ElasticDiffusion: Training-free Arbitrary Size Image Generation

Moayed Haji-Ali, Guha Balakrishnan, Vicente Ordonez
Rice University
{mh155, guha, vicenteor}@rice.edu

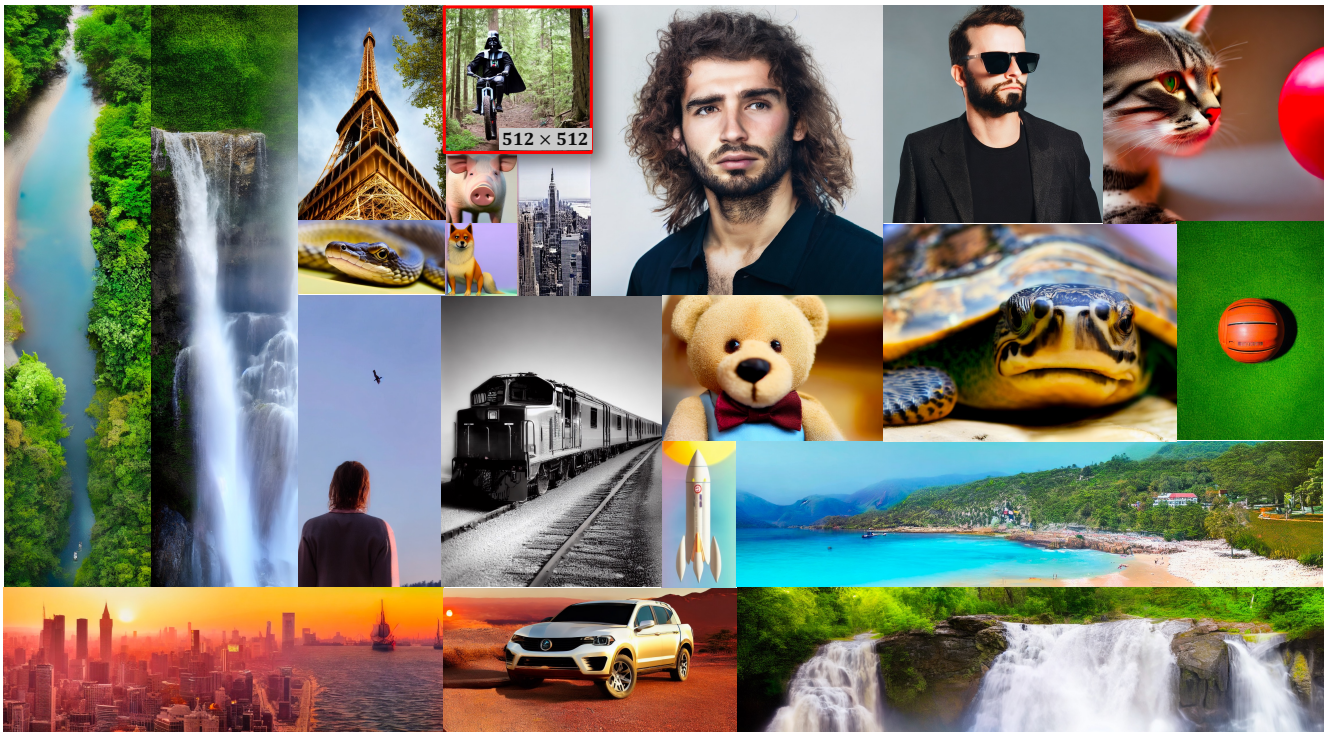


Figure 1. **ElasticDiffusion** generates high quality images at arbitrary sizes using a pretrained diffusion model trained on a single image size, with equivalent memory footprint and no further training. These results are based on Stable Diffusion$_{1.4}$, which was trained to generate 512×512 images. The examples shown in this collage are presented without any image cropping, stretching, or post-processing.

Abstract

Diffusion models have revolutionized image generation in recent years, yet they are still limited to a few sizes and aspect ratios. We propose ElasticDiffusion, a novel training-free *decoding method that enables pretrained text-to-image diffusion models to generate images with various sizes. ElasticDiffusion attempts to decouple the generation trajectory of a pretrained model into local and global signals. The local signal controls low-level pixel information and can be estimated on local patches, while the global signal is used to maintain overall structural consistency and is estimated with a reference image. We test our method on CelebA-HQ (faces) and LAION-COCO (objects/indoor/outdoor scenes). Our experiments and qualitative results show superior image coherence quality across aspect ratios compared to MultiDiffusion and the standard decoding strategy of Stable Diffusion.*

1. Introduction

Diffusion models are a powerful family of algorithms that achieve remarkable quality and obtain the current state-of-the-art performance on various image synthesis tasks. As is the case with most deep neural networks, diffusion models are typically trained on one or a few image resolutions. For instance, Stable Diffusion [31], one of the most widely adopted diffusion models, is trained on a square shape of size 512 x 512, yet fails to maintain its performance at different aspect ratios during inference time. In practice, many applications require a wide aspect ratio or portrait mode, such as digital billboards, wearable devices, artistic renderings, automotive displays, and any application relying on a computer screen. Recent studies address the issue of variable image size in different ways. SDXL [27] and Any-Size-Diffusion [45] explicitly finetune models on images with a

range of aspect ratios, which requires extensive computation, a quadratic memory footprint, and larger training data. In addition, this strategy requires a set of resolutions to be specified up-front during training time, while the models still struggle to generalize to new resolutions during inference, often resulting in artifacts. Recent works also show remarkable results in generating panoramic images using pretrained diffusion models by overlapping generated patches into a larger image [3, 46]. These methods work well for landscape images with repetitive patterns. However, their lack of global guidance limits their abilities to generate images of single objects or faces where global structure is important.

In this work, we propose ElasticDiffusion, a novel decoding strategy that takes a pretrained diffusion model and generate images at arbitrary sizes during inference using a constant memory footprint. To achieve this, ElasticDiffusion revisits the guidance mechanism of conditional diffusion models to decouple global and local content generation. Global content controls the high-level aspects of the image, whereas local content adds finer, more granular details. This separation facilitates generating the local content in patches for images of varying sizes, all while being guided with global content that we derive from a reference image at the diffusion model pretraining resolution. This enables the synthesis of images at diverse resolutions while adhering to the diffusion model's initial training size and ensuring a global coherence on the entire image. To aid in this task, we introduce several techniques including an efficient patch fusion technique that enforces smooth boundaries, a novel guidance strategy to reduce image artifacts, and a global content resampling technique that amplifies the resolution of diffusion models up to 2X the training resolution.

Figure 1 shows a diverse array of images generated with ElasticDiffusion. Several of the examples presented include a single object and some of them were generated with extreme examples of aspect ratios, showcasing our method's ability to produce coherent images under various sizes. We conduct evaluations on the CelebA-HQ dataset consisting of images of faces from celebrities and the LAION-COCO benchmark containing a more diverse array of image content. We evaluate our results quantitatively using Fréchet Inception Distance scores [11] which measure perceptual quality and CLIP-score [10] to measure adherence to textual prompts. ElasticDiffusion consistently outperforms naive resizing of the latent space using the standard Stable Diffusion$_{1.4}$ model trained on images of size 512×512. More importantly, ElasticDiffusion obtains comparable FID (228.87 vs 230.21) and CLIP scores (26.07 vs 28.06) as SDXL when generating images at 1024×1024, the native resolution of SDXL, despite relying on Stable Diffusion$_{1.4}$. We also evaluate across aspect ratios and compare favorably to the previously proposed MultiDiffusion method [3].

2. Related Work

Diffusion Models have been widely adopted for their high-quality outputs in generative tasks [5, 6, 8, 14, 16, 27, 29, 31, 35, 42]. These models involve iterative decoding with many steps leading to high compute and memory requirements. Recent work addressed these issues by devising faster sampling strategies [21, 39], hierarchical models [13, 29, 35], progressive training at increasing resolutions [25, 27, 31], and two-stage models [27, 31, 34, 41]. Stable Diffusion (SD) [31] trains a variational auto-encoder to compress images into a low-dimensional 64×64 latent space and trains a diffusion model on this latent space. To enable generation at a higher resolution, models are initially trained at a 256×256 resolution, before fine-tuning them at 512×512 and 1024×1024 in the case of SDXL [27]. SD is one of the few large-scale diffusion models that released trained parameters, making it the building block for many subsequent work [4, 9, 32, 41]. However, these models, including SD, are confined to specific resolutions and do not generalize well to aspect ratios or resolutions unseen during training. Interestingly, despite being presented with a 256×256 resolution during their early training stages, both SD and SDXL fail to generate realistic images at this specific resolution after being fine-tuned for larger outputs. ElasticDiffusion enables high quality generation at unseen resolutions including re-enabling consistent high quality outputs for SD at 256×256.

Mixture of diffusers [46] and *MultiDiffusion* [3] generate panoramic images using a pre-trained diffusion model by generating overlapping crops and combining the generation signal of the overalpping regions. *SDXL* [27] and *Any-Size-Diffusion* [45] fine-tune a pre-trained SD on a fixed set of resolutions. In contrast to these methods, ElasticDiffusion extends a pre-trained SD to generate at various image sizes at a constant memory requirement and without additional training, all while ensuring global coherence.

