Any-Size-Diffusion: Toward Efficient Text-Driven Synthesis for Any-Size HD Images

Qingping Zheng[1, 2*], Yuanfan Guo[2*], Jiankang Deng[2], Jianhua Han[2], Ying Li[1†], Songcen Xu[2], Hang Xu[2†]

[1]Northwestern Polytechnical University
[2]Huawei Noah's Ark Lab
zhengqingping2018@mail.nwpu.edu.cn, jiankangdeng@gmail.com, lybyp@nwpu.edu.cn,
{guoyuanfan1, hanjianhua4, xusongcen}@huawei.com, chromexbjxh@gmail.com

Abstract

Stable diffusion, a generative model used in text-to-image synthesis, frequently encounters resolution-induced composition problems when generating images of varying sizes. This issue primarily stems from the model being trained on pairs of single-scale images and their corresponding text descriptions. Moreover, direct training on images of unlimited sizes is unfeasible, as it would require an immense number of text-image pairs and entail substantial computational expenses. To overcome these challenges, we propose a two-stage pipeline named ***Any-Size-Diffusion*** (**ASD**), designed to efficiently generate well-composed HD images of any size, while minimizing the need for high-memory GPU resources. Specifically, the initial stage, dubbed Any Ratio Adaptability Diffusion (ARAD), leverages a selected set of images with a restricted range of ratios to optimize the text-conditional diffusion model, thereby improving its ability to adjust composition to accommodate diverse image sizes. To support the creation of images at any desired size, we further introduce a technique called Fast Seamless Tiled Diffusion (FSTD) at the subsequent stage. This method allows for the rapid enlargement of the ASD output to any high-resolution size, avoiding seaming artifacts or memory overloads. Experimental results on the LAION-COCO and MM-CelebA-HQ benchmarks show that ASD can produce well-structured images of arbitrary sizes, cutting down the inference time by 2× compared to the traditional tiled algorithm. The source code is available at https://github.com/ProAirVerse/Any-Size-Diffusion.

Introduction

In text-to-image synthesis, Stable Diffusion (SD) (Rombach et al. 2022) has emerged as a significant advancement. Existing SD models (Ruiz et al. 2023; Meng et al. 2023) transform text aligned with image components into high-quality images, typically sized at 512×512 pixels. Despite these models having the ability to handle varying sizes, they noticeably struggle with resolution changes, resulting in poor composition (*e.g.*, improper cropping and unnatural appearance), a problem demonstrated in Figure 1(a). The root of this issue lies in the models trained mainly on pairs of text and images of a uniform size, overlooking the complexities of handling images at multiple resolutions. Consequently, this leads to observed deficiencies in image composition.

Figure 1: Resolution-induced Poor Composition. Given the text, (a) $SD_{2.1}$ and (b) $MD_{2.1}$, a MultiDiffusion model, raise poor composition issues in red boxes when synthesizing images of varying sizes, as opposed to (c) our ASD.

In pursuit of generating well-structured images at arbitrary aspect ratios, guided by textual descriptions, the MultiDiffusion methodology (Bar-Tal et al. 2023) leverages a pretrained text-conditional diffusion (*e.g.*, stable diffusion), as a reference model and controls image synthesis through the utilization of several reference diffusion processes. Remarkably, the entire process is realized without requiring further training or fine-tuning. While efficient, it does not completely resolve the limitations associated with handling the reference model's multi-resolution images. As a result, the production of images may exhibit suboptimal composi-

*These authors contributed equally.
†Corresponding authors.

tional quality. The underlying reason is also tied to the reference model's training on images constrained to a single-scale size, as illustrated in Figure 1(b).

A direct and appealing solution to the problem is to train the SD model to cope with every possible image size. Yet, this approach encounters an immediate and significant barrier: the infinite diversity of image ratios, which makes it practically unfeasible. Furthermore, it's challenging to gather an extensive collection of high-resolution images and corresponding text pairs. Even with a plethora of high-quality datasets available, the intrinsic pixel-based nature of SD requires substantial computational resources, particularly when dealing with high-resolution images of various sizes. The problem is further aggravated when considering the use of megapixel images for SD training, as this involves extensive repeated function equations and gradient computations in the high-dimensional space of RGB images (Ho, Jain, and Abbeel 2020). Even when a trained model is ready, the inference step is also time-consuming and memory-intensive. Through empirical observation, we have found that attempts to generate 4K HD images using the SD model trigger out-of-memory errors when executed on a GPU with a 32GB capacity.

The key insight of this paper is to introduce a pioneering ***A**ny-**S**ize-**D**iffusion* (**ASD**) model, executed in two stages, which can synthesize high-resolution images of arbitrary sizes from text prompts. In its dual-phase approach, our ASD not only efficiently handles the resolution-induced poor composition but also successfully circumvents out-of-memory challenges. At the outset, we are faced with the complexity of accommodating all conceivable image sizes, a challenge that might seem intractable. To address this, in the first stage, we introduce a multi-aspect ratio training strategy that operates within a well-defined, manageable range of ratios. This strategy is used to optimize our proposed *Any Ratio Adaptability Diffusion* (ARAD) model. As a result, it enables the production of well-composed images that are adaptable to any size within a specified range, while also ensuring a reduced consumption of GPU memory. To yield images that can fit any size, in the second stage, we propose an additional method called *Fast Seamless Tiled Diffusion* (FSTD) to magnify the image output originating from the preceding ARAD. Contrary to the existing tiled diffusion methods (Álvaro Barbero Jiménez 2023), which address the seaming issue but compromise on the speed of inference, our proposed FSTD designs an implicit overlap within the tiled sampling diffusion process. This innovation manages to boost inference speed without the typical seaming problems, achieving the high-fidelity image magnification. To sum up, the contributions of this paper are as follows:

