ElasticDiffusion: Training-free Arbitrary Size Image Generation through Global-Local Content Separation

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 [4] at various resolutions including full-HD. Our code can be accessed at https://github.com/MoayedHajiAli/ElasticDiffusion-official.git

1. ElasticDiffusion Symbols and Implementation.

To simplify the notation in this paper, we have employed specific symbols to denote the key elements within our framework. To clarify the associations between each symbol and their meanings, we provide in Tab. 1, a detailed explanation of each symbol that we used in the paper. Additionally, we describe in Algorithm 1 the full generation process of an image or arbitrary size $\bar{H} \times \bar{W}$ using a pre-trained base diffusion models that operates on images of size $H \times W$

Table 1. Table of symbols used in this paper.

Symbol	Correspondence
$H \times W$	training resolution of the base diffusion model
$\bar{H} \times \bar{W}$	target resolution of the generated image
$N \times M$	chosen resolution of the same aspect ratio as $\bar{H} \times \bar{W}$ but smaller than $H \times W$
ϵ_θ	pre-trained diffusion model network
w	classifier-free guidance weight
x_t	diffusion latent at timestep t of size $H \times W$
$\bar{x}_t$	diffusion latent at timestep t of size $\bar{H} \times \bar{W}$
$\mathbf{x}_t$	downsampled latent from $\bar{x}_t$
$\hat{\mathbf{x}}_t$	padded $\mathbf{x}_t$ to match the training resolution $H \times W$
$\hat{x}_0^t$	a noise-free sample of x
$\hat{\mathbf{x}}_0^t$	a noise-free sample of $\mathbf{x}$
p_k	a crop of $\bar{x}_t$ smaller than the training resolution
c_k	a context crop of $\bar{x}_t$
$\mathbf{S}_u$	unconditional score for latent at size $H \times W$
$\mathbf{S}_c$	conditional score for latent at size $H \times W$
$\mathbf{S}_d$	class-direction score for latent at size $H \times W$
$\bar{\mathbf{S}}_u$	unconditional score for latent at size $\bar{H} \times \bar{W}$
$\bar{\mathbf{S}}_d$	class-direction score for latent at size $\bar{H} \times \bar{W}$

2. 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.

2.1. 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 [5]. 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 [2].

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

```
Require:
    $\epsilon_\theta$                                              ▷ 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
    $\mathbf{x}_t \leftarrow \mathrm{Downsample}(\bar{x}_t, N \times M)$
    $\mathcal{Z}_t \sim \mathcal{N}(0, I)$
    $\mathcal{A}_t \leftarrow \mathcal{Y}_{H-N, W-M}, \mathcal{Y} \sim$ Uniform$(0, 255)$
    $\hat{\mathbf{x}}_t \leftarrow \mathrm{Pad}\left(\mathbf{x}_t, \mathcal{A}_t\sqrt{\bar{\alpha}_t} + \sqrt{1-\bar{\alpha}_t} \cdot \mathcal{Z}_t\right)$   ▷ Pad to match training resolution $H \times W$
    $\mathbf{S}_c \leftarrow \mathrm{Crop}\left(\epsilon_\theta\left(\hat{\mathbf{x}}_t, c\right), N \times M\right)$   ▷ conditional score at target aspect ratio
    $\mathbf{S}_u \leftarrow \tilde{\epsilon}_\theta(\hat{\mathbf{x}}_t)$   ▷ Uncodnitional score from Eq. (3)
    $\bar{\mathbf{S}}_d^0 \leftarrow \mathrm{Upsample}(\mathbf{S_c} - \hat{\mathbf{S}}_\mathbf{u}, \bar{H} \times \bar{W})$   ▷ class-direction score
    for all $r = 1, \dots, R$ do
        $\bar{\mathbf{S}}_d^r \leftarrow \mathrm{Resample}(\bar{\mathbf{S}}_d^{r-1}, \bar{x}_t)$   ▷ 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
    $\bar{x}_{t-1} \leftarrow \bar{x}_{t-1} - \mathrm{RRG}(\bar{x}_t, \mathbf{x}_t)$   ▷ Eq. (7)
end for
return $\bar{x}_0$
```

In Fig. 1, 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 [6, 10].

2.2. 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 [1, 3, 8]. 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. 2 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. 3. 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. 4 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.

3. 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.

3.1. Additional Ablation Study.

To better understand the effect of the class-direction refinement and reduced-resolution guidance (RRG) strategy on the overall quality of the results, we analysed in Tab. 2 the performance of ElasticDiffusion when excluding these components at two different resolutions. We observe that their necessity is pronounced at higher resolutions (e.g., 4x), while their influence is limited at lower upsampling scale (e.g., 2x). Notably, even without these components, our method achieves *better* FID than baselines.

	768×768		1024×1024	
	FID ↓	CLIP ↑	FID ↓	CLIP ↑
Ours	225.86	26.66	228.87	23.74
w/o RRG	230.27	24.13	234.66	23.15
w/o RGG & refined class-direction score	233.49	23.32	263.15	20.91

Table 2. **Ablation study** on CelebAHQ Dataset using *StableDiffison*$_{1.4}$ as the base model.

3.2. 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 [9]. 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. 5, *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.

3.3. Additional Qualitative Results

We provide further qualitative analysis of ElasticDiffusion.

Figure 6 provides a comparison with *StableDiffusion* and *MultiDiffusion* using selected DrawBench [7] prompts for horizontal 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 7 shows a similar analysis on vertical 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 8 and 9 focus on showing results for the generation of Full-HD horizontal 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. 8 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.