Guided diffusion models devise strategies to enable image generation conditioned on text and other modalities [2, 2, 12, 18, 20, 24, 31, 44]. Classifier guidance [24] applies external supervision from a pre-trained classifier on noisy images to guide the generation process. Classifier-free guidance [12] eliminates the need for a pretrained classifier but requires training a conditional diffusion model, limiting its application to new modalities. Universal Guidance [2] applies the pre-trained classifier on the noise-free images produced by the DDIM [39] sampling process to eliminate the need of training the classifier on noisy images. StableSR [41] uses LoRA [15] to condition the generation process on a low-resolution version of an image, achieving impressive results on super-resolution. Inspired by this work, we propose Reduce-Resolution Guidance (Sec. 4) to guide the generation of high-resolution images from a low-resolution version, substantially reducing artifacts without additional training.

3. Background: Diffusion Models

A conditional diffusion model $\epsilon_\theta : \mathcal{X} \times \mathcal{C} \to \mathcal{X}$ predicts a less noisy version of the input image $x \in \mathbb{R}^{H \times W \times 3}$, conditioned on variable $c \in \mathbb{R}^D$ (e.g., a text embedding). Starting with $x_T \sim \mathcal{N}(0, \mathbf{I})$, the reverse diffusion process progressively denoise x_T over T steps to generate a realistic image x_0 that conforms to the input condition c through:

$$x_{t-1} = \epsilon_\theta(x_t, c) \qquad \text{for } t = T, T-1, \ldots, 1,$$

where ϵ_θ is the denoising network. Standard diffusion models operate in the pixel space, but others like Latent Diffusion Models [31] instead operate on a latent image encoding space to reduce memory footprints. A U-Net architecture is commonly used to implement the denoising network and is typically trained on images at a fixed resolution $H \times W$ [5, 31, 35]. The convolutional architecture of a U-Net allows for inputs and outputs of arbitrary spatial dimensions. In order to generate an image of a different size $\bar{H} \times \bar{W}$ at inference time, one can sample an initial noise variable $\bar{x}_T \in \mathbb{R}^{\bar{H} \times \bar{W} \times 3}$ and follow the same diffusion process. However, we find that this works poorly in practice, resulting in a significant degradation in output quality.

Denoising Diffusion Implicit Models (DDIMs) introduce a faster non-Markovian sampling strategy, bypassing denoising steps. The reverse diffusion step in DDIM is:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\frac{\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta^{(t)}(x_t)\right)}{\sqrt{\bar{\alpha}_t}}}_{\text{predicted } \hat{x}_0^t} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}}}_{\substack{\text{direction} \\ \text{pointing} \\ \text{to } x_t}}, \quad (1)$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} 1 - \beta_i$ is a cumulative product of the noise levels using a predetermined variance schedule β. $\hat{x}_0^t$ is a noise-free estimation of x_t obtained by subtracting the predicted noise scaled for step t. We assume a deterministic sampling process in the DDIM formula [39] for simplicity.

Classifier-Free Guidance updates the reverse diffusion process to condition it on a given condition c as:

$$\hat{\epsilon}_\theta^{(t)}(x_t) = \underbrace{\epsilon_\theta^{(t)}(x_t)}_{\substack{\text{unconditional} \\ \text{score}}} + (1+w) \cdot \underbrace{(\epsilon_\theta^{(t)}(x_t, c) - \epsilon_\theta^{(t)}(x_t))}_{\Delta_{\mathcal{C}}(x,c)\text{: class direction score}}. \quad (2)$$

$\epsilon_\theta(x, c)$ is a pretrained conditional diffusion model, and w is a scaling factor. The difference between the predicted $\epsilon_\theta(x, c)$ and unconditional $\epsilon_\theta(x)$ scores, denoted as *class direction score*, represents the guidance direction towards c.

4. ElasticDiffusion

The aim of this work is to develop a method capable of synthesizing images at arbitrary size $\bar{H} \times \bar{W}$, and conforming to a global condition c using a pre-trained diffusion model that is limited at inference to its training resolution $H \times W$. To achieve this, we use two key insights. First, the class direction score primarily dictates the image's overall composition, while the unconditional score enhances detail at the pixel level in a more local manner. Second, the unconditional score requires a pixel-specific precision, contributing to the image's fine-grained details, while class direction score affects pixels collectively, defining the image's overall composition. Leveraging these insights, we propose a method to decouple the generation of local and global content during the diffusion process. Specifically, we compute the unconditional score on local patches of size $H \times W$ while simultaneously resizing a class direction score, originally derived for a reference latent of the same dimensions. This dual strategy facilitates the generation of images at varied sizes using a pretrained diffusion model at its native resolution, all while maintaining the same memory requirement and without further training. We first present our approach for computing the unconditional score. We then detail our method to estimate and upscale the resolution of the class direction score. Finally, we combine the two estimated scores with a novel guidance strategy to generate images at arbitrary sizes.

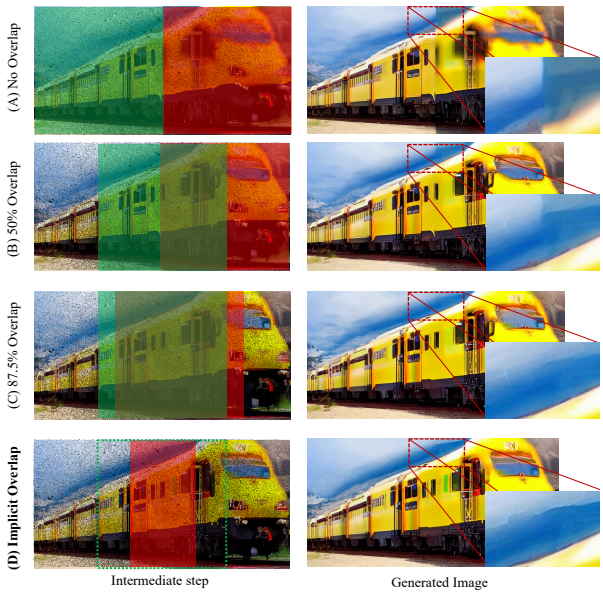


Figure 2. Comparison of various strategies for calculating diffusion model score on a local patch. No overlap between adjacent patches (A) leads to discontinuities at the boundaries. Strategies (B) and (C), explicitly overlap nearby patches, necessitating substantial overlap to be effective. Our implicit overlapping method (D) achieves superior results with computational demand similar to (B).

4.1. Estimating the Unconditional Score

Building upon previous work [3, 41, 46], we estimate the unconditional score for a large latent signal $\bar{x}_t \in \mathbb{R}^{\bar{H} \times \bar{W} \times 3}$ by concatenating scores derived from local patches. Specif-

ically, we divide image $\bar{x}_t$ to K patches $P_k \in \mathbb{R}^{H\times W\times 3}$ and compute the score as $\epsilon_\theta(\bar{x}_t) = \{\epsilon_\theta(P_k);\ \forall_{k\leq K}\}$. A common challenge encountered with this implementation is discontinuities at boundaries, as illustrated in Fig. 2 (A). To address this, earlier research calculated the diffusion model score on explicitly overlapping patches and averaged their scores in the intersecting regions [3, 41, 46]. While this strategy mitigates discontinuities, it requires large overlap between patches, thereby substantially increasing inference time and blurring details. To speed up the process, we introduce a more effective method that enhances boundary transitions in local patches by incorporating contextual information from adjacent patches, thus negating the need for signal averaging in overlapping areas. Specifically, we select patches smaller than the full size $p_k \in \mathbb{R}^{h\times w\times 3}$ with $h < H$ and $w < W$, and concatenate them with context pixels from adjacent patches, denoted as $c_k \in \mathbb{R}^{(H-h)\times(W-w)\times 3}$, to compute the diffusion model unconditional score $\mathbf{S}_u$ as:

$$\tilde{\epsilon}_\theta(x_t) = \{\,\epsilon_\theta(p_k \mid c_k) \mid \bar{x}_t = \{p_k;\ \forall_{k\leq K}\},\} \tag{3}$$

where p_i is a local patch and c_i are the context pixels surrounding it. This substantially increases the efficiency of the process. For instance, to generate an image of size 1024×1024, previous methods [3, 41, 46] used 87.5% overlap between adjacent patches, necessitating 81 forward diffusion calls per decoding step. In comparison, ElasticDiffusion achieves comparable results with only 9 forward calls, as demonstrated in Fig. 2 (D). Employing the same number of calls, previous techniques result in obvious boundary discontinuities, as depicted in Fig. 2 (B).

