereas its skip connections mainly introduce high-frequency features into the decoder module, causing the network to overlook the backbone semantics. Capitalizing on this discovery, we propose a simple yet effective method—termed "**FreeU**" — that enhances generation quality without additional training or finetuning. Our key insight is to strategically re-weight the contributions sourced from the U-Net's skip connections and backbone feature maps, to leverage the strengths of both components of the U-Net architecture. Promising results on image and video generation tasks demonstrate that our FreeU can be readily integrated to existing diffusion models, e.g., Stable Diffusion, DreamBooth, ModelScope, Rerender and ReVersion, to improve the generation quality with only a few lines of code. **All you need is to adjust two scaling factors during inference.** Project page: https://chenyangsi.top/FreeU/.*

1. Introduction

Diffusion probabilistic models, a cutting-edge category of generative models, have become a focal point in the research landscape, particularly for tasks related to computer vision [5, 6, 8, 10, 12, 20, 22, 26, 28, 29, 32]. Distinct from other classes of generative models [3, 7, 9, 16–19, 21, 25, 34, 35] such as Variational Autoencoder (VAE) [21], Generative Adversarial Networks (GANs) [3, 9, 16–19, 25], and vector-quantized approaches [7, 34], diffusion models introduce a novel generative paradigm. These models employ a fixed Markov chain to map the latent space, facilitating intricate mappings that capture latent structural complexities within a dataset. Recently, its impressive generative capabilities, ranging from the high level of details to the diversity of the generated examples, have fueled groundbreaking advancements in a variety of computer vision applications such as image synthesis [12, 29, 32], image editing [1, 4, 14, 24], image-to-image translation [4, 31, 36], and text-to-video generation [2, 11, 13, 23, 33, 37, 38, 40].

The diffusion models are comprised of the *diffusion process* and the *denoising process*. During the *diffusion process*, Gaussian noise is gradually added to the input data and eventually corrupts it into approximately pure Gaussian noise. During the *denoising process*, the original input data
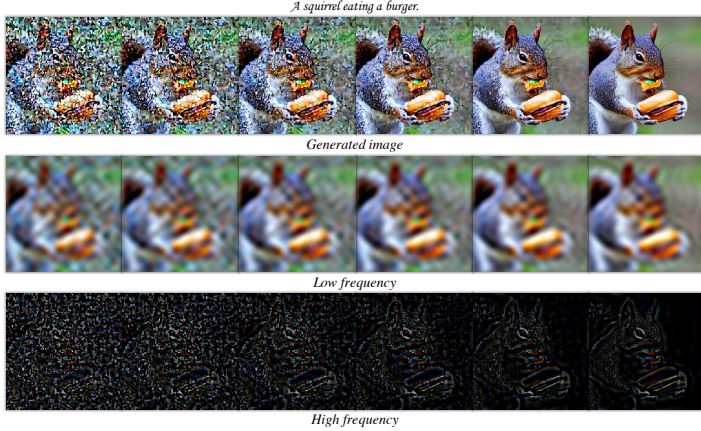
A squirrel eating a burger.

Generated image

Low frequency

High frequency

Figure 2. **The denoising process.** The top row illustrates the image's progressive denoising process across iterations, while the subsequent two rows display low-frequency and high-frequency components after the inverse Fourier Transform, matching each step. It's evident that low-frequency components change slowly, whereas high-frequency components exhibit more significant variations during the denoising process.

Figure 3. **Relative log amplitudes of Fourier with variations of the backbone scaling factor** b**.** Increasing in b correspondingly results in a suppression of high-frequency components in the images generated by the diffusion model.

is recovered from its noise state through a learned sequence of inverse diffusion operations. Usually, a U-Net is trained to iteratively predict the noise to be removed at each denoising step. Existing works focus on utilizing pre-trained diffusion U-Nets for downstream applications, while the internal properties of the diffusion U-Net, remain largely underexplored.

Beyond the application of diffusion models, in this paper, we are interested in investigating the effectiveness of diffusion U-Net for the denoising process. To better understand the denoising process, we first present a paradigm shift toward the Fourier domain to perspective the generated process of diffusion models, a research area that has received limited prior investigation. As illustrated in Fig. 2, the uppermost row provides the progressive denoising process, showcasing the generated images across successive iterations. The subsequent two rows exhibit the associated low-frequency and high-frequency spatial domain information after the inverse Fourier Transform, aligning with each respective step.

Evident from Fig. 2 is the gradual modulation of low-frequency components, exhibiting a subdued rate of change, while their high-frequency components display more pronounced dynamics throughout the denoising process. These findings are further corroborated in Fig.3. This can be intuitively explained: 1) Low-frequency components inherently embody the global structure and characteristics of an image, encompassing global layouts and smooth color. These components encapsulate the foundational global elements that constitute the image's essence and representation. Its rapid alterations are generally unreasonable in denoising
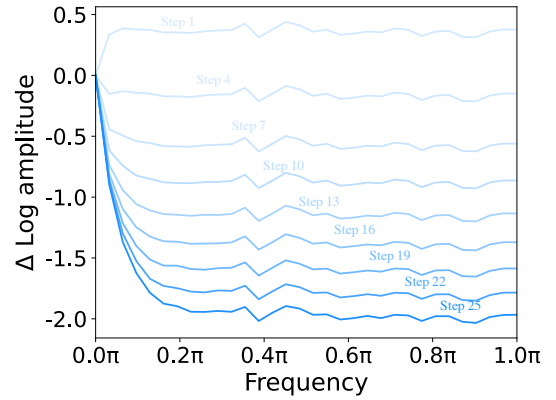
processes. Drastic changes to these components could fundamentally reshape the image's essence, an outcome typically incompatible with the objectives of denoising processes. 2) Conversely, high-frequency components contain the rapid changes in the images, such as edges and textures. These finer details are markedly sensitive to noise, often manifesting as random high-frequency information when noise is introduced to an image. Consequently, denoising processes need to expunge noise while upholding indispensable intricate details.