- We are the first to develop the ***A**ny-**S**ize-**D**iffusion* (**ASD**) model, a two-phase pipeline that synthesizes high-resolution images of any size from text, addressing both composition and memory challenges.
- We introduce a multi-aspect ratio training strategy, implemented within a defined range of ratios, to optimize ARAD, allowing it to generate well-composed images adaptable to any size within a specified range.
- We propose an implicit overlap in FSTD to enlarge images to arbitrary sizes, effectively mitigating the seaming problem and simultaneously accelerating the inference time by $2\times$ compared to the traditional tiled algorithm.

Related Work

Stable Diffusion. Building upon the foundations laid by the Latent Diffusion Model (LDM) (Rombach et al. 2022), diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2021) have achieved substantial success across various domains, including text-to-image generation (Nichol et al. 2022; Ramesh et al. 2022; Saharia et al. 2022), image-to-image translation (Dhariwal and Nichol 2021; Nichol and Dhariwal 2021), and multi-modal generation (Ruan et al. 2023). Owing to their robust ability to capture complex distributions and create diverse, high-quality samples, diffusion models excel over other generative methods (Goodfellow et al. 2014). In the field, Stable Diffusion (SD) (Rombach et al. 2022) has emerged as a leading model for generating photo-realistic images from text. While adept at producing naturalistic images at certain dimensions (*e.g.*, 512×512), it often yields unnatural outputs with sizes beyond this threshold. This constraint principally originates from the fact that existing stable diffusion models are exclusively trained on images of a fixed size, leading to a deficiency in high-quality composition on other sizes.

Diffusion-based Image Super-Resolution. The objective of image super-resolution is to infer a high-resolution image from a corresponding low-resolution counterpart. The utilization of generative models to magnify images often omits specific assumptions about degradation, leading to challenging situations in real-world applications. Recently, diffusion-based methods (Sahak et al. 2023; Saharia et al. 2023; Li et al. 2023; Ma et al. 2023) have shown notable success in real-world SR by exploiting generative priors within these models. Though effective, these approaches introduce considerable computational complexity during training, with a quadratic increase in computational demands as the latent space size increases. An optimized method, known as StableSR(Wang et al. 2023), was developed to enhance performance while reducing GPU memory consumption. However, this method can become time-inefficient when processing images divided into numerous overlapping regions.

Method

To resolve the issue of resolution-induced poor composition when creating high-fidelity images of various sizes from any text prompt, we propose a straightforward yet efficient approach called *Any Size Diffusion* (ASD). This approach simplifies the process of text-to-image synthesis by breaking it down into two stages (see Figure 2).

- **Stage-I, termed as Any Ratio Adaptability Diffusion (ARAD)**, trains on multiple aspect-ratio images and generates an image conditioned on a textual description and noise size, avoiding poor composition issues.
- **Stage-II, known as Fast Seamless Tiled Diffusion (FSTD)**, magnifies the image from Stage-I to a predeter-