4.2. Estimating the Class Direction Score

A simple approach to estimate a class direction score of an intermediate latent signal $\bar{x}_t \in \mathbb{R}^{\bar{H}\times\bar{W}\times 3}$ is to upsample the score from a reference latent $x_t \in \mathbb{R}^{N\times M\times 3}$ to $\bar{H} \times \bar{W}$. This is possible due to our observation that the class direction score represents a latent direction that can be shared between nearby pixels. We validate this observation empirically in Supplementary. We choose $N < H$ *and* $M < W$ such that $N \times M$ is as close as possible to $H \times W$ and $\frac{\bar{H}}{\bar{W}} = \frac{N}{M}$. This is important to maintain the aspect ratio and prevent stretching the global content. Formally, we compute the class direction score $\mathbf{S}_d$ as:

$$\begin{aligned} \mathbf{x}_t &\leftarrow \text{Downsample}(\bar{x}_t, N \times M), \\ \Delta_{\mathcal{C}}(\bar{x}_t, c) &= \text{Upsample}(\Delta_{\mathcal{C}}(\mathbf{x}_t, c), \bar{H} \times \bar{W}), \end{aligned} \tag{4}$$

where $\Delta_{\mathcal{C}}(.)$ is the class direction score from Eq. 2, Downsample and Upsample are downsampling and upsampling operations. We use a nearest-neighbors approach to prevent altering the statistics of the latent signal. In order to maintain the input to the diffusion models at the size $H \times W$, we dynamically pad the downsampled latent $\mathbf{x}_t$ using a random background with a constant color. This encourages the model to concentrate content generation within the center area. We then crop the extended parts from the predicted noise to the target image resolution $N \times M$. Formally we modify the forward call for the reverse diffusion step as:

$$\begin{aligned} \hat{\mathbf{x}}_t &\leftarrow \text{Pad}\left(\mathbf{x}_t, \mathcal{A}_t\sqrt{\bar{\alpha}_t} + \sqrt{1-\bar{\alpha}_t}\cdot\mathcal{Z}_t\right), \\ \epsilon_\theta\left(\mathbf{x}_t, c\right) &= \text{Crop}\left(\epsilon_\theta\left(\hat{\mathbf{x}}_t, c\right), N \times M\right), \end{aligned} \tag{5}$$

where $\mathcal{Z}_t \sim \mathcal{N}(0, I)$ represents the injected Gaussian noise at each step, and $\mathcal{A}_t$ represents a background image of size $(H-N)\times(W-M)$ with a constant color value $\mathcal{Y}$ randomly sampled from a uniform distribution. This simple padding operation guarantees that the input to the diffusion model is kept at $H \times W$, while encouraging the diffusion model to keep the generated content within the cropped $N \times M$ center that we are interested in.

4.3. Refined Class Direction Score

Sharing the class direction score among nearby pixels can result in over-smoothed images. To mitigate this, we propose an iterative resampling technique that increases the resolution of the estimated class direction score by extrapolating missing image components from their surrounding context, following [23]. Our technique involves a gradual enhancement of the class direction score's resolution by estimating and integrating it for newly sampled pixels. Specifically, in each iteration, we replace $n\%$ of the pixels in $\mathbf{x}_t$ with newly sampled ones from x_t to get an updated version $\tilde{\mathbf{x}}_t$. Following each update, the direction score is recalculated and blended with the previously calculated score. Formally, we consider $\mathbf{S}_d^0 = \mathbf{S}_d$ and define the iterative resampling as:

$$\mathbf{S}_d^{r+1} = \mathbf{S}_d^r \odot m + \tilde{\mathbf{S}}_d \odot (1-m), \tag{6}$$

where $m \in \{0,1\}^{\bar{H}\times\bar{W}}$ is a mask with a value of 1 at the positions of the newly sampled pixels and 0 elsewhere. $\tilde{\mathbf{S}}_d$ represents the recalculated class direction score on $\tilde{\mathbf{x}}_t$ as per Eq. 4. This method estimates the class direction score of $n\%$ new pixels while retaining the information of the previously estimated score, thereby increasing the score's overall resolution. In our experiments, we set n to 20% and repeat the process R times, depending on the target resolution of the generated image. Fig. 4 demonstrates the effectiveness of our method in enhancing the details of the generated images.

4.4. Reduced-Resolution Guidance (RRG)

We effectively estimate the unconditional score signal and concurrently steer the image generation using the class direction score. However, inaccuracies in unconditional score estimation or fluctuations in local content, especially from distant patches, can lead to inconsistencies or emerging artifacts. To enhance image coherence and diminish artifacts, we consider a downsampled version of the latent at each decoding step as a reference and aim to align the decoded

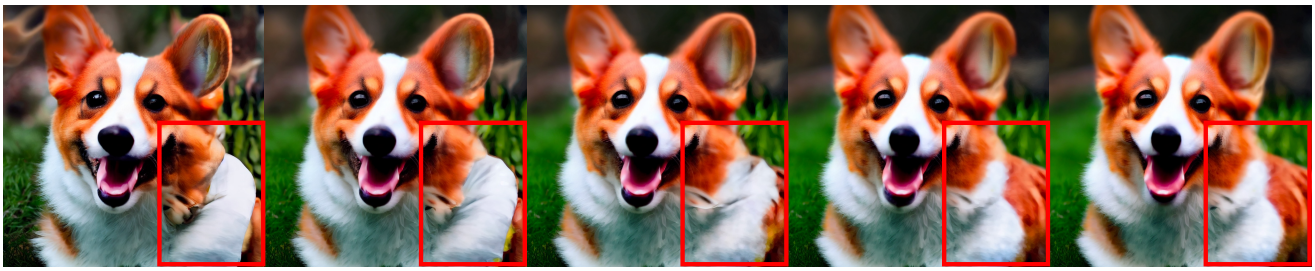


Figure 3. **The Effect of Reduced-Resolution Guidance (RRG).** Applying RRG at an increased weight effectively eliminatse emerging artifacts albeit at the cost of slightly blurrier outputs. $\delta = 200$ strikes a good balance. Improvements are more noticeable when zooming in.

Figure 4. **Effect of Resampling.** Applying more resampling steps noticeably enhances detail in the generated images.

Algorithm 1 Sampling Algorithm for Image at $\bar{H} \times \bar{W}$

Require:
ϵ_θ ▷ pre-trained DM at $H \times W$
c, w ▷ text condition and CFG weight
$\bar{x}_T \sim \mathcal{N}(0, I)$ ▷ noise at $\bar{H} \times \bar{W}$
for $t = T$ **down to** 1 **do**
  $\mathbf{x}_t \leftarrow \text{Downsample}(\bar{x}_t, N \times M)$
  $\mathbf{S}_c, \hat{\mathbf{x}}_t \leftarrow \text{Pad-and-Crop}\,(\epsilon_\theta, \mathbf{x}_t, c, N, M)$ ▷ Eqs. 5
  $\mathbf{S}_u \leftarrow \tilde{\epsilon}_\theta(\hat{\mathbf{x}}_t)$ ▷ Eq. 3
  $\bar{\mathbf{S}}_d^0 \leftarrow \text{Upsample}(\mathbf{S_c} - \mathbf{S_u}, \bar{H} \times \bar{W})$
  for all $r = 1, \ldots, R$ **do**
    $\bar{\mathbf{S}}_d^r \leftarrow \text{Resample}(\bar{\mathbf{S}}_d^{r-1}, \bar{x}_t)$ ▷ Eq. 6
  end for
  $\bar{\mathbf{S}}_\mathbf{u} \Leftarrow \tilde{\epsilon}_\theta(\bar{x}_t)$ ▷ Eq. 3
  $\bar{x}_{t-1} \leftarrow \bar{\mathbf{S}}_u + (1 + w) \cdot \bar{\mathbf{S}}_d^R$ ▷ diffusion update
  $\bar{x}_{t-1} \leftarrow \bar{x}_{t-1} - \text{RRG}(\bar{x}_t, \mathbf{x}_t)$ ▷ Eq. 7
end for
return $\bar{x}_0$

latent with it through our reduced-resolution latent update strategy. Specifically, we utilize the noise-free sample $\hat{x}_0^t$ of the latent x_t from Eq. 1 and generate a corresponding noise-free sample $\hat{\mathbf{x}}_0^t$ from its downsampled counterpart $\mathbf{x}^t$ in Eq. 4 as:

$$\hat{\mathbf{x}}_0^t = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t} \cdot (\epsilon_\theta(\mathbf{x}_t) + (1+w) \cdot \Delta_\mathcal{C}(\mathbf{x}_t, c))),$$

Here, both $\hat{x}_0^t$ and $\hat{\mathbf{x}}_t^0$ corresponds to the same latent update at different resolutions. Due to its smaller dimension, $\hat{\mathbf{x}}_t^0$ has a broader context when computing the unconditional signal. To guide x_t with its downsampled reference, we refine the latent x_t with the direction that minimizes $L2$ difference between $\hat{\mathbf{x}}_t^0$ and $\hat{x}_0^t$. Formally,

$$\bar{x}_{t-1} \leftarrow \bar{x}_{t-1} - \delta_t \nabla_{x_t} ||\text{Upsample}(\hat{\mathbf{x}}_t^0, \bar{H} \times \bar{W}) - \hat{x}_0^t||, \quad (7)$$

where δ_t represent the weight of the guidance. Since the overall image structure is determined in the early diffusion steps, we start with $\delta_t = 200$ and linearly decrease this weight until 60% for the diffusion steps are completed. This scheduling strategy mitigates potential quality degradation from matching the generated image with a lower-resolution version while allowing the model to fill-in higher-resolution details in the later decoding stages. Fig. 3 illustrates how RRG eliminates emerging artifacts. Our full algorithm for generating images at an arbitrary size is summarized in Alg. 1

5. Experiments

ElasticDiffusion facilitates image generation across diverse resolutions and aspect ratios. Given that pre-trained Stable Diffusion models are trained on square images of fixed size, we organize our experiments into two parts: (1) generation of square images at multiple resolutions, and (2) generation of images with varying aspect ratios and resolutions.

Datasets. We evaluate the generation of square images on the Multi-Modal CelebAHQ dataset [43], which includes high-resolution square face images accompanied with text descriptions. We evaluate different aspect ratio generation using the LAION-COCO dataset [37], derived from the web-crawled LAION-5B dataset [36], which includes a variety of image aspect ratios and contains landscapes, people, objects, and everyday scenes, each paired with a synthetic caption generated using BLIP [22]. We consider four common aspect ratios: 9:16, 16:9, 3:4 and 4:3.