In light of these observations between low-frequency and high-frequency components during the denoising process, we extend our investigation to ascertain the specific contributions of the U-Net architecture within the diffusion framework. In each stage of the U-Net decoder, the skip features from the skip connection and the backbone features are concatenated together. Our investigation reveals that the main backbone of the U-Net primarily contributes to denoising. Conversely, the skip connections are observed to introduce high-frequency features into the decoder module. These connections propagate fine-grained semantic information to make it easier to recover the input data. However, an unintended consequence of this propagation is the potential weakening of the backbone's inherent denoising capabilities during the inference phase. This can lead to the generation of abnormal image details, as illustrated in the first row of Fig. 1.

Building upon this revelation, we propel forward with the introduction of a novel strategy, denoted as "**FreeU**", which holds the potential to improve sample quality without necessitating the computational overhead of additional

training or fine-tuning. During the inference stage, we instantiate two specialized modulation factors designed to balance the feature contributions from the U-Net architecture's primary backbone and skip connections. The first, termed the backbone feature factors, aims to amplify the feature maps of the main backbone, thereby bolstering the denoising process. However, we find that while the inclusion of backbone feature scaling factors yields significant improvements, it can occasionally lead to an undesirable oversmoothing of textures. To mitigate this issue, we introduce the second factor, skip feature scaling factors, aiming to alleviate the problem of texture oversmoothing.

Our FreeU framework exhibits seamless adaptability when integrated with existing diffusion models, encompassing applications like text-to-image generation and text-to-video generation. We conduct a comprehensive experimental evaluation of our approach, employing Stable Diffusion [29], DreamBooth [30], ReVersion [15], ModelScope [23], and Rerender [39] as our foundational models for benchmark comparisons. By employing FreeU during the inference phase, these models indicate a discernible enhancement in the quality of generated outputs. The visualization illustrated in Fig. 1 substantiates the efficacy of FreeU in significantly enhancing both intricate details and overall visual fidelity within the generated images. Our contributions are summarized as follows:

- We investigate and uncover the potential of U-Net architectures for denoising within diffusion models and identify that its main backbone primarily contributes to denoising, whereas its skip connections introduce high-frequency features into the decoder module.

- We further introduce a simple yet effective method, denoted as "**FreeU**", which enhances U-Net's denoising capability by leveraging the strengths of both components of the U-Net architecture. It substantially improves the generation quality without requiring additional training or fine-tuning.

- The proposed FreeU framework is versatile and seamlessly integrates with existing diffusion models. We demonstrate significant sample quality improvement across various diffusion-based methods, showing the effectiveness of FreeU at no extra cost.

2. Methodology

2.1. Preliminaries

Diffusion models such as Denoising Diffusion Probabilistic Models (DDPM) [12], encompass two fundamental processes for data modeling: a diffusion process and a denoising process. The diffusion process is characterized by a sequence of T steps. At each step t, Gaussian noise is incrementally introduced into the data distribution $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$ via a Markov chain, following a prescribed variance sched-

ule denoted as $\beta_1, \ldots, \beta_T$:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\mathcal{I}) \qquad (1)$$

The denoising process reverses the above diffusion process to the underlying clean data $\boldsymbol{x}_{t-1}$ given the noisy input $\boldsymbol{x}_t$:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)) \qquad (2)$$

The $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ determined through estimation procedures involving a denoising model denoted as ϵ_θ. Typically, this denoising model is implemented using a time-conditional U-Net architecture. It is trained to eliminate noise from data samples while concurrently enhancing the overall fidelity of the generated samples.

2.2. How does diffusion U-Net perform denoising?

Building upon the notable disparities observed between low-frequency and high-frequency components throughout the denoising process illustrated in Fig. 2 and Fig. 3, we extend our investigation to delineate the specific contributions of the U-Net architecture within the denoising process, to explore the internal properties of the denoising network. As depicted in Fig. 4, the U-Net architecture comprises a primary backbone network, encompassing both an encoder and a decoder, as well as the skip connections that facilitate information transfer between corresponding layers of the encoder and decoder.

The backbone of U-Net. To evaluate the salient characteristics of the backbone and lateral skip connections in the denoising process, we conduct a controlled experiment wherein we introduce two multiplicative scaling factors—denoted as b and s—to modulate the feature maps generated by the backbone and skip connections, respectively, prior to their concatenation. As shown in Fig. 5, it is evident that elevating the scale factor b of the backbone distinctly enhances the quality of generated images. Conversely, variations in the scaling factor s, which modulates the impact of the lateral skip connections, appear to exert a negligible influence on the quality of the generated images.