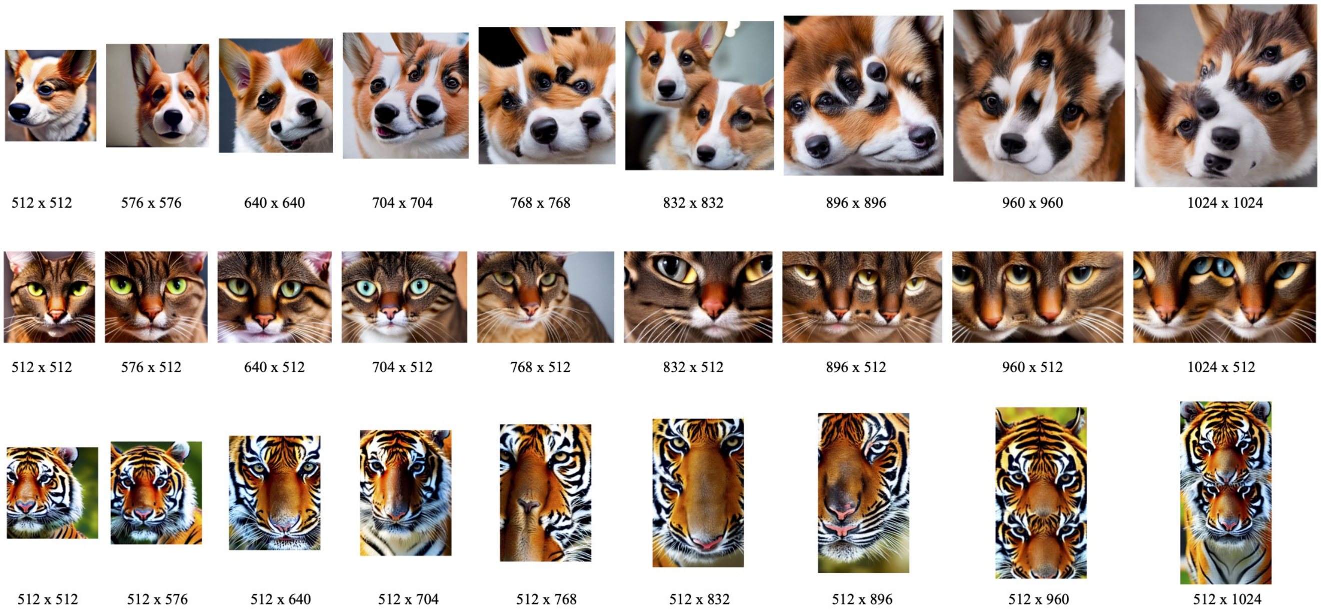


Figure 1. **Degradation of image quality in StableDiffusion$_{1.4}$ with varying resolutions.** We illustrate the progressive decrease in image quality as the resolution deviates from the model's training size of 512×512. The results on square, horizontal, and vertical resolutions show a significant reduction in global coherence for image sizes beyond 64 pixels from the training resolution.

Figures 10 and 11 provide a similar analysis for Full-HD vertical 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. 10, or stretched parts like the woman face in the first example of Fig. 11. In contrast, our method generates detailed objects that fit the vertical aspect ratio while avoiding any stretching or element repetition.

3.4. Limitations

Fig. 12 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).

References

[1] 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. 2

[2] Hugging Face. Stable diffusion xl. `https://huggingface.co/docs/diffusers/main/en/using-diffusers/sdxl`, 2023. Accessed: [Insert Date Here]. 1

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[4] 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. 1

[5] Suraj Patil, Pedro Cuenca, Nathan Lambert, and Patrick von Platen. Stable diffusion with diffusers. *Hugging Face Blog*, 2022. `https://huggingface.co/blog/stable_diffusion`. 1