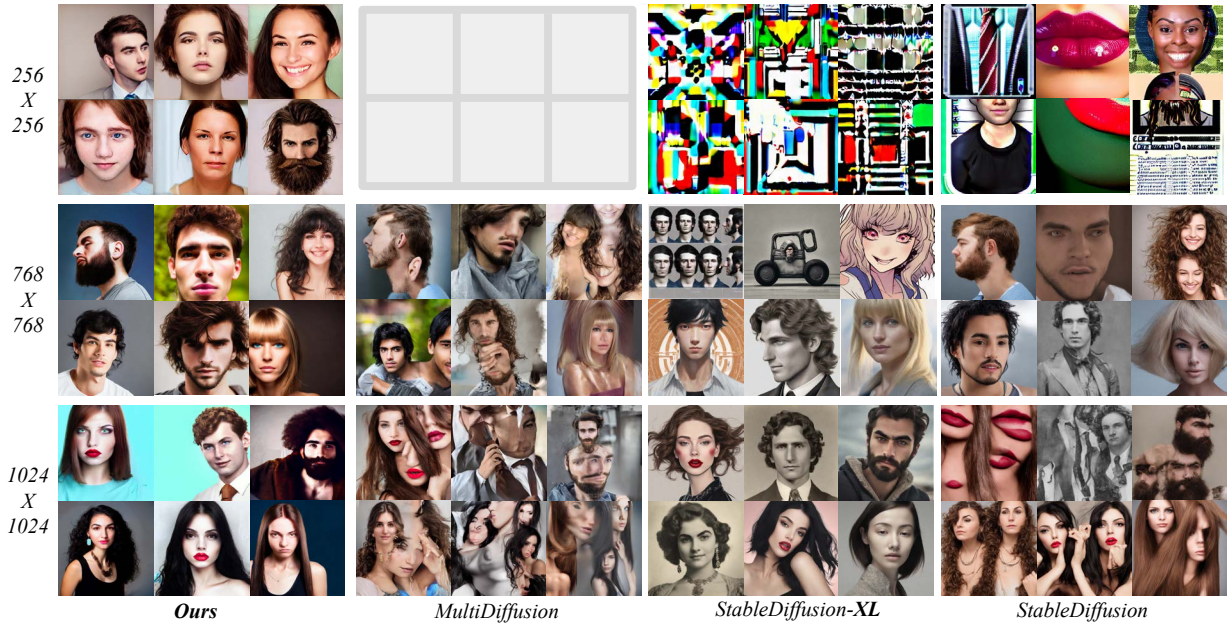


Figure 5. **Qualitative comparison at various resolutions on CelebAHQ faces.** *ElasticDiffusion* consistently generates coherent images at all resolutions. *StableDiffusion*, and *MultiDiffusion* produce repeating body parts at higher resolutions, while *StableDiffusion-XL* fails to maintain its quality at lower resolutions, resulting in noisy outputs at 256×256. We exclude *MultiDiffusion* results at 256×256 as it is not designed to produce images at lower resolutions.

Evaluation Metrics. Following prior text-to-image synthesis works [3, 25, 27, 30, 35], we use automatic evaluation metrics *Frechet Inception Distance (FID)* and *CLIP-score*. *FID* [11] measures both the realism and diversity of the generated images by calculating the difference between features of the real and generated images computed using Inception-V3 [40] pretrained on ImageNet [33]. *CLIP-score* uses a pretrained text-image CLIP model [28] to measure alignment between the generated images and input prompts. We compute all evaluation metric values using 10,000 images.

Baselines. We compare our approach against prior diffusion model generation methods, specifically focusing on *Stable Diffusion (SD)* and *MultiDiffusion (MD)*. *SD* follows the standard reverse diffusion process on an image latent space. *MultiDiffusion* uses a pretrained *SD* model for panoramic image generation, by creating smaller, overlapping patches and averaging the Diffusion Model scores in intersecting areas. For baseline comparisons, we fix the pre-trained diffusion model to the $SD_{1.4}$ version of Stable Diffusion, which is trained to generate images at a resolution of 512×512. Additionally, we compare our method with *SDXL*, an enhanced Stable Diffusion model three times larger than the standard one. SDXL is trained at a higher resolution of $1024p$, and fine-tuned on a specific set of aspect ratios with pixel sums close to 1024^2. We exclude the latent refinement module in *SDXL* to focus our analysis on the base model. Throughout our experiments, we employ a DDIM sampling strategy with 50 steps and use 7.0 for the classifier-free guidance scaling factor. We also ensure consistency in seeds and captions across all baseline comparisons for a fair evaluation.

5.1. Qualitative Results

We show square image generation samples in Fig. 5 generated from the CelebAHQ dataset at various resolutions. Both *MultiDiffusion* and *Stable Diffusion* have a tendency to replicate textures and body parts at higher resolutions, resulting in images that lack coherence. At resolutions lower than its training resolution 512×512, *StableDiffusion* aligns poorly with the provided captions and produces unappealing images. *SDXL*, trained for 1024×1024 resolution, has a similar trend of reduced perceptual quality at lower sizes, eventually producing complete noise at a resolution of 256×256. In contrast, ElasticDiffusion maintains image coherence across all tested resolutions, and yields results comparable to *SDXL* at 1024×1024, despite using a less powerful base model and lower memory. We excluded results at 512×512 because both our method and *MultiDiffusion* generate same outputs as the base *Stable Diffusion* model. We also omitted the results of *MultiDiffusion* at 256×256 since it is not designed to generate images at sizes smaller than the base model. Fig. 6 presents generated samples at various aspect ratios. We observe a similar trend of pattern repetition and reduced perceptual quality for images generated by the baselines, in contrast to those generated by ElasticDiffusion.

Finally, we apply our method using *SDXL* as the base model to enable Full-HD image generation (1920×1080).

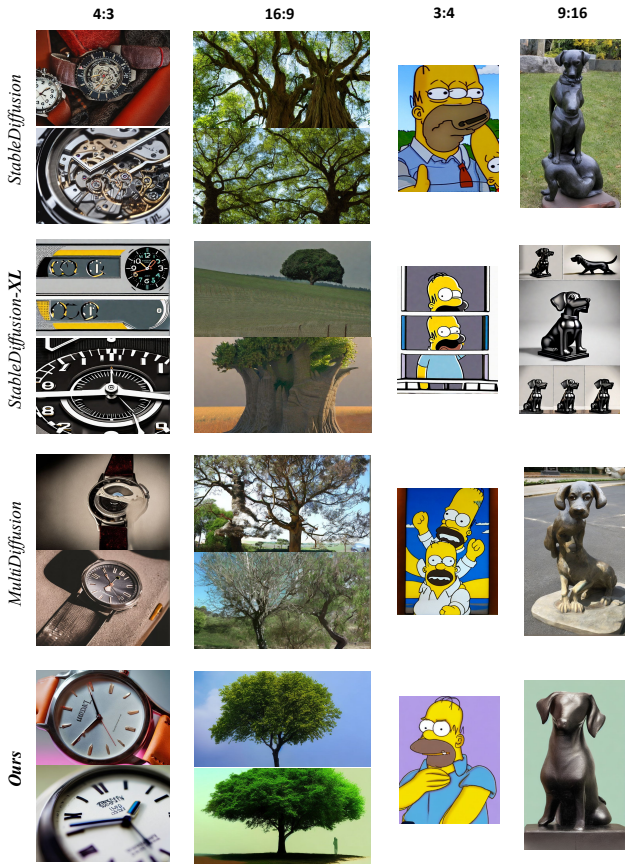


Figure 6. **Qualitative comparison on various aspect ratios.** *ElasticDiffusion* effectively handles a variety of aspect ratios. In comparison, *SD* and *MultiDiffusion* generate unrealistic images, while *SDXL* outputs exhibit a decline in the perceptual quality.

We show an example of our approach vs. *SDXL* in Fig. 7. While our method successfully generates coherent images at Full-HD resolution, *SDXL* struggles with texture and part repetition, despite being fined-tuned on 40 different aspect ratios. This is because *SDXL* was restricted to resolution with pixel counts near 1024×1024 during fine-tuning stages.

5.2. Quantitative Results

We provide quantitative evaluations in Table 1. *StableDiffusion* shows increasing image quality degradation, as indicated by the *FID* metric when processing images of sizes different from its training resolution 512×512. *MultiDiffusion* slightly improves *FID* at the expense of a substantially more forward calls. *SDXL* demonstrates similar declines in quality at resolutions far from its fine-tuning size of 1024×1024. ElasticDiffusion, however, improves *FID* while maintaining a comparable *CLIP-score* to the base model. ElasticDiffusion also significantly improves the performance for lower resolutions 256×256, while achieving comparable results to *SDXL* at its training resolution 1024×1024, with only $\sim 31\%$ of the memory requirement for *SDXL*.

Table 1. Quantitative comparison of on CelebA-HQ and LAION-COCO at different resolution. We indicate the best performances in **bold**, and underline the second best ones. $\#Calls$ represent the number of diffusion model calls required at each decoding steps.