Building upon these observations, we subsequently probed the underlying mechanisms that account for the enhancement in image generation quality when the scaling factor b associated with the backbone feature maps is augmented. Our analysis reveals that this quality improvement is fundamentally linked to an amplified denoising capability imparted by the U-Net architecture's backbone. As delineated in Fig. 6, a commensurate increase in b correspondingly results in a suppression of high-frequency components in the images generated by the diffusion model. This implies that enhancing backbone features effectively bolsters the denoising capability of the U-Net architecture, thereby contributing to a superior output in terms of both fidelity and detail preservation.
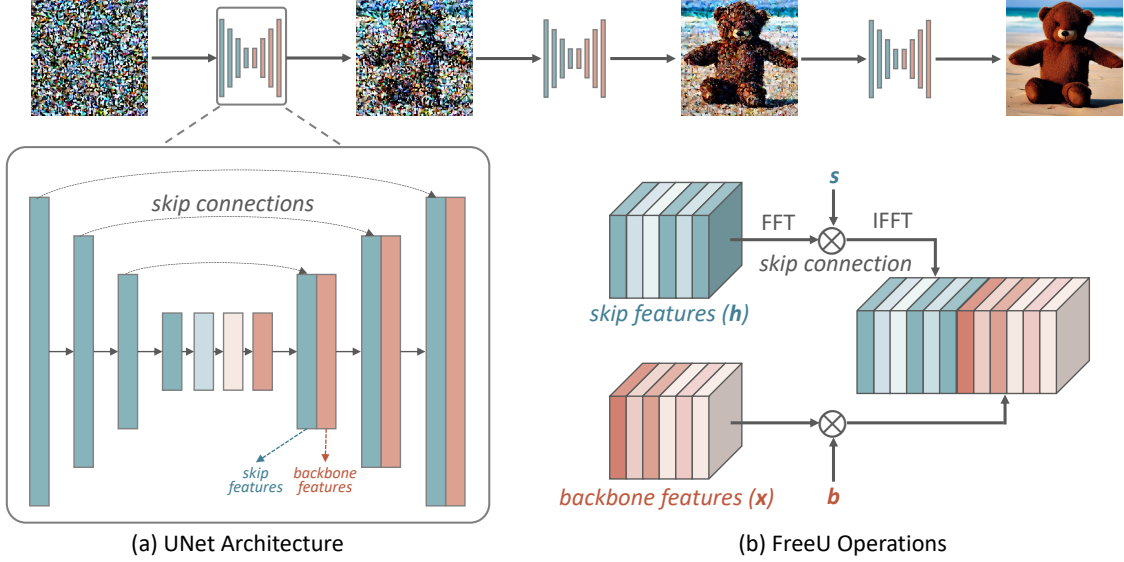
Figure 4. **FreeU Framework**. **(a) U-Net Skip Features and Backbone Features**. In U-Net, the skip features and backbone features are concatenated together at each decoding stage. We apply the FreeU operations during concatenation. **(b) FreeU Operations.** The factor b aims to amplify the backbone feature map x, while factor s is designed to attenuate the skip feature map h.
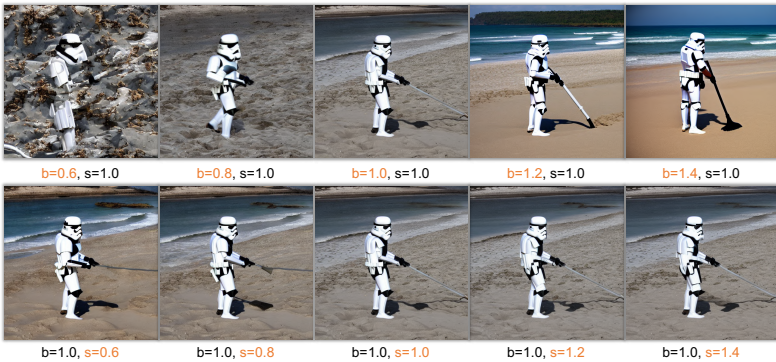

Figure 5. **Effect of backbone and skip connection scaling factors (b and s).** Increasing the backbone scaling factor b significantly enhances image quality, while variations in the skip scaling factor s have a negligible influence on image synthesis quality.
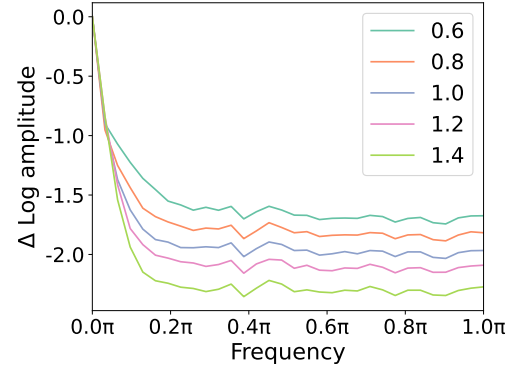
Figure 6. **Relative log amplitudes of Fourier with variations of the backbone scaling factor b.** Increasing in b correspondingly results in a suppression of high-frequency components in the images generated by the diffusion model.

The skip connections of U-Net. Conversely, the skip connections serve to forward features from the earlier layers of encoder blocks directly to the decoder. Intriguingly, as evidenced in Fig. 7, these features primarily constitute high-frequency information. Our conjecture, grounded in this observation, posits that during the training of the U-Net architecture, the presence of these high-frequency features may inadvertently expedite the convergence toward noise prediction within the decoder module. Furthermore, the limited impact of modulating skip features in Fig. 5 also indicates that the skip features predominantly contribute to the decoder's information. This phenomenon, in turn,

could result in an unintended attenuation of the efficacy of the backbone's intrinsic denoising capabilities during inference. Thereby, this observation prompts pertinent questions about the counterbalancing roles played by the backbone and the skip connections in the composite denoising performance of the U-Net framework.

2.3. Free lunch in diffusion U-Net

Capitalizing on the above discovery, we propel forward with the introduction of simple yet effective method, denoted as "**FreeU**", which effectively bolsters the denoising capability of the U-Net architecture by leveraging the strengths
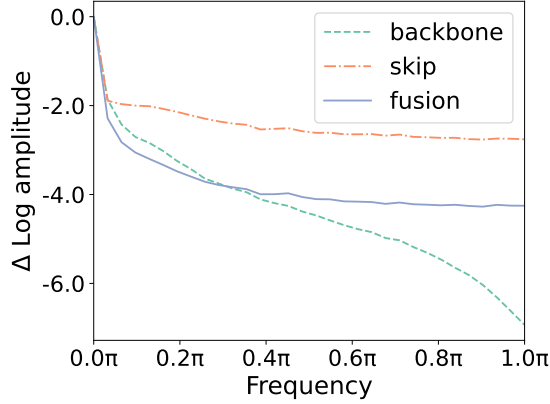
Figure 7. **Fourie relative log amplitudes of backbone, skip, and their fused feature maps.** The features, forwarded by skip connections directly from earlier layers of the encoder block to the decoder contain a large amount of high-frequency information.
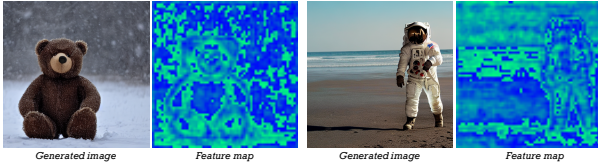

Figure 8. **Visualization of the average feature maps from the second stage in the decoder**.

of both components of the U-Net architecture. It substantially improves the generation quality without requiring additional training or fine-tuning.