[6] 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. 2

[7] 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. 3

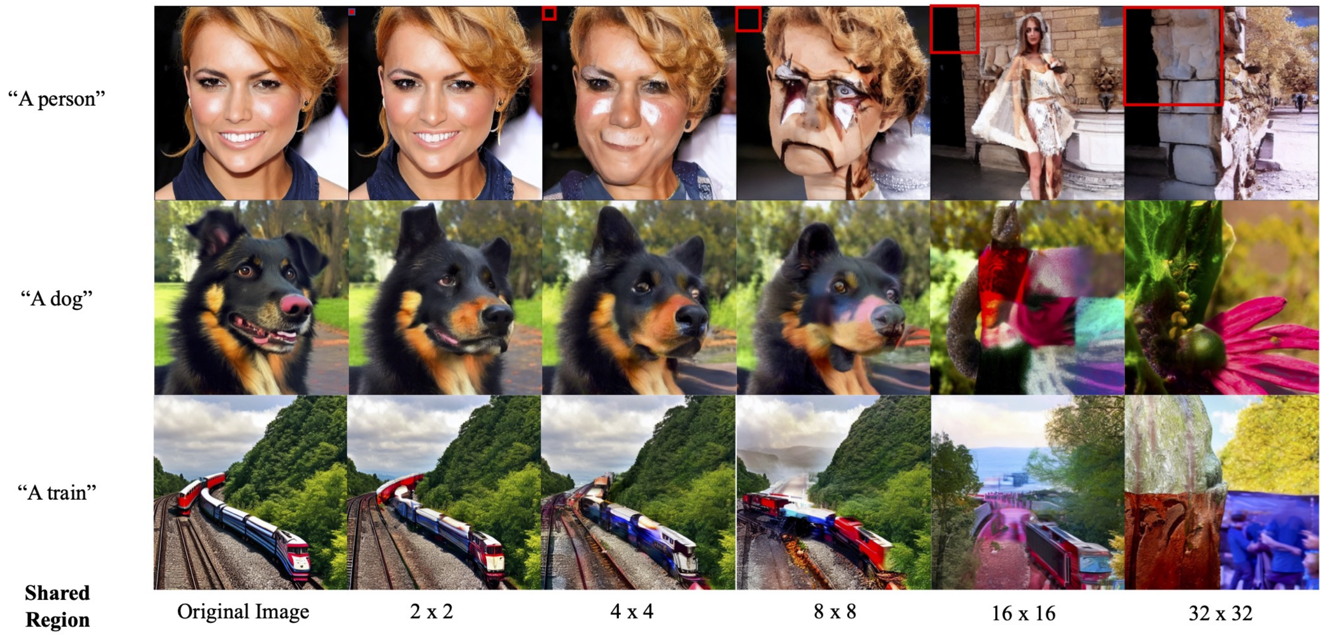


Figure 2. **Effect of sharing *class direction score* between nearby pixels.** We highlight that sharing the score within a group of neighboring pixels preserves the global content and coherence of the image, despite a reduction in the perceptual quality when selecting larger group sizes (as denoted 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.

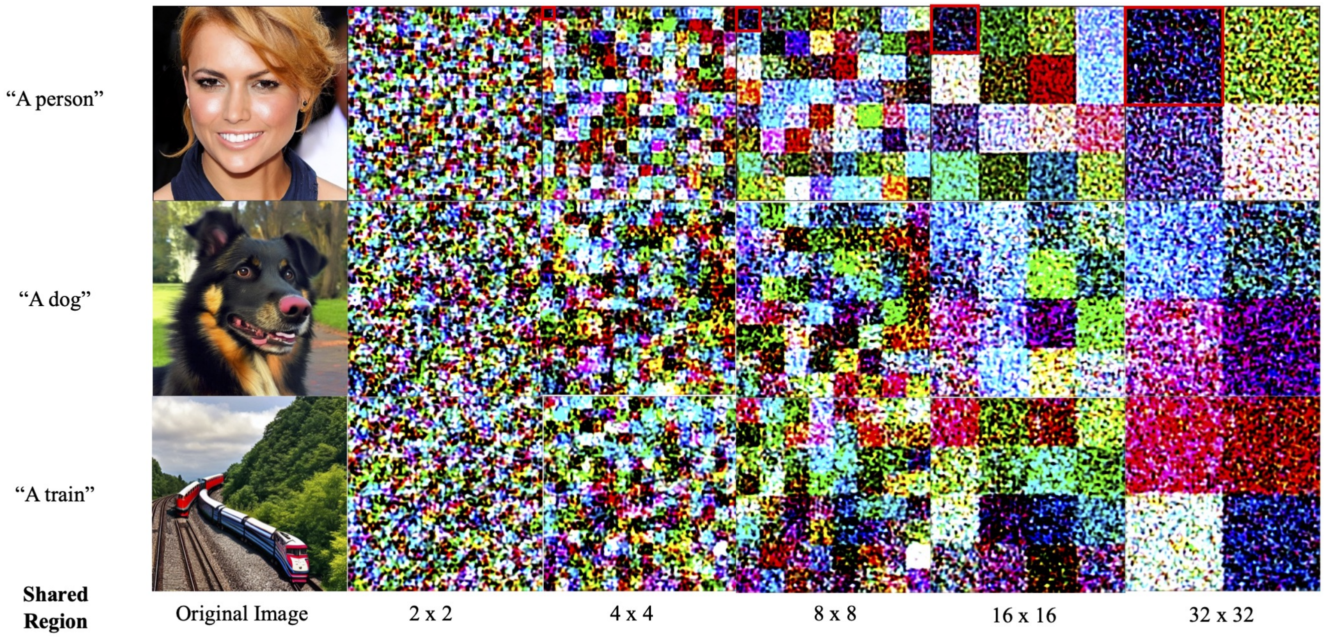


Figure 3. **Effect of sharing *unconditional score* between nearby pixels.** We show that sharing the unconditional score, even in small groups of pixels, leads to the generation of complete noise. This indicates that *uncondtional score*, in contrast to the *class direction score*, requires pixel-level precision to generate local details.

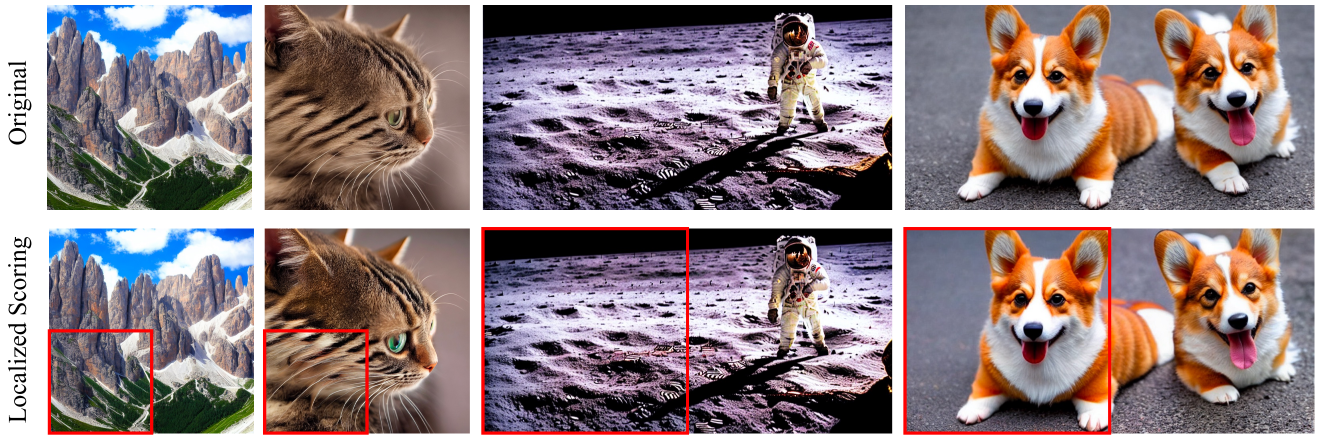


Figure 4. **Unconditional score computation on localized regions.** We show that computing the diffusion model *unconditional score* on local patches (corresponding to the size of the red boxes in the second row) results in images that are visually similar to those produced by a global score computation (displayed in the first row).