Res.	Method	CelebA-HQ		LAION-COCO		#Calls	Mem.
		FID ↓	CLIP ↑	FID ↓	CLIP ↑		
256 × 256	$SD_{1.4}$	258.43	20.14	54.06	21.43	2	7.2GB
	SDXL	368.06	14.40	175.87	14.60	2	18.5GB
	Ours$_{1.4}$	**235.23**	**23.88**	**23.77**	**26.30**	2	8.6GB
512 × 512	$SD_{1.4}$	233.40	24.00	20.50	27.33	2	8.6GB
	SDXL	240.20	21.57	42.58	25.34	2	21.6GB
768 × 768	$SD_{1.4}$	238.87	23.45	29.89	27.01	2	11.1GB
	$MD_{1.4}$	240.56	22.82	29.98	27.31	50	8.6GB
	SDXL	**225.48**	24.23	**23.31**	**27.88**	2	24.5GB
	Ours$_{1.4}$	225.86	**26.66**	25.78	25.93	17	8.6GB
1024 × 1024	$SD_{1.4}$	266.01	21.90	47.01	25.70	1	14.7GB
	$MD_{1.4}$	264.57	21.55	37.70	26.96	162	8.6GB
	SDXL	230.21	**24.62**	**25.58**	**28.06**	1	27.5GB
	Ours$_{1.4}$	**228.87**	23.74	27.76	26.07	33	8.6GB

We also provide an evaluation of our method on a variety of aspect ratios and resolutions for both landscape and portrait images in Table 2. *StableDiffusion* exhibits a consistently worse *FID* score with increasing resolutions, while paradoxically maintaining or even enhancing its *CLIP-score*. We believe that this is because *CLIP-score* quantifies the agreement between the input prompt and the generated image, and *StableDiffusion* benefits from the repetitive textures and element artifacts it generates which align closely with the prompt. For example, a photo of repeated lipsticks might align more with the caption *"lipstick"* despite its poor composition. *MultiDiffusion* generally enhances image quality, though it does not achieve satisfactory performance. *MultiDiffusion* struggles with objects that span the entire image (e.g. Fig 6). Our method consistently improves the *FID* metric over the baseline model on landscape resolutions and most portrait ones while preserving fidelity to the input prompts, thereby attaining comparable or superior *CLIP-scores*. Remarkably, even with a larger model size and explicit fine-tuning at a similar resolution of 768×1280, *SDXL* only marginally surpasses our method at 768×1024 resolution.

5.3. Ablation study

Table 3 presents results of our method when excluding key components, demonstrating that each element improves performance. Fig. 2 illustrates the effectiveness of our proposed

Table 2. **Quantitative comparison of on LAION-COCO datasets at various aspect ratios (A) and resolutions (R).** We indicate the best performances in **bold**, and underline the second best. Portrait means the resolution is transposed from H:W to W:H.

A	R	Method	Landscape		Portrait	
			FID ↓	CLIP↑	FID ↓	CLIP↑
3:4	384 × 512	$SD_{1.4}$	38.86	24.63	17.66	26.54
		$MD_{1.4}$	–	–	–	–
		SDXL	104.84	21.33	68.84	22.40
		$Ours_{1.4}$	**35.10**	**24.91**	**15.50**	**27.33**
	512 × 680	$SD_{1.4}$	45.54	**24.96**	**16.81**	**27.33**
		$MD_{1.4}$	43.44	24.56	19.13	26.80
		SDXL	62.80	24.20	28.23	26.17
		$Ours_{1.4}$	**41.06**	24.40	18.90	26.83
	768 × 1024	$SD_{1.4}$	71.00	24.09	28.83	26.30
		$MD_{1.4}$	54.89	25.02	26.35	26.95
		SDXL	47.05	**25.79**	**19.50**	**27.41**
		$Ours_{1.4}$	**47.03**	24.91	22.52	25.80
9:16	288 × 512	$SD_{1.4}$	23.50	24.69	24.01	24.89
		$MD_{1.4}$	–	–	–	–
		SDXL	121.83	17.65	112.41	18.54
		$Ours_{1.4}$	**23.23**	**25.26**	**22.86**	**26.30**
	512 × 904	$SD_{1.4}$	29.86	25.34	27.45	26.01
		$MD_{1.4}$	26.35	**25.70**	26.70	25.28
		SDXL	29.60	25.40	27.27	26.08
		$Ours_{1.4}$	**22.85**	25.01	**26.68**	**26.12**

Table 3. **Ablation analysis of ElasticDiffusion** on 500 images.

Model Details	FID ↓	CLIP ↑
ElasticDiffusion	133.67	25.82
w/o *Resampling*	150.01	23.82
w/o *RRG*	150.64	24.34
w/o *Imp. overlap*, w/ exp. overlap	141.42	25.82

implicit overlap method in resolving boundary discontinuities at a reduced computational cost. Fig. 3 highlights the effectiveness of *Reduced-Resolution Guidance* in removing emerging artifacts. Fig. 4 shows the effectiveness of our iterative resampling technique in enhancing image details.

6. Discussion and Conclusion

Experimental results highlight the adaptability and effectiveness of ElasticDiffusion at steering diffusion models to produce an array of resolutions and aspect ratios. ElasticDiffusion requires no fine-tuning, consumes a constant memory footprint, enables both increased and reduced resolutions,

Figure 7. **Results at Full-HD.** Our method enables *SDXL* to produce coherent Full-HD images (1920 × 1080). In contrast to its standard process which produces repeated textures and elements, ElasticDiffusion applies global guidance from a lower resolution to enforce global structure.

and can generate a variety of aspect ratios.

ElasticDiffusion, however, does have several practical limitations. First, inaccuracies in the estimation of the global and local signals may result in artifacts. Although we attempt to mitigate artifacts with our *Reduced-resolution guidance*, it can still generate blurrier outputs. Second, since the global content guidance is initially estimated at the original training resolution of the underlying diffusion model, our method is less effective in generating images at significantly *extended* sizes beyond the training resolution. In particular, at extreme resolutions beyond 2X, our method produces artifacts and images of a lower perceptual quality. We provide examples of these failure cases in the supplementary materials.

The main insight underpinning our method is a novel reinterpretation of *classifier-free guidance* in somewhat disentangling both global and local content. Our comprehensive evaluations demonstrate the feasibility of disentangling these signals, yet the full extent of their separation offers an avenue for further exploration in this direction. We hope that our findings inspire future work in investigating the separation of global and local content guidance signals for image synthesis. This separation holds potential for various applications such as selectively manipulating local or global content or enhancing image style transfer by decoupling style from content generation in diffusion models.

References

[1] DeepFloyd. https://www.deepfloyd.ai/. Accessed: 2023. 12

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models, 2023. 2

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023. 2, 3, 4, 6

[4] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, 2023. 2

[5] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2, 3

[6] Jiatao Gu, Qingzhe Gao, Shuangfei Zhai, Baoquan Chen, Lingjie Liu, and Josh Susskind. Control3diff: Learning controllable 3d diffusion models from single-view images. In *3DV24*, 2023. 2

[7] Moayed Haji-Ali, Andrew Bond, Tolga Birdal, Duygu Ceylan, Levent Karacan, Erkut Erdem, and Aykut Erdem. Vidstyleode: Disentangled video editing via stylegan and neuralodes. In *ICCV*, 2023. 11

[8] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation, 2023. 2

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2022. 2

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. 2

[11] 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. In *NIPS*, 2017. 2, 6

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 2

[13] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *NeurIPS*, 2022. 2

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[16] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *ICML*, 2023. 2

[17] Hugging Face. Stable diffusion xl. https://huggingface.co/docs/diffusers/main/en/using-diffusers/sdxl, 2023. Accessed: [Insert Date Here]. 11

[18] Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *ICCV*, 2023. 2

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 11

[20] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 2

[21] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. In *ICML*, 2021. 2

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 5

[23] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 4

[24] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[25] 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. In *ICML*, 2022. 2, 6, 11

[26] Suraj Patil, Pedro Cuenca, Nathan Lambert, and Patrick von Platen. Stable diffusion with diffusers. *Hugging Face Blog*, 2022. https://huggingface.co/blog/stable_diffusion. 11

[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, 2023. 1, 2, 6, 11

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 6

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3

[32] 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. 2

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 6

[34] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021. 2

[35] 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, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2, 3, 6, 12

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 5

[37] Christoph Schuhmann, Andreas A. Köpf, Richard Vencu, Theo Coombes, and Ross Beaumont. Laion coco: 600m synthetic captions from laion2b-en, 2023. 5

[38] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *ICCV*, 2019. 11

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2015. 6

[41] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution, 2023. 2, 3, 4

[42] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023. 2

[43] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation, 2021. 5

[44] Guy Yariv, Itai Gat, Lior Wolf, Yossi Adi, and Idan Schwartz. Audiotoken: Adaptation of text-conditioned diffusion models for audio-to-image generation, 2023. 2

[45] Qingping Zheng, Yuanfan Guo, Jiankang Deng, Jianhua Han, Ying Li, Songcen Xu, and Hang Xu. Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images, 2023. 1, 2, 11

[46] Álvaro Barbero Jiménez. Mixture of diffusers for scene composition and high resolution image generation, 2023. 2, 3, 4

Appendices

In this supplementary document, we extended the discussion on existing diffusion models generation processes, highlighting their constraints in adapting to diverse image sizes and the potential for separating global and local content generation. We also provide further qualitative comparisons with baselines using the DrawBench benchmark [25] at various resolutions including full-HD. Our code can be accessed at https://github.com/MoayedHajiAli/ElasticDiffusion-official.git

A. Discussion on Diffusion Models

In this section, we discuss the generation process of diffusion models, focusing on their performance across various image sizes and our analysis of their capacity to separate global and local content generation.