Technically, for the l-th block of the U-Net decoder, let $\boldsymbol{x}_l$ represent the backbone feature map from the main backbone at the preceding block, and let $\boldsymbol{h}_l$ denote the feature map propagated through the corresponding skip connection. To modulate these feature maps, we introduce two scalar factors: a backbone feature scaling factor b_l for $\boldsymbol{x}_l$ and a yet-to-be-defined skip feature scaling factor s_l for $\boldsymbol{h}_l$. Specifically, the factor b_l aims to amplify the backbone feature map $\boldsymbol{x}_l$, while factor s_l is designed to attenuate the skip feature map $\boldsymbol{h}_l$.

For the backbone features, we introduce a novel method known as structure-related scaling, which dynamically adjusts the scaling of backbone features for each sample. Unlike a fixed scaling factor applied uniformly to all samples or positions within the same channel, our approach adjusts the scaling factor adaptively based on the specific characteristics of the sample features. We first computer the average feature map along the channel dimension:

$$\bar{\boldsymbol{x}}_l = \frac{1}{C} \sum_{i=1}^{C} \boldsymbol{x}_{l,i}, \qquad (3)$$

where $\boldsymbol{x}_{l,i}$ represents the i-th channel of the feature map $\boldsymbol{x}_l$. C denotes the total number of channels in $\boldsymbol{x}_l$. Subsequently, the backbone factor map is determined as follows:

$$\boldsymbol{\alpha}_l = (b_l - 1) \cdot \frac{\bar{\boldsymbol{x}}_l - Min(\bar{\boldsymbol{x}}_l)}{Max(\bar{\boldsymbol{x}}_l) - Min(\bar{\boldsymbol{x}}_l)} + 1, \qquad (4)$$

where $\boldsymbol{\alpha}_l$ represents the backbone factor map. b_l is a scalar constant. Then, upon experimental investigation, we discern that indiscriminately amplifying all channels of $\boldsymbol{x}_l$ through multiplication with $\boldsymbol{\alpha}_l$ engenders an oversmoothed texture in the resulting synthesized images. The reason is the enhanced U-Net compromises the image's high-frequency details while denoising. Consequently, we confine the scaling operation to the half channels of $\boldsymbol{x}_l$ as follows:

$$\boldsymbol{x}'_{l,i} = \begin{cases} \boldsymbol{x}_{l,i} \odot \boldsymbol{\alpha}_l, & \text{if } i < C/2 \\ \boldsymbol{x}_{l,i}, & \text{otherwise} \end{cases} \qquad (5)$$

Indeed, as illustrated in Fig. 8, the average feature map $\bar{\boldsymbol{x}}_l$ inherently contains valuable structural information. Consequently, the backbone factor map $\boldsymbol{\alpha}_l$ is instrumental in amplifying the backbone feature map $\boldsymbol{x}_l$ in a manner that aligns with its structural characteristics. This strategic approach serves to mitigate the issue of oversmoothing. Importantly, this strategy offers a dual benefit. Firstly, it enhances the denoising capabilities of the backbone feature map, allowing it to filter out noise more effectively. Secondly, it avoids the adverse effects associated with the indiscriminate application of scaling across the entire feature map, thereby achieving a more nuanced equilibrium between noise reduction and texture preservation.

To further mitigate the issue of oversmoothed texture due to enhancing denoising, we further employ spectral modulation in the Fourier domain to selectively diminish low-frequency components for the skip features. Mathematically, this operation is performed as follows:

$$\mathcal{F}(\boldsymbol{h}_{l,i}) = \text{FFT}(\boldsymbol{h}_{l,i}) \qquad (6)$$
$$\mathcal{F}'(\boldsymbol{h}_{l,i}) = \mathcal{F}(\boldsymbol{h}_{l,i}) \odot \boldsymbol{\beta}_{l,i} \qquad (7)$$
$$\boldsymbol{h}'_{l,i} = \text{IFFT}(\mathcal{F}'(\boldsymbol{h}_{l,i})) \qquad (8)$$

where FFT($\cdot$) and IFFT($\cdot$) are Fourier transform and inverse Fourier transform. $\odot$ denotes element-wise multiplication, and $\boldsymbol{\beta}_{l,i}$ is a Fourier mask, designed as a function of the magnitude of the Fourier coefficients, serving to implement the frequency-dependent scaling factor s_l:

$$\boldsymbol{\beta}_{l,i}(r) = \begin{cases} s_l & \text{if } r < r_{\text{thresh}}, \\ 1 & \text{otherwise}. \end{cases} \qquad (9)$$

where r is the radius. r_{thresh} is the threshold frequency. Then, the augmented skip feature map $\boldsymbol{h}'_l$ is then concatenated with the modified backbone feature map $\boldsymbol{x}'_l$ for subsequent layers in the U-Net architecture, as shown in Fig. 4.

5

| SD | SD + FreeU | SD | SD + FreeU | SD | SD + FreeU |

a blue car is being filmed   *Mother rabbit is raising baby rabbits*   *A bridge is depicted in the water*

a baby in a red shirt   *a attacks an upset cat and is then chased off*   *A teddy bear walking in the snowstorm*

A cat riding a motorcycle.   *A panda standing on a surfboard in the ocean*   *A boy is playing pokemon*

Figure 9. **Samples generated by Stable Diffusion [29] with or without FreeU.**

Remarkably, the proposed FreeU framework does not require any task-specific training or fine-tuning. Adding the backbone and skip scaling factors can be easily done with just a few lines of code. Essentially, the parameters of the architecture can be adaptively re-weighted during the inference phase, which allows for a more flexible and potent denoising operation without adding any computational burden. This makes FreeU a highly practical solution that can be seamlessly integrated into existing diffusion models to improve their performance.

3. Experiments

3.1. Implementation details

To assess the effectiveness of the proposed FreeU, we systematically conduct a series of experiments, aligning our benchmarks with state-of-the-art methods such as Stable Diffusion [29], DreamBooth [30], ModelScope [23], and Rerender [39]. Importantly, our approach seamlessly integrates with these established methods without imposing any additional computational overhead associated with supplementary training or fine-tuning. We meticulously adhere to the prescribed settings of these methods and exclusively introduce the backbone feature factors and skip feature factors during the inference.