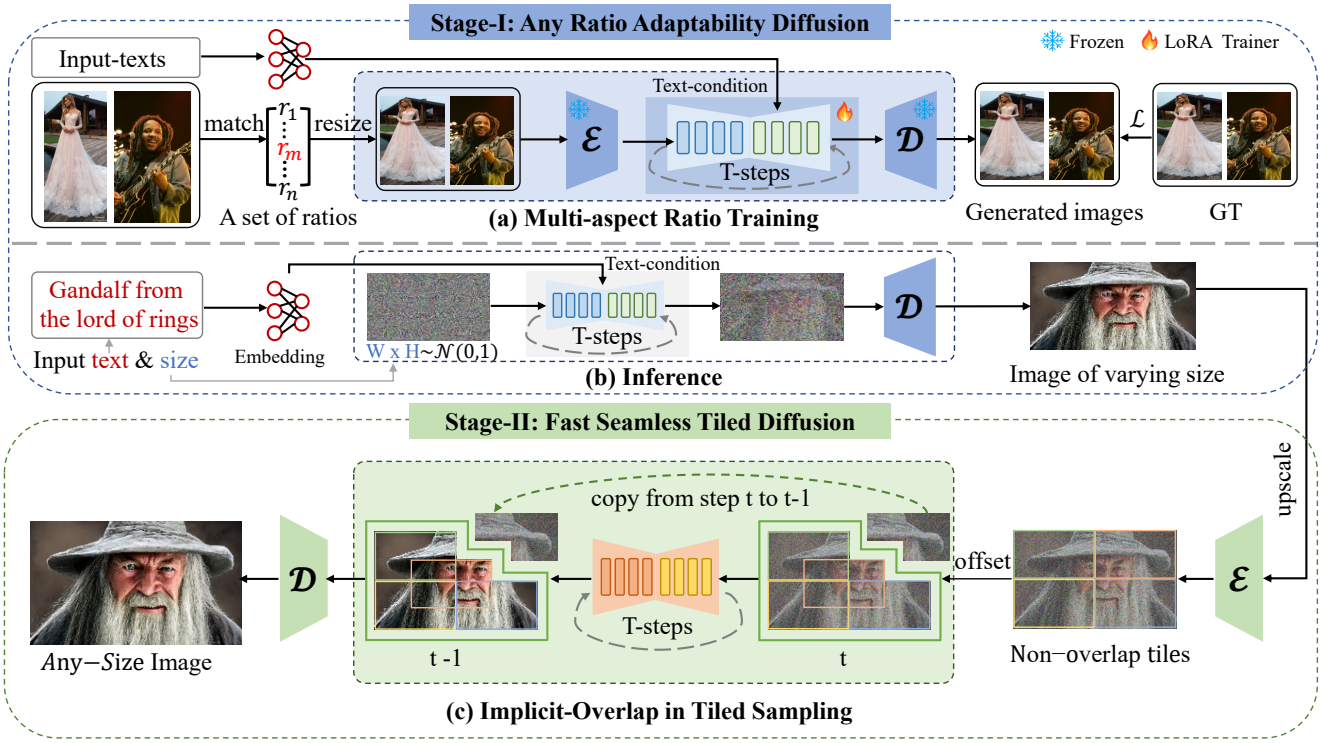


Figure 2: The Any-Size-Diffusion (ASD) pipeline, including: 1) Stage-I translates text into images, adapting to various aspect ratios, and 2) Stage-II is responsible for transforming low-resolution images from the Stage-I into high-resolution versions of any specified size. For procedure (c), the implicit overlap in tiled sampling, only the solid green line region is sent to the UNetModel for current denoising. At Stage-II, the dashed green arrow represents regions that are directly copied from the preceding denoised latents, potentially enhancing efficiency and consistency within the overall process.

mined larger size, ultimately producing a high-resolution synthesized image, adjustable to any specified size.

Pipeline

Given a user-defined text prompt (*e.g.,* "Gandalf from the lord for the rings") and noise size, ARAD employs the pretrained text encoder (Cherti et al. 2023) to generate a contextual embedding $\tau_\theta(y)$ and initializes random noise ϵ at the base resolution. The noisy input conditioned on the textual embedding $p(\epsilon|y)$ is progressively denoised by the UNetModel (Cherti et al. 2023). This process is iterated through T times, leveraging the DDPM algorithm (Song, Meng, and Ermon 2020) to continuously remove noises and restore the latent representation z. Ultimately, a decoder $\mathcal{D}$ converts the denoised latent back into an image $I \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the image's height and width respectively.

Afterward, FSTD takes the resulting image as input and performs inference based on the image-conditional diffusion (Wang et al. 2023). In detail, the image is magnified by a specified size. A pretrained visual encoder $\mathcal{E}$ is employed to map the resulting image $I' \in \mathbb{R}^{H'\times W'\times 3}$ into a latent representation $z = \mathcal{E}(I')$. A normal distribution-based noise $\epsilon \sim \mathcal{N}(0,1)$ is then added to it, yielding the noisy latent variable $z' = \mathcal{A}(z)$. The image, conditioned on itself $p(z'|z)$, undergoes progressive iterations by the UNetModel, utilizing our proposed tiled sampling $I \in \mathbb{R}^{H\times W\times 3}$ for T cycles. Lastly, the decoder $\mathcal{D}$ is employed to project the denoised latent variable into the final output, effectively transforming the latent space back into the image domain.

Any Ratio Adaptability Diffusion (ARAD)

In this stage, ARAD is proposed to make the model have the capability of generating an image, adjustable to varying aspect ratios, resolving the issue of resolution-induced poor composition. This stage is mainly achieved by our designed multi-aspect ratio training strategy.

Multi-aspect ratio training. Instead of directly training on the original image and text pairs, we employ our aspect-ratio strategy to map the original image into an image with a specific ratio. To be more precise, we define a set of ratios $\{r_1, r_2, ..., r_n\}$, each corresponding to specific sizes $\{s_1, s_2, ..., s_n\}$, where n represents the number of predefined aspect ratios. For each training image $x \in \mathbb{R}^{H\times W\times 3}$, we calculate the image ratio as $r = H/W$. This ratio r is then compared with each predefined ratio, selecting the one m with the smallest distance as the reference ratio. The index m is determined by

$$\arg\min_m f(m) = \{|r_1 - r|, \cdots, |r_m - r|, \cdots, |r_n - r|\}, \quad (1)$$

where $f(m)$ represents the smallest distance between the

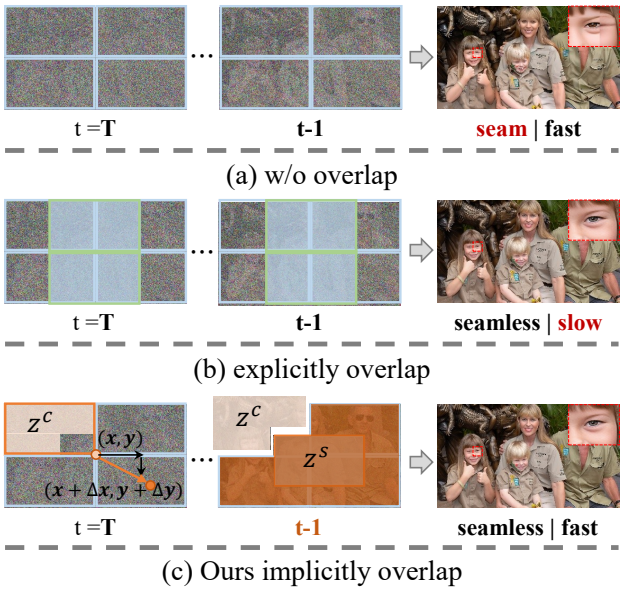


Figure 3: Comparison of various tiling strategies: (a) without overlapping, (b) with explicit overlapping, and (c) with implicit overlapping. Green tiles are explicit overlaps, and the orange tile is our implicit overlap at step t-1.

current ratio and the predefined ratio. Therefore, if the image has a ratio similar to the m^{th} predefined size s_m, the original size of the training image is resized to s_m.