A.1. Diffusion Models Adaptability Across Sizes

Pretrained diffusion models such as *StableDiffusion*$_{1.4}$ are technically capable of handling various image sizes. Accordingly, the official implementation provides parameters for adjusting the size of the generated images. However, our experiments show a significant decline in image quality when these models operate at resolutions outside those seen during training. These observations are confirmed in the Stable Diffusion blog post on Hugging Face which warns that deviating from the trained resolution may compromise image quality [26]. Specifically, it notes that going below the training resolution results in lower quality images, while exceeding it in both the height and width directions causes repetitive image areas, leading to a loss of global coherence. Similar findings were noted in the *StableDiffusion-XL* official blog post [17].

In Fig. 8, we qualitatively analyze how generating images larger than the training resolution impacts image coherence. We generate these results using *StableDiffusion*$_{1.4}$ which was pretrained on 512×512 images. For smaller dimensions, the model tends to stretch the generated objects, whereas for larger dimensions, such as 1024×1024, it often creates repetitive elements. Notice the stretch in the cat and lion faces in the third and fourth columns. Additionally, observe how artifacts and repetition regarding nose and eye parts tend to happen more frequently as we increase the resolution.

Notably, the model maintains its output quality within a narrow margin of 64 pixels from its training resolution, suggesting a limit to the generalization capabilities of diffusion models with respect to various image sizes. This observation also shows the potential limitations of the solutions based only on a fine-tuning process for a fixed set of aspect ratios such as those proposed in prior work [27, 45].

A.2. Global and Local Content Generation

In the domain of generative adversarial networks (GANs), the disentanglement style and content in the synthesized images has been widely explored, paving the way for advancements in diverse generation and editing applications [7, 19, 38]. However, the precise definitions of 'style' and 'content' remain fluid, with no consensus on the definition in the literature. Previous works often define the content and style based on manually pre-defined attributes. In this work, we opt to avoid such ambiguity by denoting the overall composition of the image as global content and the fine-grained details as local content. Subsequently, we conceive ElasticDiffusion based on two key insights: First, the *class direction score* (Eq. 2 in the main paper) collectively influences pixels to shape the overall composition of the generated image, denoted as global content. This global score can be effectively shared among neighboring pixels. Fig. 9 demonstrates that sharing the *class direction score* between nearby pixels maintains the global content and coherence of the generated image, although increasing the sharing extent decreases the perceptual quality. In contrast, the *unconditional score* requires pixel-level precision and it may not be feasible to share it between nearby pixels, as illustrated in Fig. 10. Second, the *unconditional score* dictates the fine-grained details of the generated image, denoted as local content. This suggests that the score can be computed effectively on localized regions without necessitating global information from the entire image. Fig. 11 shows that computing the *unconditional score* in localized regions, corresponding to the size 512×512 of the generated image, produces similar results to those obtained when computing the score globally.

B. Analysis of ElasticDiffusion

This section provides a comprehensive analysis of ElasticDiffusion, focusing on its application to pixel-based diffusion models and comparisons with baseline methods. We present further qualitative results to showcase ElasticDiffusion's effectiveness in enhancing the coherence of the generated image across various sizes. We finally discuss the limitations and failure cases of our method.

B.1. Generalization to Pixel-Based Diffusion Models

We apply ElasticDiffusion to a pre-trained *DeepFloyd-IF-XL-V1.0* model, which operates on the pixel-space [1]. In the first stage, *DeepFloyd-XL-V1.0* generates images at a 64×64 resolution, which are then upscaled to 256×256 and subsequently to 1024×1024 in later stages. To assess the generalization capabilities of our method, we only focus on the first stage which generates images at a low resolution. As illustrated in Fig. 12, *DeepFloyd-XL-V1.0* shows similar limitations as latent diffusion models when dealing with various resolutions, primarily characterized by repetitive elements and reduced image coherence. However, we demonstrate the effectiveness of ElasticDiffusion in enhancing the ability of the pre-trained model to handle diverse resolutions and aspect ratios by successfully generating well-structured images. This shows the applicability of our proposed generation process to diffusion models that operate on the pixel space.

B.2. Additional Qualitative Results

We provide further qualitative analysis of ElasticDiffusion.

Figure 13 provides a comparison with *StableDiffusion* and *MultiDiffusion* using selected DrawBench [35] prompts for landscape images at resolution 680×512. We highlight the tendency of baseline methods to generate repetitive elements. This not only disrupts the image's overall coherence but also makes the baselines struggle to accurately reflect object counts. For example, in the first row, both baselines produced multiple dogs for an input prompt '*one* dog on the street'. In contrast, our method effectively aligns with the given prompt, generating a single, coherent dog.

Figure 14 shows a similar analysis on portrait images of resolution 512×680. We observe a similar limitation in baselines such as element and texture repetition in the generated images. This tendency of repeating elements particularly affects the model's capacity to create coherent objects that share textures with the background. For example, In the first row, the baselines struggle to accurately depict a hamburger, whereas our method successfully generates a coherent hamburger that is separated from the background. This limitation also affects the baseline models' ability to render objects with repeating patterns, like a 'cube made of bricks' shown in the last row. Moreover, the baseline behavior of repeating patterns especially escalates when generating a single object across the majority of the image, as observed in the 4^{th} and 5^{th} rows. In contrast, our method is able to maintain image coherence across various settings.

Figures 15 and 16 focus on showing results for the generation of Full-HD landscape images using *StableDiffusion-XL* as the base model. We compare against the standard decoding process of *StableDiffusion-XL* on sampled DrawBench prompts from the Reddit Category and observe a significant improvement in image coherence when applying our method. Notice in Fig. 15 how *StableDiffusion-XL* stretch the car in the first example, or repeat the limbs of the corgi and the lion (in the 2^{nd} and 3^{rd} example). In comparison, our method successfully coherently generates the requested image without any such distortions, all while utilizing lower memory requirements.

Figures 17 and 18 provide a similar analysis for Full-HD portrait images. *StableDiffusion-XL* produces significantly distorted images which either contain incorrectly repeating elements as seen in the cat and the man faces in the first two examples of Fig. 17, or stretched parts like the woman face in the first example of Fig. 18. In contrast, our method generates detailed objects that fit the portrait aspect ratio while avoiding any stretching or element repetition.

B.3. Limitations

Fig. 19 illustrates the limitations of ElasticDiffusion in various scenarios. Our method retains some limitations of the pre-trained base model, including challenges with text-image alignment for complex prompts and occasional occurrence of artifacts. Additionally, we observe an increase in image blurriness with the application of larger *Reduced-Resolution Guidance* weights (Sec. 4.4 of the main paper). Moreover, while infrequent, there are instances where the constant-color background inadvertently blends into the generated image (as discussed in Sec. 4.2 of the main paper).

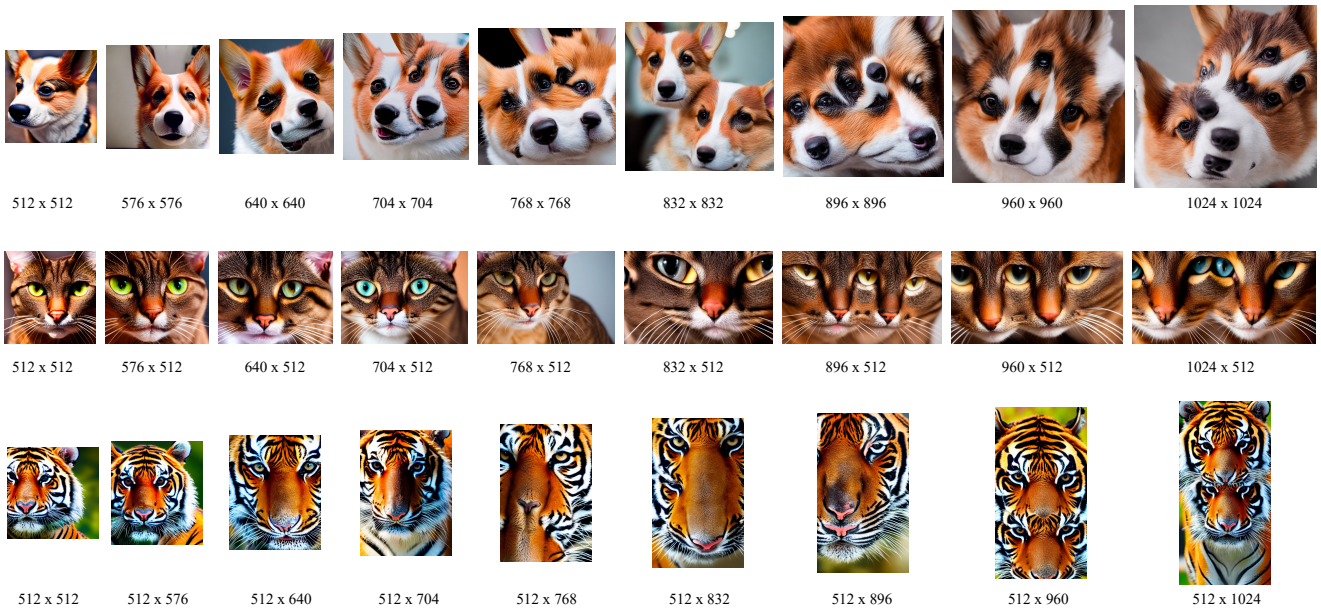


Figure 8. **Degradation of image quality in StableDiffusion$_{1.4}$ with varying resolutions.** We illustrate the progressive decrease in image quality as the resolution deviates from the model's training size of 512×512. The results on square, landscape, and portrait images show a significant reduction in global coherence for image sizes beyond 64 pixels from the training resolution.