3.2. Text-to-image

Stable Diffusion [29] is a latent text-to-image diffusion model renowned for its capability to generate photorealistic images based on textual input. It has consistently demonstrated exceptional performance in various image synthesis tasks. With the integration of our FreeU augmentation into Stable Diffusion, the results, as exemplified in Fig. 9, exhibit a notable enhancement in the model's generative capacity.

To elaborate, the incorporation of FreeU into Stable Diffusion [29] yields improvements in both entity portrayal and fine-grained details. For instance, when provided with the prompt *"a blue car is being filmed"*, FreeU refines the image, eliminating rooftop irregularities and enhancing the textural intricacies of the surrounding structures. In the case of *"Mother rabbit is raising baby rabbits"*, FreeU ensures that the generated image portrays a mother rabbit in a normal appearance caring for baby rabbits. Furthermore, In scenarios like *"a attacks an upset cat and is then chased off"* and *"A teddy bear walking in the snowstorm"*, FreeU helps generate more realistically posed cats and teddy bears. Impressively, in response to the complex prompt *"A cat riding a motorcycle"*, FreeU not only accurately renders the individual entities but also expertly captures the nu-

6

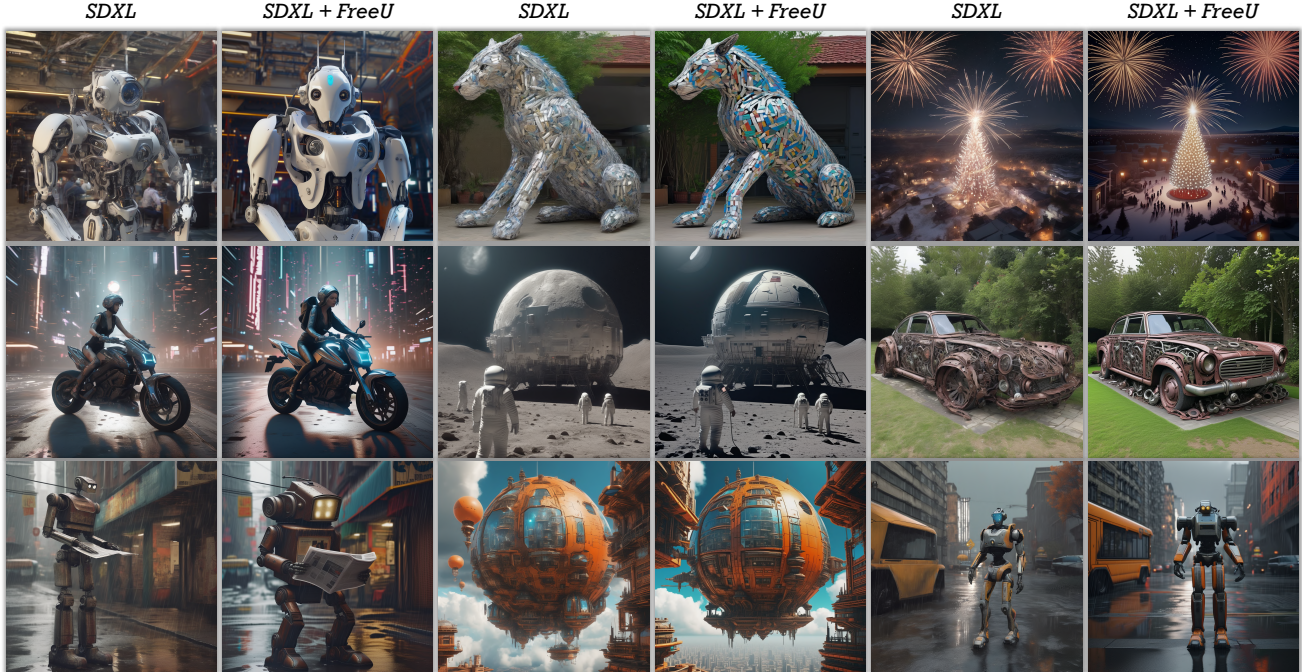| *SDXL* | *SDXL + FreeU* | *SDXL* | *SDXL + FreeU* | *SDXL* | *SDXL + FreeU* |

Figure 10. **Samples generated by Stable Diffusion-XL [27] with or without FreeU.**

anced relationship between them, ensuring that the cat is actively engaged in riding. In Figure 10, we present the generated images based on the SDXL framework [27]. It becomes evident that our proposed FreeU consistently excels in generating realistic images, especially in detail generation. These compelling results serve as a testament to the substantial qualitative enhancements engendered by the synergy of FreeU with the SD[29] or SDXL[27] frameworks.

Quantitative evaluation. We conduct a study with 35 participants to assess *image quality* and *image-text alignment*. Each participant receives a text prompt and two corresponding synthesized images, one from SD and another from SD+FreeU. To ensure fairness, we use the same randomly sampled random seed for generating both images. The image sequence is randomized to eliminate any bias. Participants then select the image they consider superior for *image-text alignment* and *image quality*, respectively. We tabulate the votes for SD and SD+FreeU in each category in Table 1. Our analysis reveals that the majority of votes go to SD+FreeU, indicating that FreeU significantly enhances the Stable Diffusion text-to-image model in both evaluated aspects.