Forward ARAD process. During the training process, a pair consisting of an image and its corresponding text (x, y) is processed, where x represents an image in the RGB space $\mathbb{R}^{H\times W\times 3}$, and y denotes the associated text. A fixed visual encoder, $\mathcal{E}$, is used to transform the resized image s_m into a spatial latent code z. Meanwhile, the corresponding text is converted into a textual representation $\tau_\theta(y)$ via OpenCLIP (Cherti et al. 2023). For the total steps T, conditional distributions of the form $p(z_t|y)$, $t = 1\cdots T$, can be modeled using a denoising autoencoder $\epsilon_\theta(z_t, t, y)$. Consequently, the proposed ARAD can be learned using an objective function

$$L_{ARAD} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]. \quad (2)$$

Fast Seamless Tiled Diffusion (FSTD)

In the second stage, we propose FSTD, a training-free approach built on StableSR (Wang et al. 2023) that amplifies the ARAD-generated image to any preferred size. To achieve efficient image super-resolution without heavy computational demands during inference, we devise an implicit overlap technique within the tiled sampling method.

Tiled sampling. For clarity, consider an upscaled image $\mathcal{I} \in \mathbb{R}^{H'\times W'\times 3}$, partitioned into M petite tiles, symbolized as $\{P_i \in \mathbb{R}^{h\times w\times 3} \mid 1 \le i \le M\}$, where w and h denote the width and height of each tile. We initially encode each tile P_i using an encoder function $\mathcal{E}$, adding the random noise, to generate a set of noisy latent representations $\mathcal{Z} = \{\mathcal{Z}_i = \mathcal{E}(P_i) + \epsilon_i \mid \epsilon_i \sim \mathcal{N}(0,1),\ 1 \le i \le M\}$. Subsequently, each noisy tile is processed by the UNetModel conditioned on the original tile for T steps, resulting in a set of denoised latents $\mathcal{Z}' = \{\mathcal{Z}'_i \mid \epsilon_i \sim \mathcal{N}(0,1),\ 1 \le i \le M\}$. Finally, a decoder $f_\mathcal{D}$ is applied to convert them back into image space, culminating in the reconstructed image

$$\mathcal{I}' = \{P'_i \in \mathbb{R}^{h\times w\times 3} \mid P'_i = f_\mathcal{D}(\mathcal{Z}'_i), 1 \le i \le M\}. \quad (3)$$

Herein, P'_i represents the i^{th} tile decoded from its corresponding denoised latent tile.

However, a seaming problem emerges when any two tiles in the set are disjoint, as depicted in Figure 3(a). To tackle this, we implement overlaps between neighboring tiles that share common pixels (Figure 3(b)). While increasing explicit overlap can effectively mitigate this issue, it substantially escalates the denoising time. As a consequence, the inference time quadratically increases with the growth in overlapping patches. Indeed, it's practically significant to strike a balance between inference time and the amount of overlap.

Implicit overlap in tiled sampling. To speed up the inference time while avoiding the seaming problem, we propose an implicit overlap in tiled sampling. As depicted in Figure 3(c), the magnified image is divided into L non-overlapping tiles and we keep the quantity of disjoint noisy latent variables constant during the reverse sampling process. Before each sampling step, we apply a random offset to each tile, effectively splitting $\mathcal{Z}$ into two components: $\mathcal{Z}^s$ (the shifted region with tiling) and $\mathcal{Z}^c$ (the constant region without tiling). This can be mathematically represented as $\mathcal{Z} = \mathcal{Z}^s \cup \mathcal{Z}^c$. Take note that at the initial time step, $\mathcal{Z}^c = \emptyset$. At each sampling, the shifted part, $\mathcal{Z}^s$, is a collection of L disjoint tiles, denoted as $\mathcal{Z}^s = \{\mathcal{Z}^s_i \mid 1 \le i \le L\}$. Here, each $\mathcal{Z}^s_i$ symbolizes a shifted tile. The shifted portion, $\mathcal{Z}^s$, comprises L disjoint tiles that change dynamically throughout the sampling process. Within this segment, each tile is expressed as $\mathcal{Z}^s_{i,x,y} = \mathcal{Z}_{y_i+\Delta y_i, x_i+\Delta x_i}$ for $1 \le i \le L$. Here, Δx_i and Δy_i denote the random offsets for tile $\mathcal{Z}^s_i$ implemented in the preceding step. As for the constant section without tiling, denoted as $\mathcal{Z}^c$, the pixel value is sourced from the corresponding latent variable in the previous sampling step. It is worth noting that after each time step, $\mathcal{Z}^c$ is non-empty, symbolically represented as $\mathcal{Z}^c \neq \emptyset$. This approach ensures implicit overlap during tiled sampling, effectively solving the seaming issue.