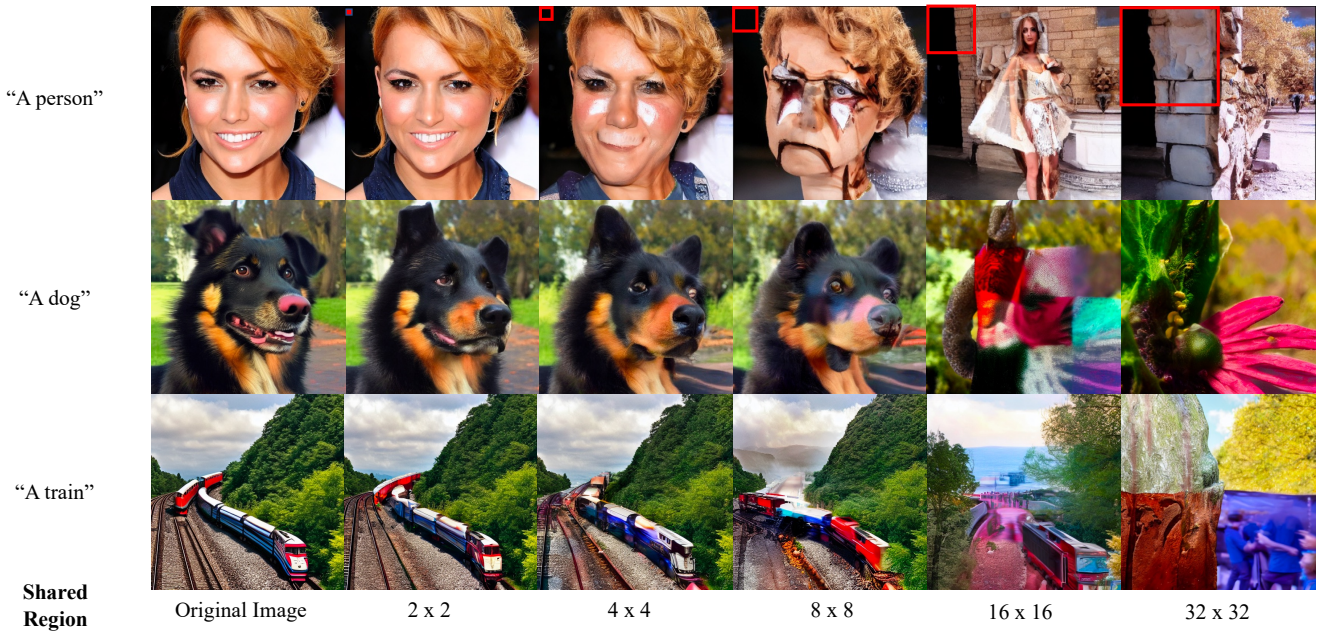


Figure 9. **Effect of sharing *class direction score* between nearby pixels.** We highlight that sharing the score within a group of neighboring pixels preserves the global content and coherence of the image, despite a reduction in the perceptual quality when selecting larger group sizes (as denoted by the red square). This supports our assumption that this score tends to be similar among neighboring pixels. To conduct this experiment, we downsample the *class direction score* of size 64×64 by a factor $N \times M$ (as specified in the last row) and upsample it back to 64×64, thus sharing the score for each $N \times M$ region. Note that as our experiments utilize a latent diffusion model, sharing the score within an $N \times M$ latent pixels during the decoding process impacts $8N \times 8M$ pixels of the final generated image.

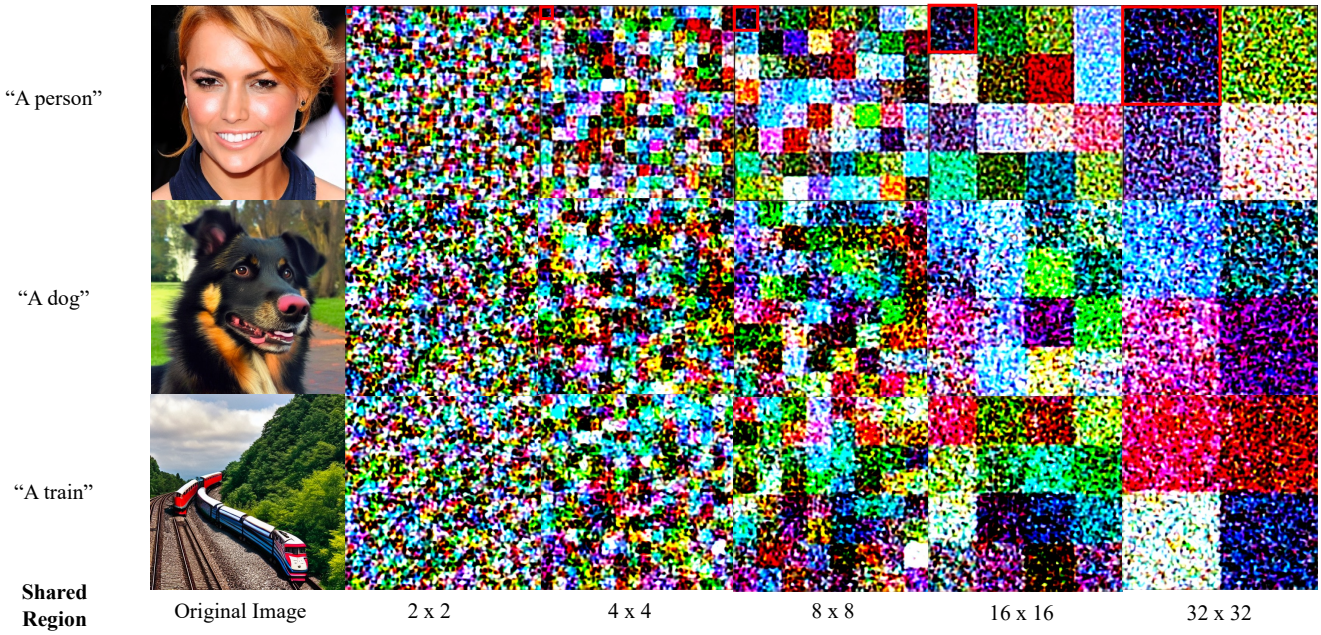


Figure 10. **Effect of sharing *unconditional score* between nearby pixels.** We show that sharing the unconditional score, even in small groups of pixels, leads to the generation of complete noise. This indicates that *uncondtional score*, in contrast to the *class direction score*, requires pixel-level precision to generate local details.

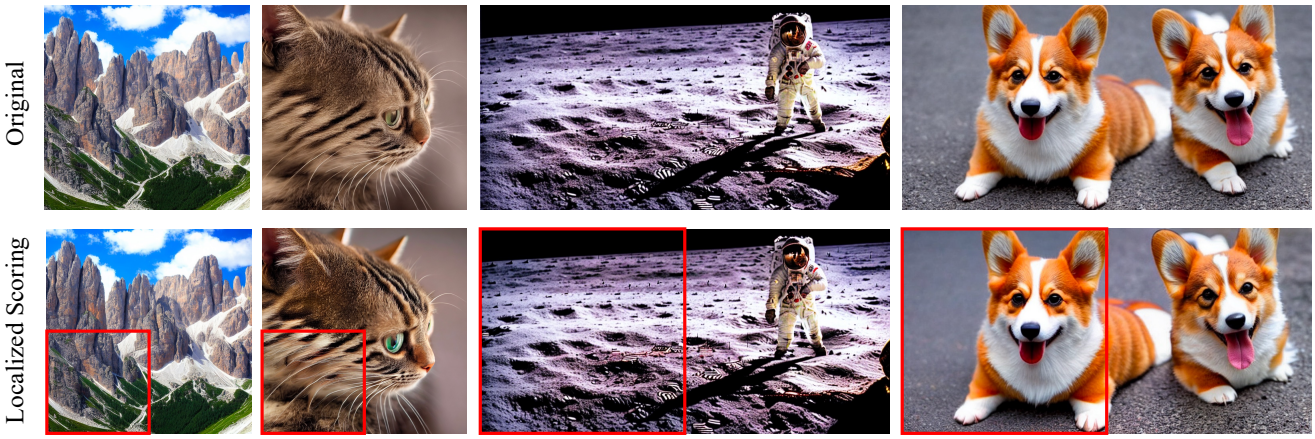


Figure 11. **Unconditional score computation on localized regions.** We show that computing the diffusion model *unconditional score* on local patches (corresponding to the size of the red boxes in the second row) results in images that are visually similar to those produced by a global score computation (displayed in the first row).