3.3. Text-to-video

ModelScope [23], an avant-garde text-to-video diffusion model, stands at the forefront of video generation from textual descriptions. The infusion of our FreeU augmentation

Table 1. **Text-to-Image Quantitative Results.** We count the percentage of votes for the baseline and our method respectively. *Image-Text* refers to *Image-Text Alignment*.
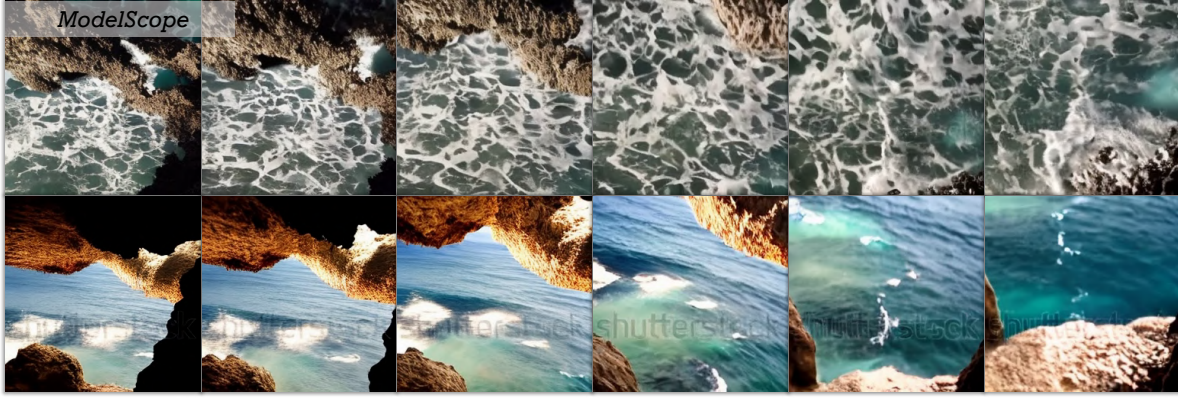
Method	Image-Text	Image Quality
SD [29]	14.12%	14.66%
SD+FreeU	**85.88%**	**85.34%**

Table 2. **Text-to-Video Quantitative Results.** We count the percentage of votes for the baseline and our method respectively. *Video-Text* refers to *Video-Text Alignment*.
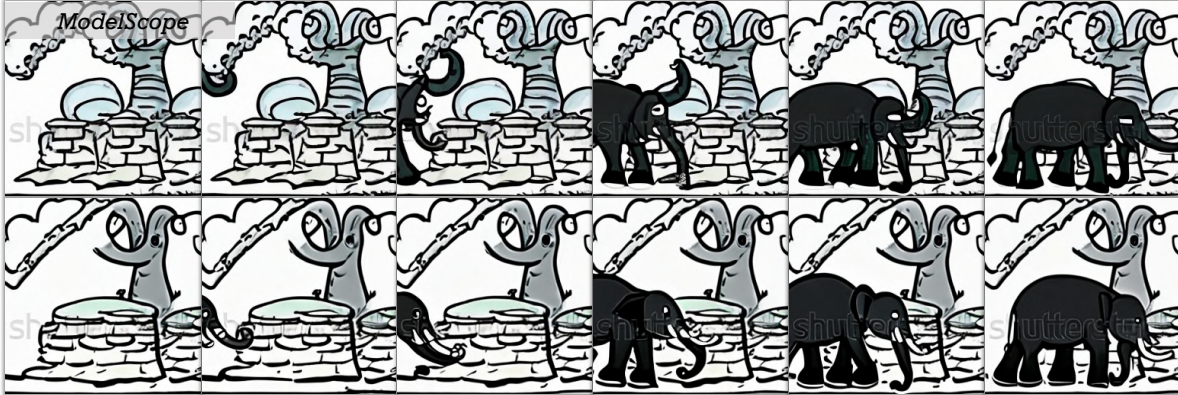
Method	Video-Text	Video Quality
ModelScope [23]	15.29%	14.33%
ModelScope+FreeU	**84.71%**	**85.67%**

into ModelScope [23] serves to further hone its video synthesis prowess, as substantiated by Fig. 11. For instance, when presented with the prompt *"A cinematic view of the ocean, from a cave"*, FreeU enables ModelScope [23] to generate the perspective "from a cave", enriching the visual narrative. In the case of *"A cartoon of an elephant walking"*, ModelScope [23] initially generates an elephant with two trunks, but with the incorporation of FreeU, it rectifies this anomaly and produces a correct depiction of an elephant in motion. Moreover, in response to the prompt *"An astronaut flying in space"*, ModelScope [23], with the assistance of FreeU, can generate a clear and vivid portrayal of an astronaut floating in the expanse of outer space.

A cinematic view of the ocean, from a cave.

A cartoon of an elephant walking.

An astronaut flying in space.

Figure 11. **Samples generated by ModelScope [23] with or without FreeU.**

These results underscore the significant improvements achieved through the synergistic application of FreeU with ModelScope [23], resulting in high-quality generated content characterized by clear motion, rich detail, and semantic alignment.

Quantitative evaluation. We conduct the quantitative evaluation for FreeU on the text-to-video task in a similar way as text-to-image. The results displayed in Table 2 indi-

cate that most participants prefer the video generated with FreeU.

3.4. Downstream tasks

FreeU presents substantial enhancements in the quality of synthesized samples across various diffusion model applications. Our evaluations extend from foundational image and video synthesis models to more specialized downstream
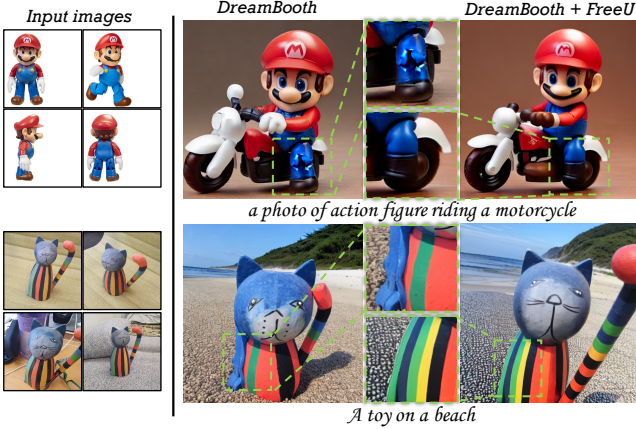
Figure 12. **Samples generated by DreamBooth [30] with or without FreeU.**

Figure 13. **Samples generated by ReVersion [15] with or without FreeU.**

applications.