Experiments

Experimental Settings

Datasets. The ARAD of our ASD is trained on a subset of LAION-Aesthetic (Schuhmann 2022) with 90k text-image pairs in different aspect ratios. It is evaluated on MA-LAION-COCO with 21,000 images across 21 ratios (selecting from LAION-COCO (Schuhmann et al. 2022)), and MA-COCO built from MS-COCO (Lin et al. 2014) containing 2,100 images for those ratios. A test split of MM-CelebA-HQ (Xia et al. 2021), consisting of 2,824 face image pairs in both low and high resolutions, is employed to evaluate our FSTD and whole pipeline.

Implementation Details. Our proposed method is implemented in PyTorch (Paszke et al. 2019). A multi-aspect ratio training method is leveraged to finetune ARAD (using

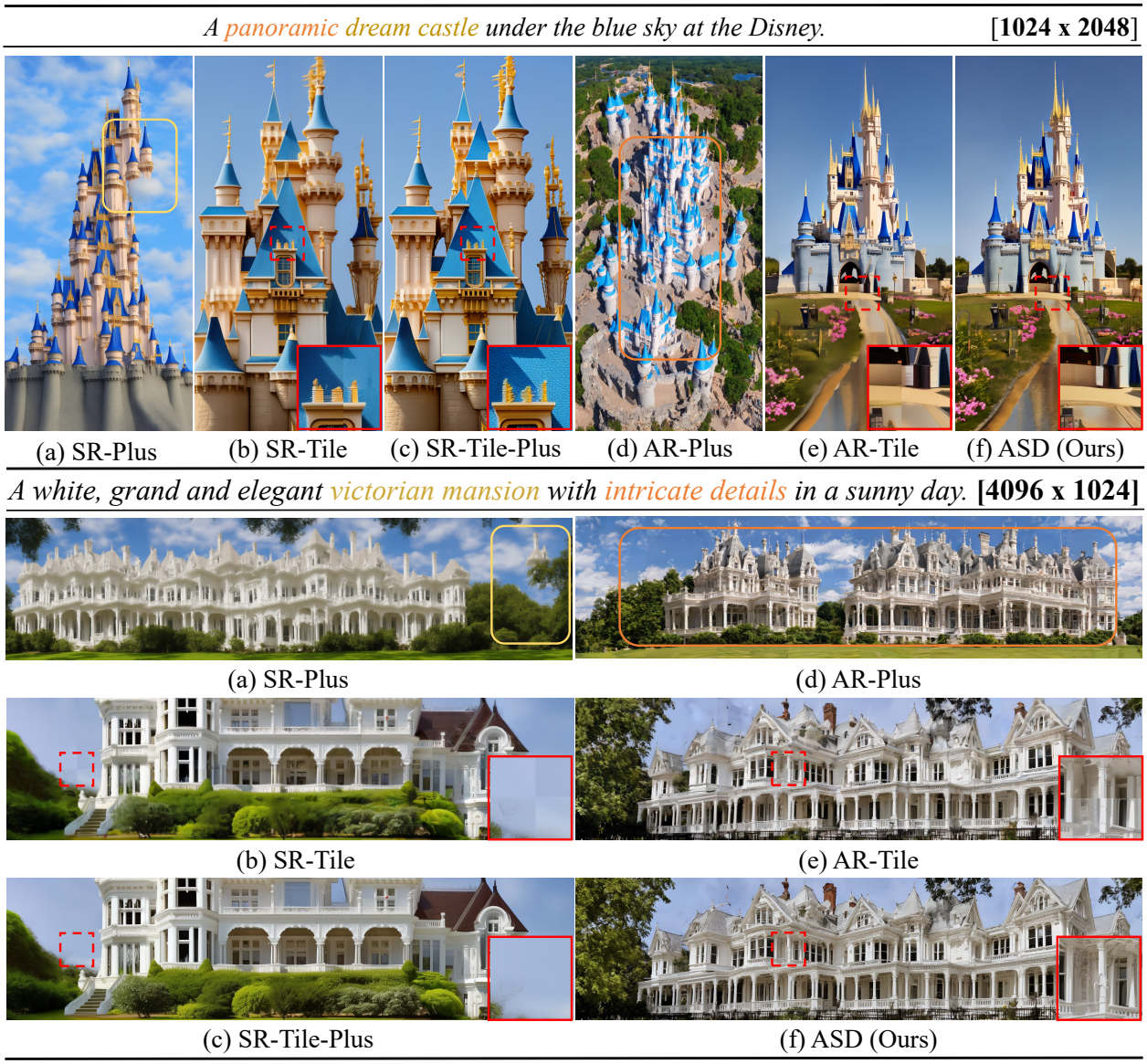


Figure 4: Qualitative comparison of our proposed ASD method with other baselines, including (a) SR-Plus, (b) SR-Tile, (c) SR-Tile-Plus, (d) AR-Plus, (e) AR-Tile and (f) our proposed ASD. The yellow box indicates the resolution-induced poor composition. The orange box indicates better composition. The red solid line box is the zoom-in of the red dashed line box, aiming to inspect if there are any seaming issues. Our ASD outperforms others in both composition quality and inference time.