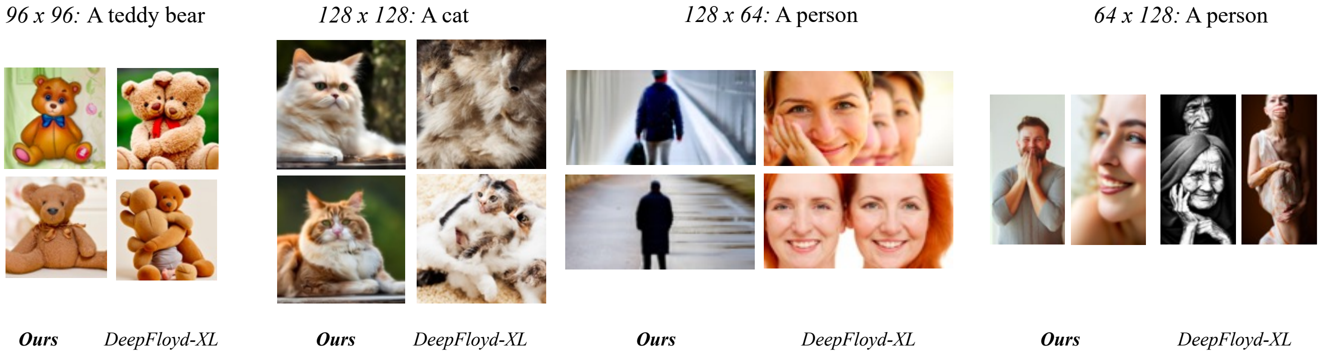


Figure 5. ***DeepFloyd-XL* across various image sizes.** We test *DeepFloyd-XL*, a diffusion model that operates on the pixel space, across multiple image resolutions. We observe a degradation in performance similar to that seen in $StableDiffusion$. The application of ElasticDiffusion significantly improves the overall composition of the generated images.

[8] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *ICCV*, 2019. 2

[9] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. DeepFloyd. `https://www.deepfloyd.ai/`, 2023. Accessed: 2023. 3