We incorporate FreeU into Dreambooth [30], a diffusion model specialized in personalized text-to-image tasks. The enhancements are evident, as demonstrated in Fig. 12, the synthesized images present marked improvements in realism. For instance, while the base DreamBooth [30] model struggles to synthesize the appearance of the action figure's legs from the prompt *"a photo of action figure riding a motorcycle"*, the FreeU-augmented version deftly overcomes this hurdle. Similarly, for the prompt *"A toy on a beach"*, the initial output exhibited body shape anomalies. FreeU's integration refines these imperfections, providing a more accurate representation and improving color fidelity.

We also integrate FreeU into ReVersion [15], a Stable Diffusion based relation inversion method, enhancing its quality as shown in Fig. 13. For example, when the relation "back to back" is to be expressed between two children, FreeU enhances ReVersion's ability to accurately represent

Figure 14. **Samples generated by Rerender [39] with or without FreeU.**

this relationship. For the "inside" relation, when a *dog* is supposed to be placed inside of a *basket*, ReVersion sometimes generates a dog with artifacts, and introducing FreeU helps eliminate these artifacts. While ReVersion effectively captures relational concepts, Stable Diffusion might occasionally struggle to synthesize the relation concept due to excessive high-frequency noises in the U-Net skip features. Adding FreeU allows better entity and relation synthesis quality by using exactly the same relation prompt learned by ReVersion.

Furthermore, we evaluated FreeU's impact on Rerender [39], a diffusion model tailored for zero-shot text-guided video-to-video translations. Fig. 14 depicts the results: clear improvements in the detail and realism of synthesized videos. For instance, when provided with the prompt *"A dog wearing sunglasses"* and an input video, Rerender [39] initially produces a dog video with artifacts related to the *"sunglasses"*. However, the incorporation of FreeU successfully eliminates such artifacts, resulting in a refined output.

In summation, these outcomes substantiate that the incorporation of FreeU leads to enhanced entity representation and synthesis quality, employing precisely the same learned prompt.

3.5. Ablation study

Effects of FreeU. FreeU is introduced with the primary aim of enhancing the denoising capabilities of the U-Net architecture within the diffusion model. To assess the impact of FreeU, we conducted analytical experiments using Sta-
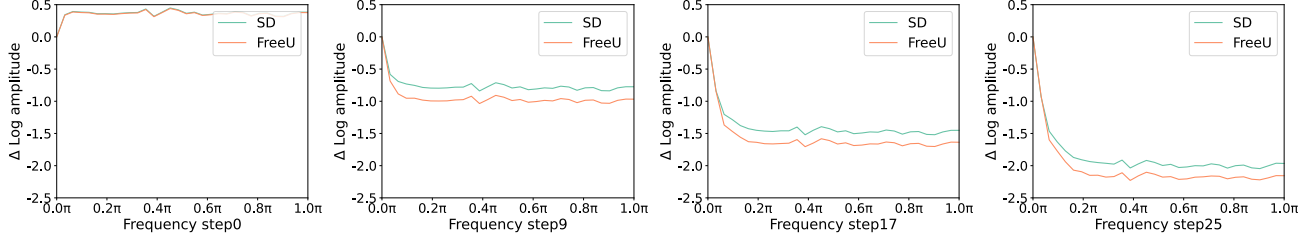
Figure 15. **Fourier relative log amplitudes of Stable Diffusion [29] with or without FreeU within the denoising process.**
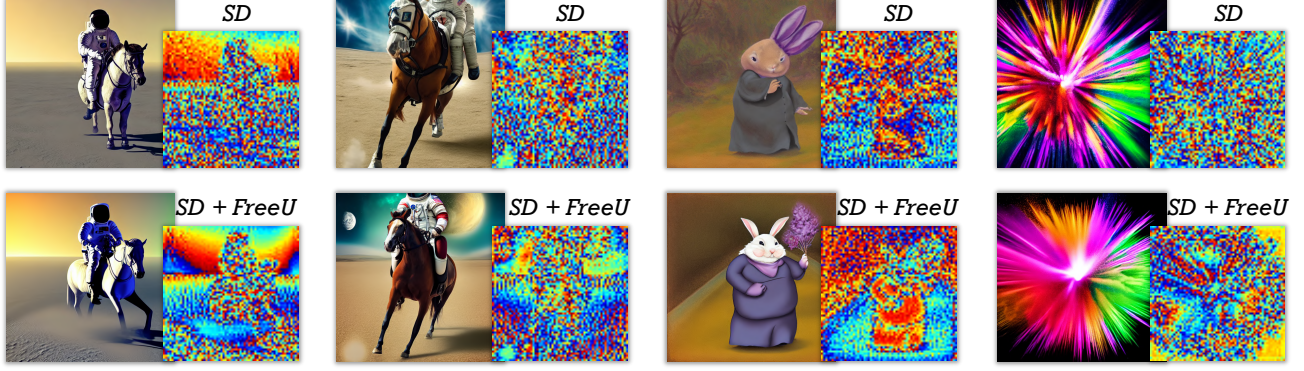

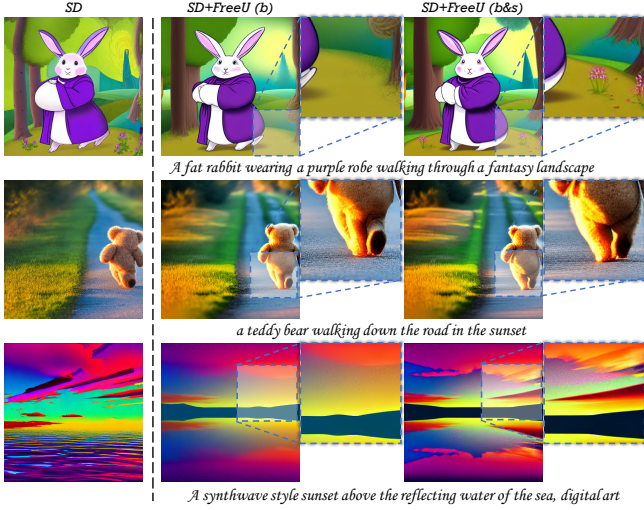Figure 16. The visualization of feature maps for Stable Diffusion [29] with or without FreeU.