Exp.	Stage-I	Stage-II		Capability			MM-CelebA-HQ		
	Ratio	Tile	Overlap	Composition	Max Resolution	Seam	FID ↓	IS ↑	CLIP ↑
(a)	**S**	✗	✗	Poor	2048^2	N	118.83	2.11	27.22
(b)	**S**	✓	✗	Poor	$1843 2^2$	Y	111.96 (- 6.87)	2.46 (+ 0.35)	27.46 (+ 0.24)
(c)	**S**	✓	✓	Poor	$1843 2^2$	N	111.06 (- 7.77)	2.53 (+ 0.42)	27.55 (+ 0.33)
(d)	**A**	✗	✗	Excellent	2048^2	N	92.80 (- 26.03)	3.97 (+ 1.86)	29.15 (+ 1.93)
(e)	**A**	✓	✗	Excellent	$1843 2^2$	Y	85.66 (- 33.17)	3.98 (+ 1.87)	29.17 (+ 1.95)
(f)	**A**	✓	✓	Excellent	$1843 2^2$	N	**85.34** (- 33.49)	**4.04** (+ 1.93)	**29.23** (+ 2.01)

Table 1: Quantitative evaluation against baselines. (a) SR-Plus, (b) SR-Tile, (c) SR-Tile-Plus, (d) AR-Plus, (e) AR-Tile and (f) our ASD. 'S' and 'A' denote single and arbitrary ratios, respectively. All tests run on a 32G GPU. Notably, under the same GPU memory, our ASD achieves at least 9× higher resolution than the original SD model.

LoRA (Hu et al. 2021)) for 10,000 steps with a batch size of 8. We use Adam (Kingma and Ba 2014) as an optimizer and the learning rate is set to 1.0e-4. Our FSTD (the second stage model) is training-free and is built upon StableSR (Wang et al. 2023). During inference, DDIM sampler (Song, Meng, and Ermon 2020) of 50 steps is adopted in ARAD to generate the image according to the user-defined aspect ratio. In the second stage, we follow StableSR to use 200 steps DDPM sampler (Ho, Jain, and Abbeel 2020) for FSTD.

Evaluation metrics. For benchmarks, we employ common perceptual metrics to assess the generative text-to-image models, including FID (Heusel et al. 2017), IS (Salimans et al. 2016) and CLIP (Radford et al. 2021). IS correlates with human judgment, important to evaluate the metric on a large enough number of samples. FID captures the disturbance level very well and is more consistent with the noise level than the IS. CLIP score is used to measure the cosine similarity between the text prompt and the image embeddings. Besides, the extra metrics (*e.g.,* PSNR, SSIM (Wang et al. 2004) and LPIPS (Zhang et al. 2018)) are employed to assess the super-resolution ability of the second stage of our ASD. PSNR and SSIM scores are evaluated on the luminance channel in the YCbCr color space. LPIPS quantifies the perceptual differences between images.

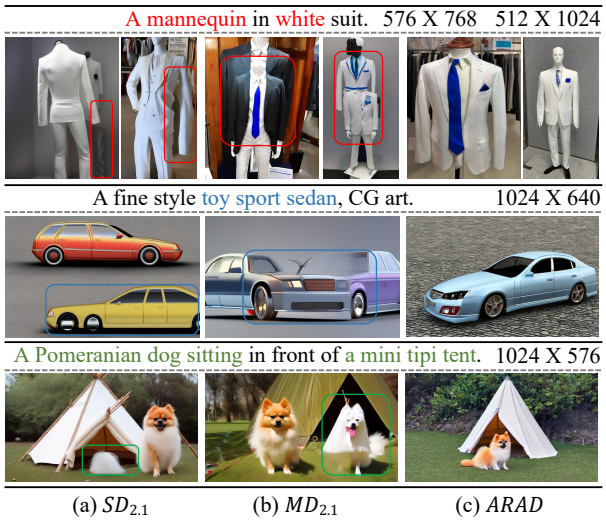
Baseline Comparisons

- **SR-Plus:** employs SD 2.1 (Rombach et al. 2022) for the direct synthesis of text-guided images with varying sizes.
- **SR-Tile:** utilizes SD 2.1 for initial image generation, magnified using StableSR (Wang et al. 2023) with a non-overlap in tiled sampling(Álvaro Barbero Jiménez 2023).
- **SR-Tile-Plus:** A two-stage method that initiates with SD 2.1 and refines the output using our proposed FSTD, facilitating the synthesis of images of arbitrary dimensions.
- **AR-Plus:** deploys our proposed ARAD model for direct, text-driven image synthesis across a spectrum of sizes.
- **AR-Tile:** commences with our ARAD model for initial image generation, followed by magnification via StableSR employing a non-overlap in tiled sampling.
- **ASD:** is our proposed novel framework, integrating ARAD in Stage I and FTSD in Stage II, designed to synthesize images with customizable dimensions.

Quantitative evaluation. As reported in Table 1, our proposed ASD method consistently outperforms the baseline methods. Specifically, our ASD model shows a 33.49 reduction in FID score compared to (a) SR-Plus, and an increase of 1.92 and 2.01 in IS and CLIP scores, respectively. On a 32GB GPU, SR-Plus fails to synthesize images exceeding 2048^2 resolution. In contrast, our ASD effectively mitigates this constraint, achieving at least 9× higher resolution than SR-Plus under identical hardware conditions. Additionally, we also have the following observations: **(i)** Utilizing multi-aspect ratio training results in notable improvements across various comparisons, specifically reducing FID scores from 118.83 to 92.80 in (a)-(d), 111.96 to 85.66 in (b)-(e), and 111.06 to 85.34 in (c)-(f). **(ii)** Introducing a tiled algorithm at the second stage enables the generation of images with unlimited resolution, while simultaneously enhancing performance, *e.g.,* FID scores improve from 92.80 to 85.66 when comparing (a)-(b) and (d)-(c). **(iii)** Implementing overlap in tiled sampling effectively addresses the seaming issue, as evidenced by the comparisons between (b)-(c) and (e)-(f).