[10] 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. 2

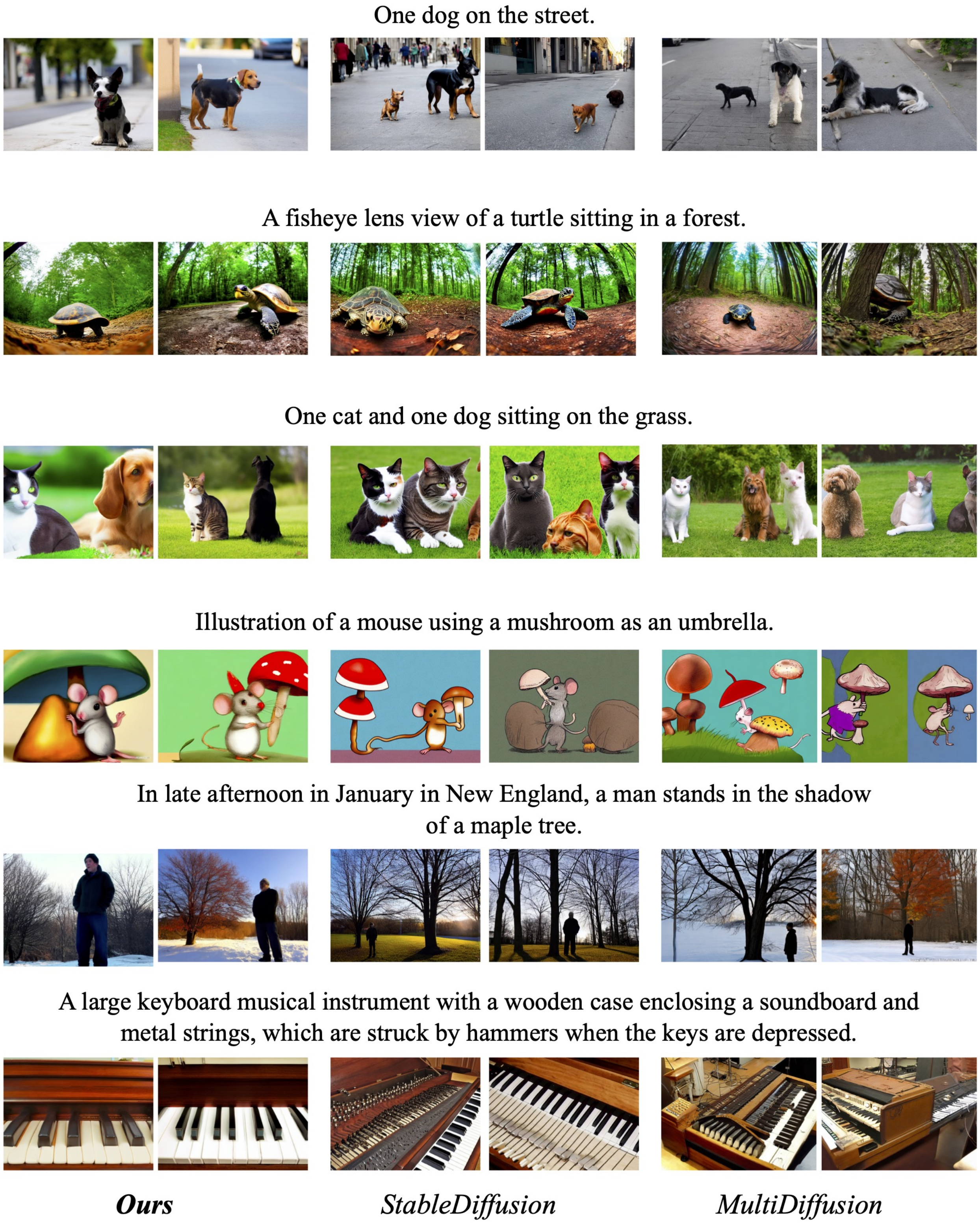

One dog on the street.
A fisheye lens view of a turtle sitting in a forest.
One cat and one dog sitting on the grass.
Illustration of a mouse using a mushroom as an umbrella.
In late afternoon in January in New England, a man stands in the shadow of a maple tree.
A large keyboard musical instrument with a wooden case enclosing a soundboard and metal strings, which are struck by hammers when the keys are depressed.
Ours
StableDiffusion
MultiDiffusion

Figure 6. **Additional qualitative comparison on horizontal 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.

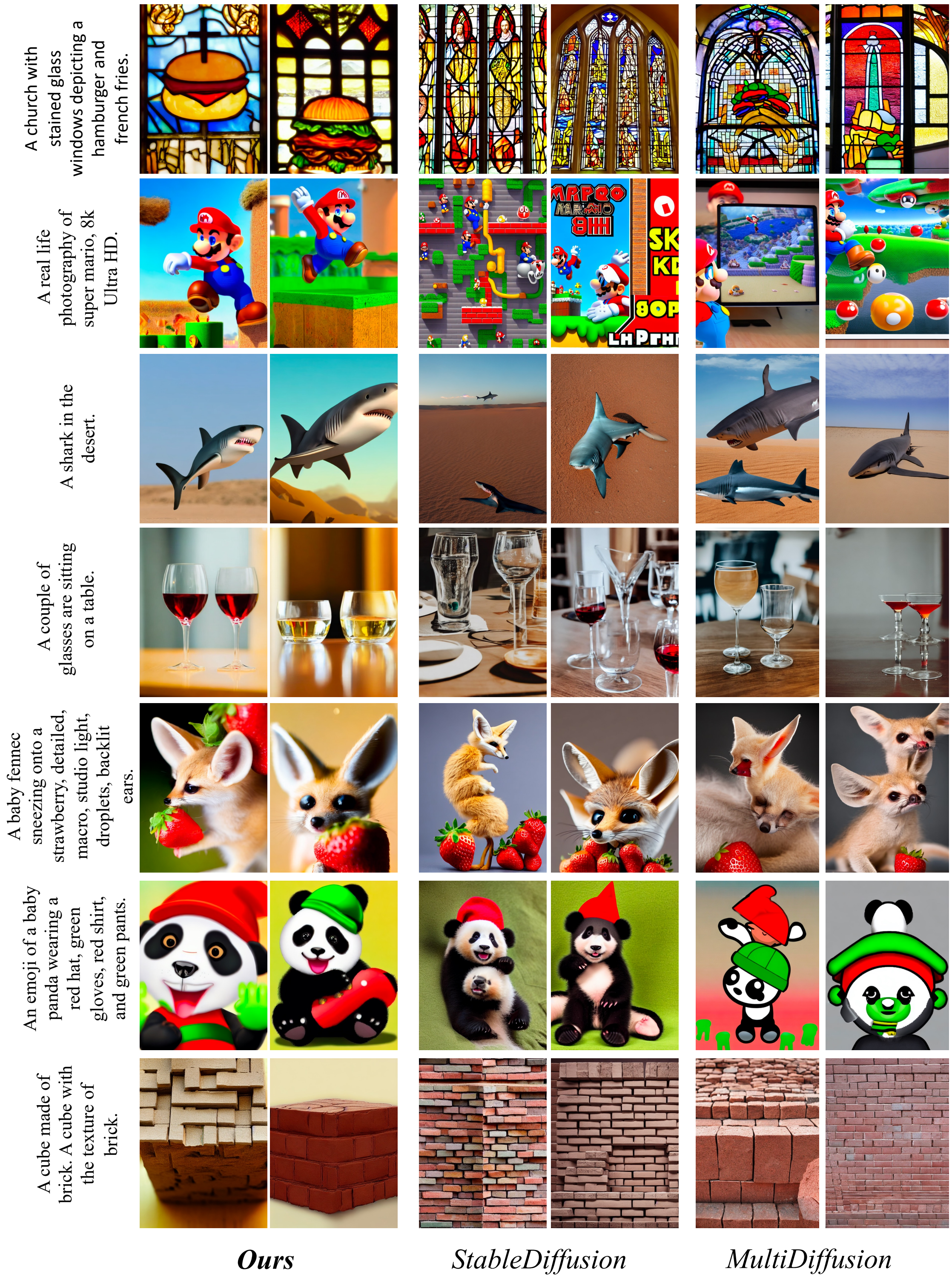

A church with stained glass windows depicting a hamburger and french fries.
A real life photography of super mario, 8k Ultra HD.
A shark in the desert.
A couple of glasses are sitting on a table.
A baby fennec sneezing onto a strawberry, detailed, macro, studio light, droplets, backlit ears.
An emoji of a baby panda wearing a red hat, green gloves, red shirt, and green pants.
A cube made of brick. A cube with the texture of brick.
Ours
StableDiffusion
MultiDiffusion

Figure 7. **Additional qualitative comparison on vertical 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 horizontal resolutions, 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.