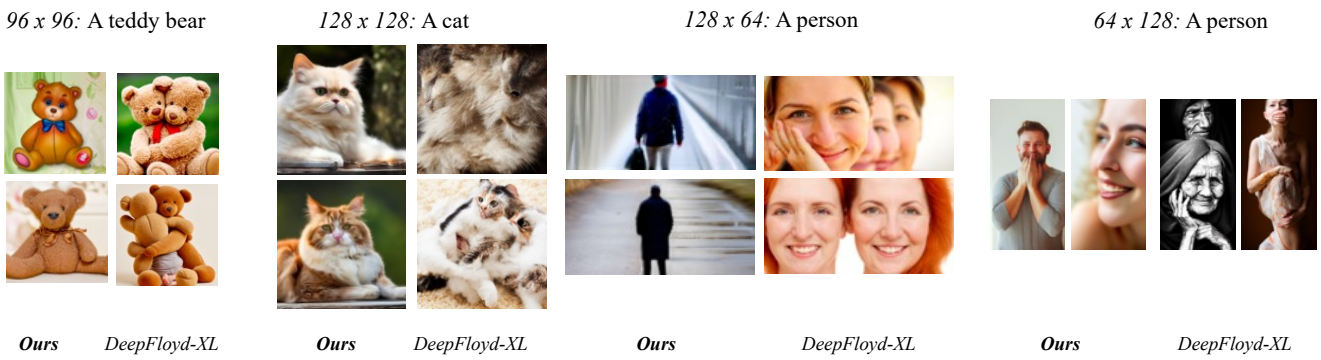
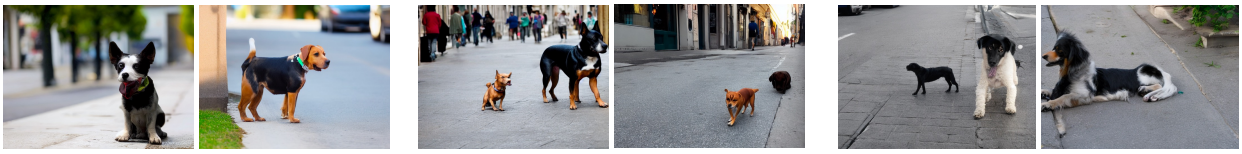


Figure 12. ***DeepFloyd-XL* across various image sizes.** We test *DeepFloyd-XL*, a diffusion model that operates on the pixel space, across multiple image resolutions. We observe a degradation in performance similar to that seen in $StableDiffusion$. The application of ElasticDiffusion significantly improves the overall composition of the generated images.

One dog on the street.



A fisheye lens view of a turtle sitting in a forest.

One cat and one dog sitting on the grass.

Illustration of a mouse using a mushroom as an umbrella.

In late afternoon in January in New England, a man stands in the shadow of a maple tree.

A large keyboard musical instrument with a wooden case enclosing a soundboard and metal strings, which are struck by hammers when the keys are depressed.

Ours *StableDiffusion* *MultiDiffusion*

Figure 13. **Additional qualitative comparison on landscape images using selected DrawBench prompts.** We use $SD_{1.4}$ as a base model for our method, *StableDiffusion*, and *MultiDiffusion* and generate images at resolution 680×512. Images produced by baselines display reduced alignment to the input prompt (1^{st} and 3^{rd} rows) and repeated elements (4^{th} and 6^{th} rows). In comparison, our method displays superior image coherence and faithfulness to the input prompts.

Figure 14. **Additional qualitative comparison on portrait images using selected DrawBench prompts.** We use $SD_{1.4}$ base model for our method, *StableDiffusion*, and *MultiDiffusion* and generate at the resolution 512×680. Similar to the landscape images, baseline methods exhibit several limitations such as poor text-image alignment (1^{st} and 2^{nd} rows), repeated elements (3^{rd}, 4^{th} and 5^{th} rows), and generated artifacts (6^{th} and 7^{th} rows). In comparison, our method consistently maintains better image coherence and fidelity to the input prompts.

Figure 15. **Additional qualitative comparison with SDXL on Full-HD landscape images**. We use randomly sampled DrawBench prompts from the Reddit Category. Despite its fine-tuning process, *StabelDiffusion-XL* produces images with repetitive textures and elements in full-HD resolution. Our method achieves a more cohesive composition while maintaining a comparable level of details, all while requiring less memory.

Figure 16. **Additional qualitative comparison with SDXL on Full-HD landscape images**. We use randomly sampled DrawBench prompts from the Reddit Category. *StabelDiffusion-XL* produce images that tend to repeat body parts and texture in full-HD resolution. In comparison, our method achieves better image coherence and maintains a similar perceptual quality, all while requiring less memory.

Figure 17. **Additional qualitative comparison with SDXL on Full-HD portrait images**. We use randomly sampled DrawBench prompts from the Reddit Category. Similar to landscape images, *StabelDiffusion-XL* produce repeated elements in full-HD resolution. In comparison, our method achieves more coherent images and generates content that fits the frame aspect ratio.

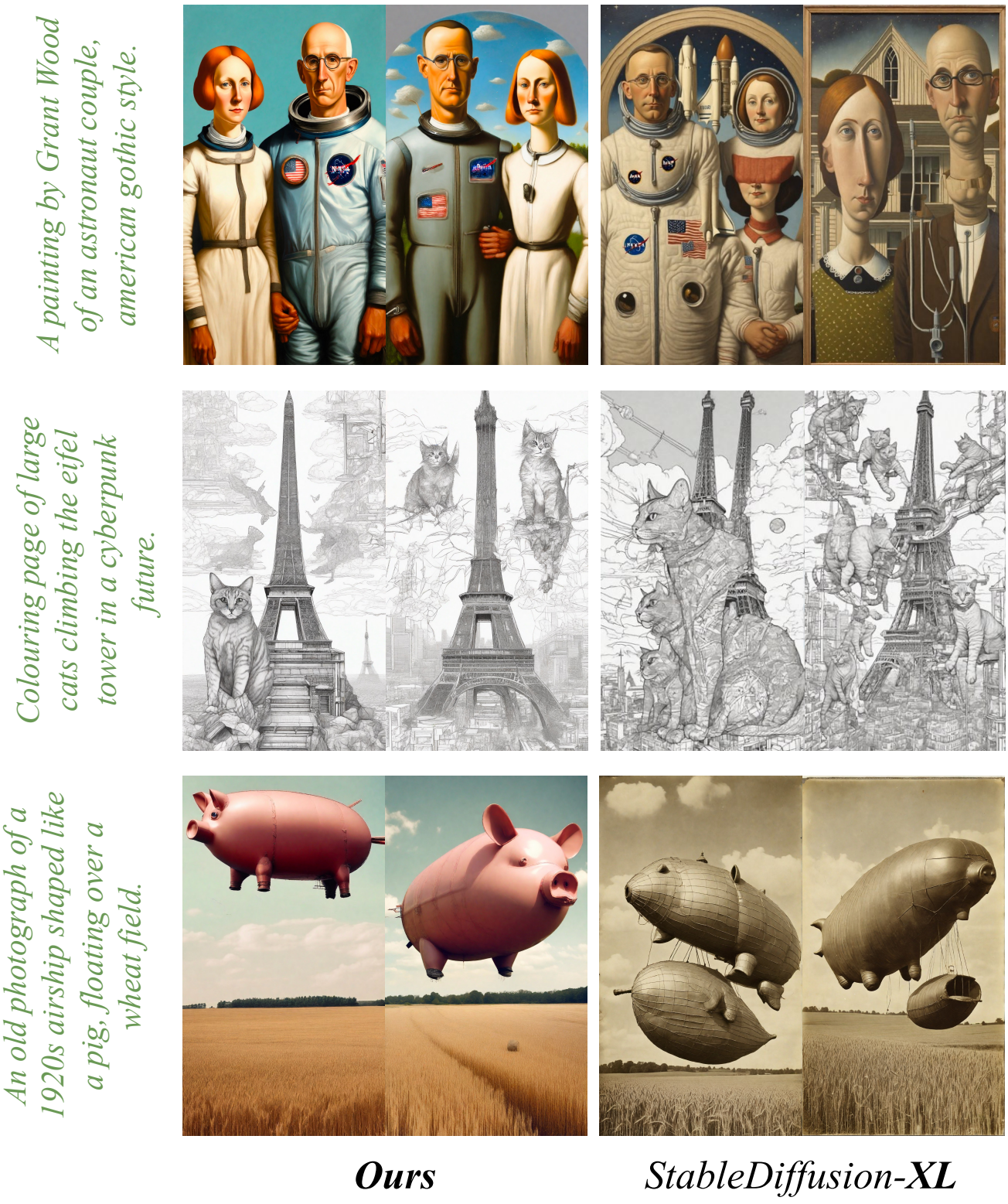
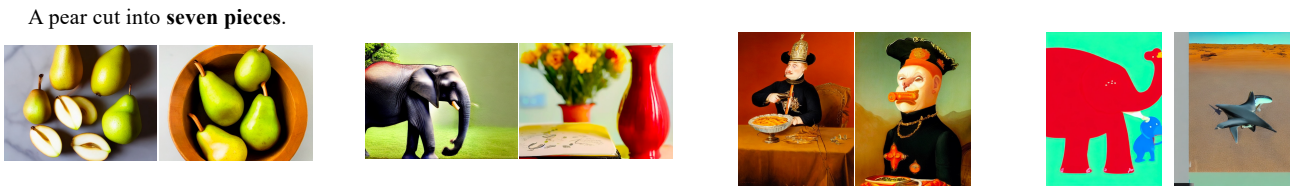


Figure 18. **Additional qualitative comparison with SDXL on Full-HD portrait images**. We use randomly sampled DrawBench prompts from the Reddit Category. Similar to landscape images, *StabelDiffusion-XL* produce repeated elements that significantly affect the image coherence. In comparison, our method achieves superior composition while maintaining a similar level of detail.



(A) Poor Image-text alignment (C) Blurry Outputs (B) Emerging Artifacts (D) Background Bleedthrough

Figure 19. **Limitations of ElasticDiffusion.** (A) poor text-image alignment for complex prompts, inherited from the base diffusion model, (B) increased blur in outputs with higher RRG weight, (C) emerging artifacts in complex images, and (D) rare background bleed-through where the color-constant background is unintentionally included in the generated image.