Method	MA-LAION-COCO			MA-COCO		
	FID ↓	IS ↑	CLIP ↑	FID ↓	IS ↑	CLIP ↑
$SD_{2.1}$	14.32	31.25	31.92	42.50	30.20	31.63
$MD_{2.1}$	14.57	28.95	32.11	43.25	28.92	30.92
ARAD	**13.98**	**34.03**	**32.60**	**40.28**	29.77	**31.87**

Table 2: Comparison of our ARAD and other diffusion-based approaches. We compare their compositional ability to handle the synthesis of images across 21 different sizes.



(a) $SD_{2.1}$ (b) $MD_{2.1}$ (c) $ARAD$

Figure 5: Comparison of visual results. Composition quality of the text-to-image synthesis using (a) $SD_{2.1}$, a stable diffusion 2.1, (b) $MD_{2.1}$, a multi-diffusion based on SD 2.1, and (c) our ARAD. Color boxes indicate poor composition.

Qualitative comparison. As depicted in Fig. 4, the images synthesized by ASD exhibit superior composition quality (*e.g.* proper layout) when compared to other baseline methods. Additionally, ASD can generate 4K HD images that are not only well-composed but also free from seaming artifacts. Specifically, when guided by a text description, the AR-Plus method is observed to generate a more complete castle than SR-Plus, as demonstrated in Fig.4(a) *vs.* Fig.4(d). Compared with SR-Plus, AR-Tile can produce realistic images but is hindered by the presence of seaming issues (see Fig. 4(e)). In contrast, Fig. 4(f) shows that our ASD successfully eliminates seaming artifacts and ensures the production of well-composed images, while minimizing GPU memory usage.

ARAD Analysis

Impact of ARAD. Table 2 highlights the performance of ARAD, showing improvements of 13.98, 34.03, and 32.60 in FID, IS, and CLIP, respectively, on MA-LAION-COCO

Types	MA-LAION-COCO			MA-COCO		
	FID ↓	IS ↑	CLIP ↑	FID ↓	IS ↑	CLIP ↑
3	14.36	32.53	32.38	41.28	29.58	31.71
5	14.10	33.61	32.58	40.25	29.63	31.80
All	**13.98**	**34.03**	**32.60**	**40.28**	**29.77**	**31.87**

Table 3: Performance on ARAD trained on the various types of aspect ratios. "All" denotes the 9 aspect ratios.

Method	MM-CelebA-HQ				Time
Overlap & Offset	PSNR↑	SSIM↑	LPIPS↓	FID↓	1/fps
w/o overlap	26.89	0.76	0.09	22.80	75.08s
explicit	27.49	0.76	0.09	24.15	166.8s
implicit & fixed	26.83	0.75	0.08	21.37	**75.01s**
implicit & random	**27.53**	0.76	0.08	22.25	75.19s

Table 4: The versatility of tiled sampling in FSTD. "w/o", "explicit", and "implicit" describe non-overlapping, explicit, and implicit overlap in tile sampling respectively. "fixed", and "random" refer to different tile offset strategies. Here, the overlap of two adjacent tiles is 32×32.

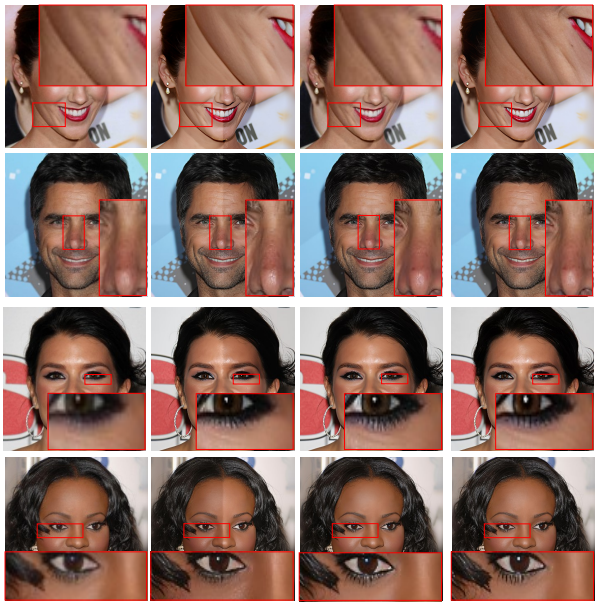
over original SD 2.1 and MultiDiffusion (Bar-Tal et al. 2023) ($MD_{2.1}$). This superiority is further illustrated in Fig. 5. While $SD_{2.1}$ and $MD_{2.1}$ exhibit composition problems, our ASD produces images that are consistent with user-defined textual descriptions. For example, $MD_{2.1}$ incorrectly generates two overlapped blue suits from a prompt for a white suit, a mistake not present in our ASD's results.