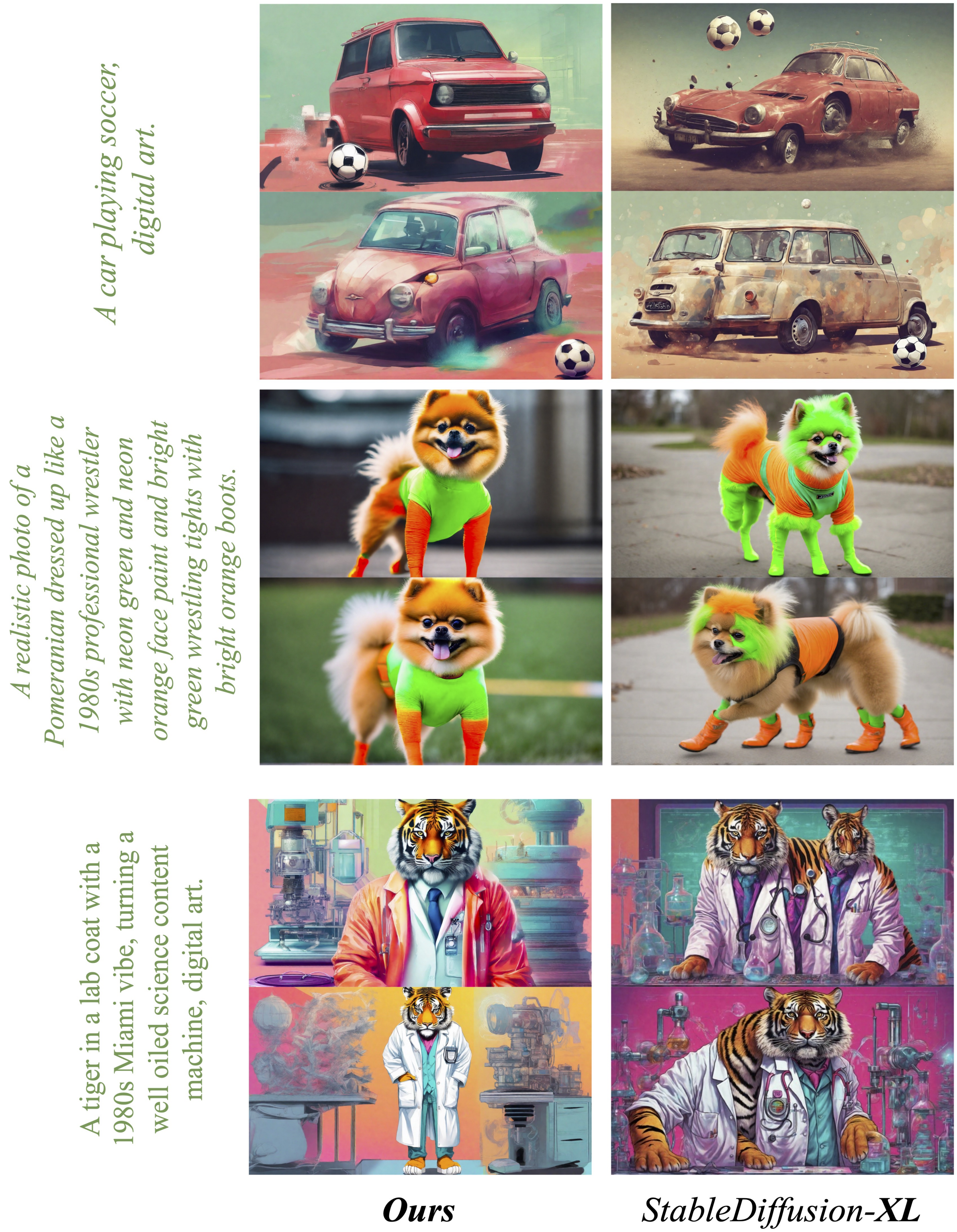

A car playing soccer, digital art.
A realistic photo of a Pomeranian dressed up like a 1980s professional wrestler with neon green and neon orange face paint and bright green wrestling tights with bright orange boots.
A tiger in a lab coat with a 1980s Miami vibe, turning a well oiled science content machine, digital art.
Ours
StableDiffusion-XL

Figure 8. **Additional qualitative comparison with SDXL on Full-HD horizontal 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.