Figure 17. **The ablation study of backbone scaling factor and skip scaling factor.**

Figure 18. **The ablation study of backbone scaling factor. (a) The generated images of SD. (b) The generated images of FreeU with a constant factor. (c) The generated images of FreeU with the structure-related scaling factor map.**

ble Diffusion [29] as the base framework. In Fig. 15, we present visualizations of the relative log amplitudes of the Fourier transform of Stable Diffusion [29], comparing cases with and without the incorporation of FreeU. These visualizations illustrate that FreeU exerts a discernible influence in reducing high-frequency information at each step of the denoising process, which indicates FreeU's capacity to ef-
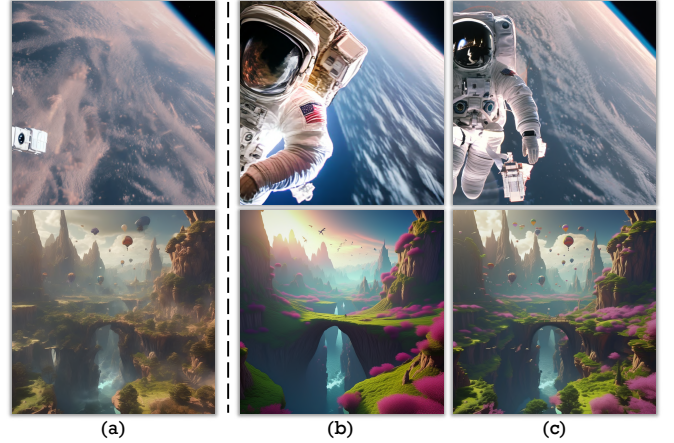
fectively denoising. Furthermore, we extended our analysis by visualizing the feature maps of the U-Net architecture. As shown in Fig. 16, we observe that the feature maps generated by FreeU contain more pronounced structural information. This observation aligns with the intended effect of FreeU, as it preserves intricate details while effectively removing noise, harmonizing with the denoising objectives of

10

the model.

Effects of components in FreeU. We evaluate the effects of the proposed FreeU strategy, *i.e.* introducing backbone feature scaling factors and skip feature scaling factors to intricately balance the feature contributions from the U-Net architecture's primary backbone and skip connections. In Fig. 17, we present the results of our evaluations. In the case of *SD+FreeU(b)*, where backbone scaling factors are integrated during inference, we observe a noticeable improvement in the generation of vivid details compared to *SD* [29] alone. For instance, when given the prompt *"A fat rabbit wearing a purple robe walking through a fantasy landscape"*, *SD+FreeU(b)* generates a more realistic rabbit with normal arms and ears, as opposed to *SD* [29]. However, it is imperative to note that while the inclusion of feature scaling factors yields significant improvements, it can occasionally lead to an undesirable oversmoothing of textures. To mitigate this issue, we introduce skip feature scaling factors, aiming to reduce low-frequency information and alleviate the problem of texture oversmoothing. As demonstrated in Fig. 17, the combination of both backbone and skip feature scaling factors in *SD+FreeU(b & s)* leads to the generation of more realistic images. For instance, in the prompt *"A synthwave style sunset above the reflecting water of the sea, digital art"*, the generated sunset sky in *SD+FreeU(b & s)* exhibits enhanced realism compared to *SD+FreeU(b)*. This highlights the efficacy of the comprehensive FreeU strategy in balancing features and mitigating issues related to texture smoothing, ultimately resulting in more faithful and realistic image generation.

Effects of backbone structure-related factor. We evaluate the effects of the proposed backbone scaling strategy, structure-related scaling, on the delicate balance between noise reduction and texture preservation. Illustrated in Figure 18, when compared to the results generated by *SD* [29], we observe a substantial enhancement in the image quality generated by FreeU when utilizing a constant scaling factor. However, it is pertinent to highlight that the utilization of a fixed scaling factor can engender adverse consequences, manifesting as pronounced oversmoothing of textures and undesirable color oversaturation. Conversely, FreeU with the structure-related scaling factor map employs an adaptive scaling approach, leveraging structural information to guide the assignment of the backbone factor map. Our observations indicate that FreeU with the structure-related scaling factor map effectively mitigates these issues and achieves significant improvements in generating vivid and intricate details.

4. Conclusion

In this study, we introduce the elegantly simple yet highly effective approach, termed **FreeU**, which substantially enhances the sample quality of diffusion models without in-

curring any additional computational costs. Motivated by the fundamental role played by both skip connections and backbone features in U-Net architectures, we conduct an in-depth analysis of their effects in diffusion U-Net. Our investigation reveals that the primary backbone primarily contributes to denoising, while the skip connections predominantly introduce high-frequency features into the decoder, potentially leading to a neglect of essential backbone semantics. To address this, we strategically re-weight the contributions originating from the U-Net's skip connections and backbone feature maps. This re-weighting process capitalizes on the unique strengths of both U-Net components, resulting in a substantial improvement in sample quality across a wide range of text prompts and random seeds. Our proposed **FreeU** can be seamlessly integrated into various diffusion foundation models and their downstream tasks, offering a versatile means of enhancing sample quality.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 1

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 1

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *ICCV*, 2021. 1

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

[6] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. ImageBART: Bidirectional context with multinomial diffusion for autoregressive image synthesis. In *NeurIPS*, 2021. 1

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

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 1

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

[10] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 1

[11] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity

video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 1

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

[13] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 1

[14] Ziqi Huang, Kelvin C.K. Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *CVPR*, 2023. 1

[15] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin C.K. Chan, and Ziwei Liu. ReVersion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 3, 9

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 1

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

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

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

[20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 1

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[22] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 1

[23] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. VideoFusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023. 1, 3, 6, 7, 8

[24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 1

[25] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[26] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 7

[28] 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

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 6, 7, 10, 11

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3, 6, 9

[31] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH*, 2022. 1

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[33] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1

[34] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 1

[35] Pei Wang, Yijun Li, and Nuno Vasconcelos. Rethinking and improving the robustness of image style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 124–133, 2021. 1

[36] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 1

[37] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 1

[38] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022. 1

[39] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 3, 6, 9

[40] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation, 2023. 1