Influence on the number of aspect ratios. Table 3 reveals the model's performance across various aspect ratios. The data shows that increasing the number of aspect ratios in the training dataset improves performance, with FID scores falling from 14.36 to 13.98. A comparison between 3 and 5 aspect ratios highlights a significant improvement, as the FID score drops from 14.36 to 14.10. Further increasing the aspect ratios continues this trend, reducing the FID score to 13.98. This pattern emphasizes the importance of aspect ratios in enhancing model performance.

FSTD Analysis

Importance of tiles with overlap. The first two lines from Table 4 reveal a comparison between the perceptual performance of explicit overlap and non-overlap in tiled sampling. Specifically, the explicit overlap exhibits superior performance (*e.g.,* 27.49 *vs.* 26.89 on PSNR). However, non-overlap tiled sampling offers an approximately 2× faster inference time compared to the explicit overlap. Despite this speed advantage, Fig. 6(b) exposes the seaming problem associated with non-overlap tiled sampling, highlighting the trade-off between performance and efficiency.

Implicit vs. explicit overlap. Table 4 and Fig.6(c)-(d) confirm that the use of implicit overlap in tiled sampling yields the best performance across both perceptual metrics and visual representation. Further examination of the last column in Table 4 demonstrates that the inference time for implicit



(a) Zoomed LR (b) w/o overlap (c) Explicit **(d) Implicit**

Figure 6: The super-resolved results of ×4 for different methods, including (a) Zoomed LR (bicubic), tiled diffusion with (b) non-overlap and (c) explicit overlap tiles; and (d) our FSTD which uses implicit overlap in tiled sampling.

overlap in tiled sampling is nearly equivalent to that of tiling without overlap. Moreover, the implementation of implicit overlap successfully reduces the inference time from 166.8s to approximately 75.0s. This validates our FSTD's superiority in balancing quality and inference time optimally.

Effect of various offset strategies. The last two lines of Table 4 demonstrate the advantage of using a random offset in implicit overlap tiled sampling. Specifically, when comparing the fixed and random offset methods in implicit overlap, the random offset yields a PSNR value of 27.53, outperforming the fixed offset, which registered at 26.83. The results for other perceptual metrics and visual performance indicators are found to be nearly identical, further emphasizing the preference for a random offset in this context.

Conclusion

In this study, we address the challenge of resolution-induced poor composition in creating high-fidelity images from any text prompt. We propose *Any Size Diffusion* (ASD), a method consisting of ARAD and FSTD. Trained with multi-aspect ratio images, ARAD generates well-composed images within specific sizes. FSTD, utilizing implicit overlap in tiled sampling, enlarges previous-stage output to any size, reducing GPU memory consumption. Our ASD is validated both quantitatively and qualitatively in real-world scenes.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62271400), and the Shaanxi Provincial Key R&D Program, China (2023-GHZD-02).

References

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *ICML*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *CVPR*, 2818–2829.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat Gans on Image Synthesis. *NIPS*, 34: 8780–8794.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS*, volume 27.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *NIPS*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *NIPS*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Li, R.; Zhou, Q.; Guo, S.; Zhang, J.; Guo, J.; Jiang, X.; Shen, Y.; and Han, Z. 2023. Dissecting Arbitrary-scale Super-resolution Capability from Pre-trained Diffusion Generative Models. *arXiv preprint arXiv:2306.00714*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, 740–755.

Ma, Y.; Yang, H.; Yang, W.; Fu, J.; and Liu, J. 2023. Solving Diffusion ODEs with Optimal Boundary Conditions for Better Image Super-Resolution. *arXiv*.

Meng, C.; Rombach, R.; Gao, R.; Kingma, D.; Ermon, S.; Ho, J.; and Salimans, T. 2023. On Distillation of Guided Diffusion Models. In *CVPR*, 14297–14306.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic models. In *ICML*, 8162–8171.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, 16784–16804.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32.

Radford, A.; Wook Kim, J.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 8821–8831.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10684–10695.

Ruan, L.; Ma, Y.; Yang, H.; He, H.; Liu, B.; Fu, J.; Yuan, N. J.; Jin, Q.; and Guo, B. 2023. Mm-diffusion: Learning Multi-modal Diffusion Models for Joint Audio and Video Generation. In *CVPR*, 10219–10228.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*, 22500–22510.

Sahak, H.; Watson, D.; Saharia, C.; and Fleet, D. 2023. Denoising Diffusion Probabilistic Models for Robust Image Super-Resolution in the Wild. *arXiv preprint arXiv:2302.07864*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *NIPS*, 35: 36479–36494.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2023. Image Super-Resolution via Iterative Refinement. *TPAMI*, 45(4): 4713–4726.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. *NIPS*, 29.

Schuhmann, C. 2022. LAION-AESTHETICS. https://laion.ai/blog/laion-aesthetics/. Accessed: 2022-8-16.

Schuhmann, C.; Köpf, A.; Vencu, R.; Coombes, T.; and Beaumont, R. 2022. LAION COCO: 600M Synthetic Captions from LAION2B-EN. https://laion.ai/blog/laion-coco/. Accessed: 2022-9-15.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *ICLR*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *arXiv preprint arXiv:2305.07015*.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image Quality Assessment: from Error Visibility to Structural Similarity. *TIP*, 13(4): 600–612.

Xia, W.; Yang, Y.; Xue, J.-H.; and Wu, B. 2021. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. In *CVPR*, 2256–2265.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 586–595.

Álvaro Barbero Jiménez. 2023. Mixture of Diffusers for Scene Composition and High Resolution Image Generation. *arXiv preprint arXiv:2302.02412*.