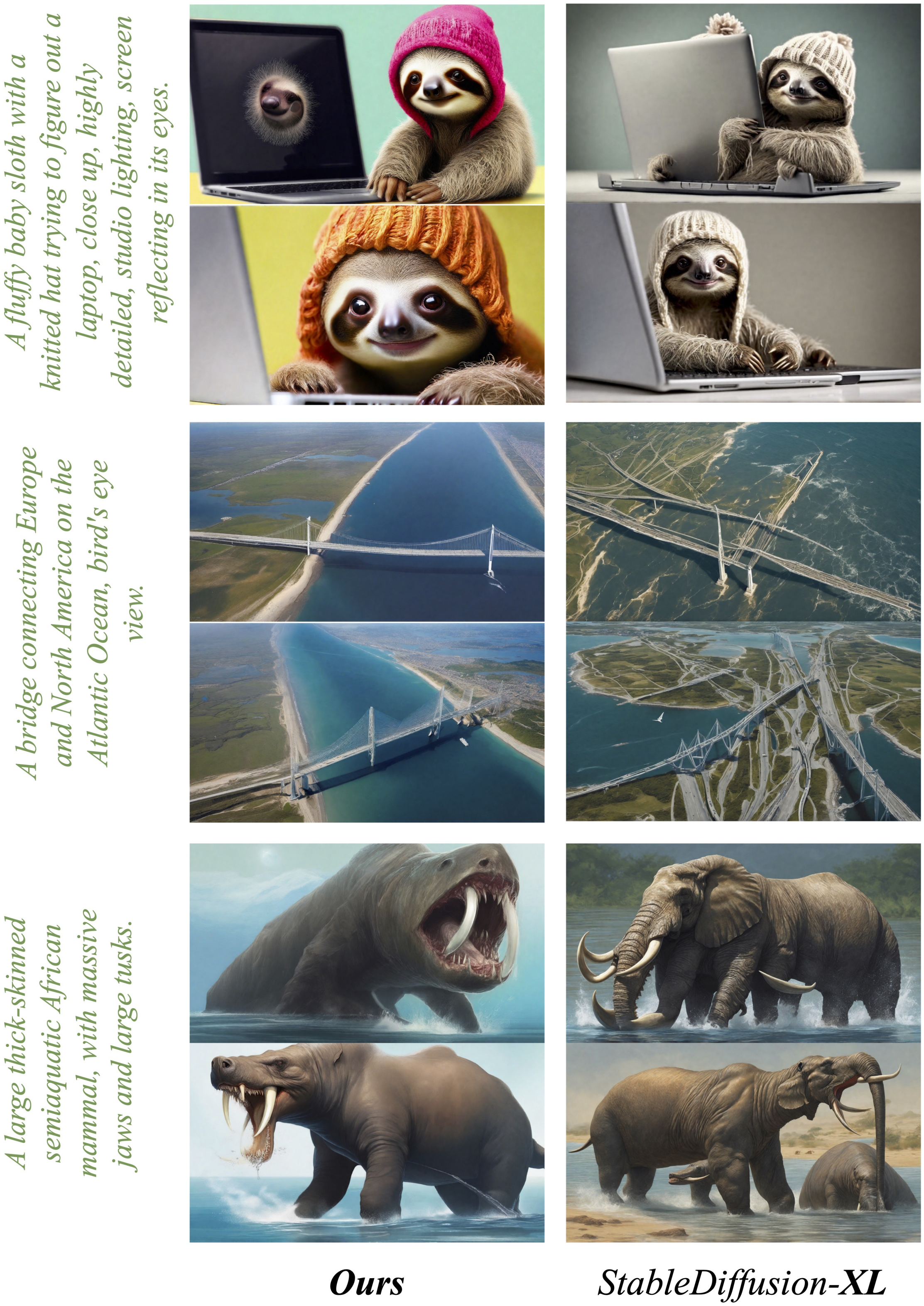

A fluffy baby sloth with a knitted hat trying to figure out a laptop, close up, highly detailed, studio lighting, screen reflecting in its eyes.
A bridge connecting Europe and North America on the Atlantic Ocean, bird's eye view.
A large thick-skinned semiaquatic African mammal, with massive jaws and large tusks.
Ours
StableDiffusion-XL

Figure 9. **Additional qualitative comparison with SDXL on Full-HD horizontal 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.

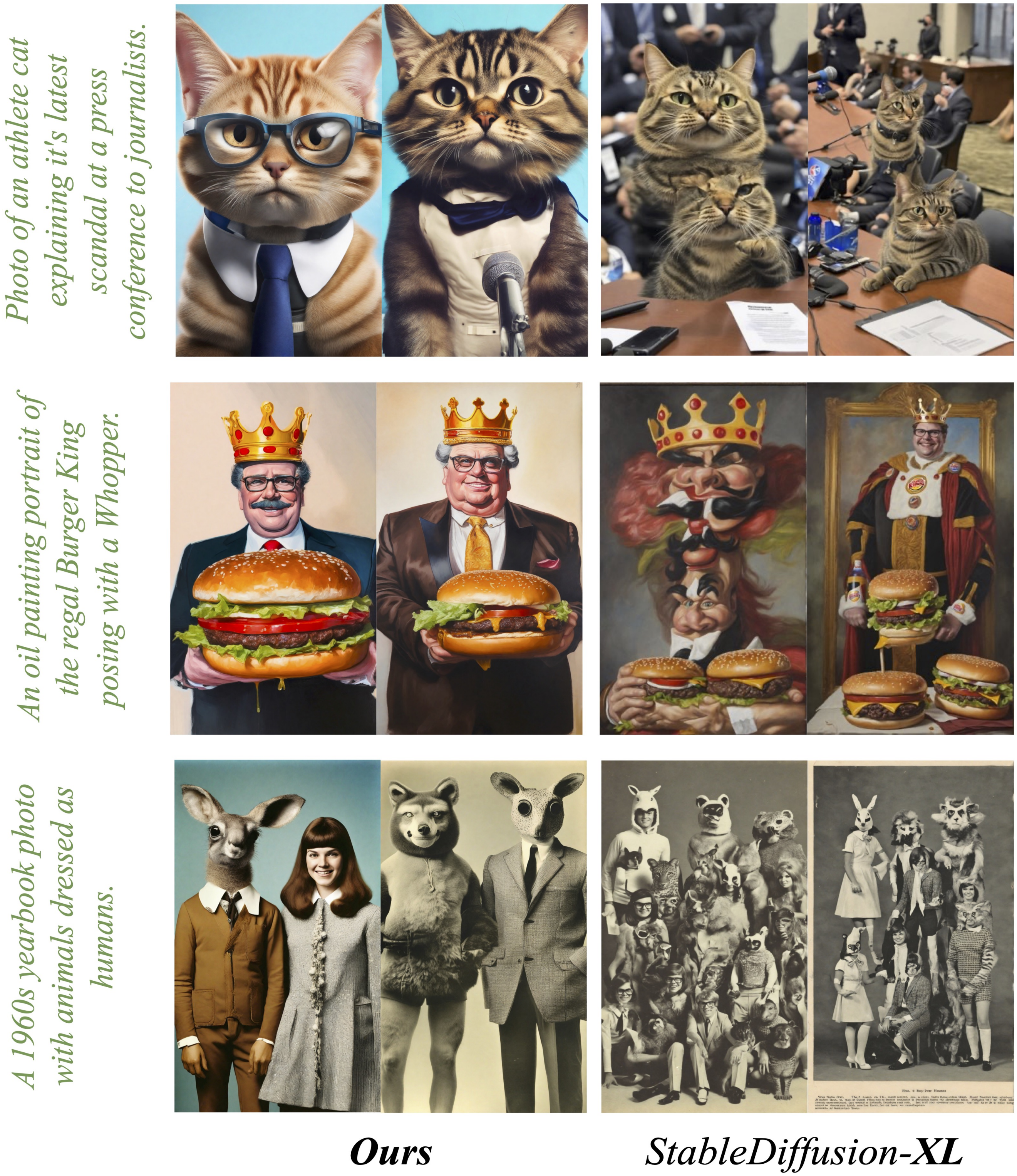


Photo of an athlete cat explaining it's latest scandal at a press conference to journalists.
An oil painting portrait of the regal Burger King posing with a Whopper.
A 1960s yearbook photo with animals dressed as humans.
Ours
StableDiffusion-XL

Figure 10. **Additional qualitative comparison with SDXL on Full-HD vertical images**. We use randomly sampled DrawBench prompts from the Reddit Category. Similar to horizontal resolutions, *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.

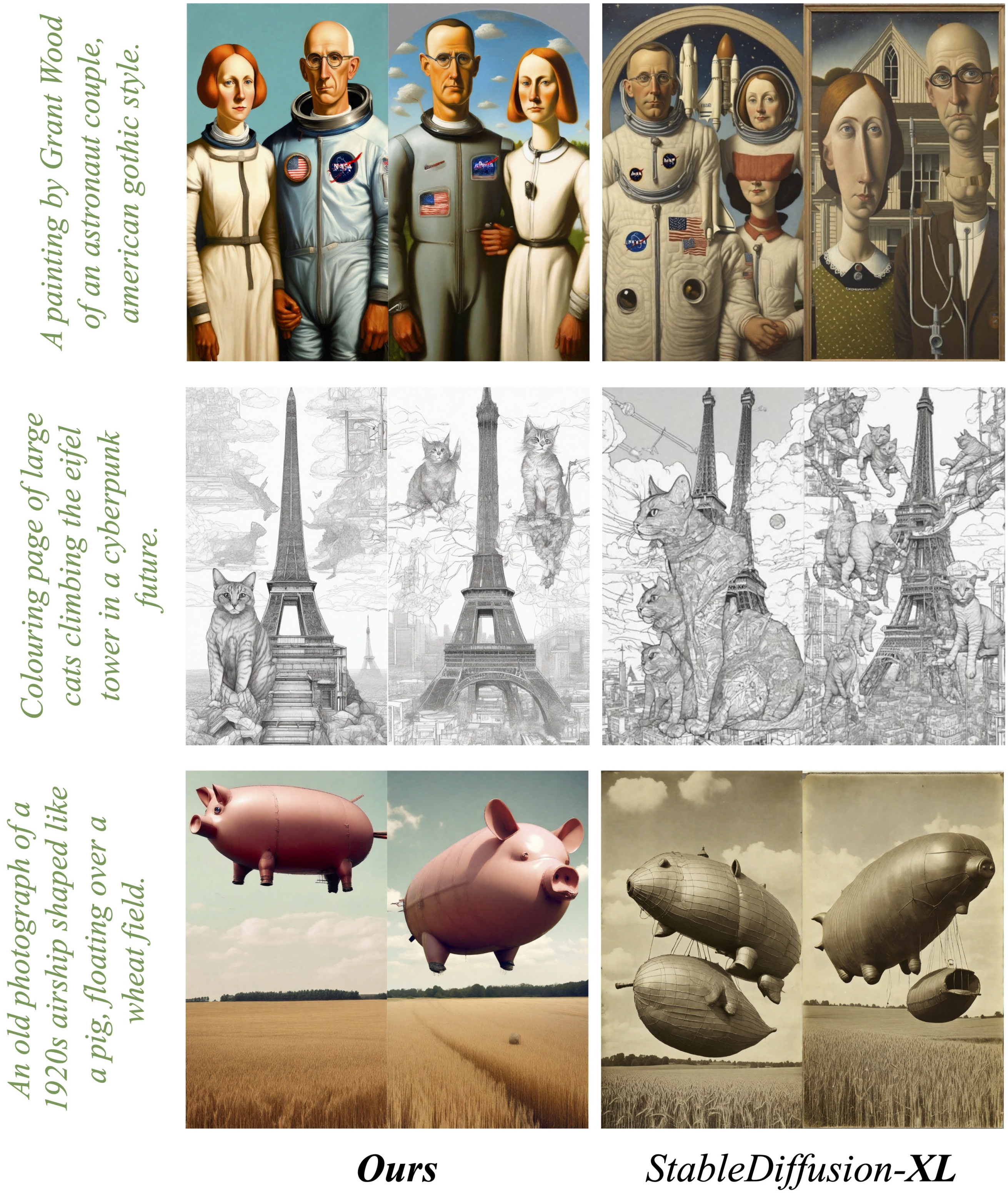
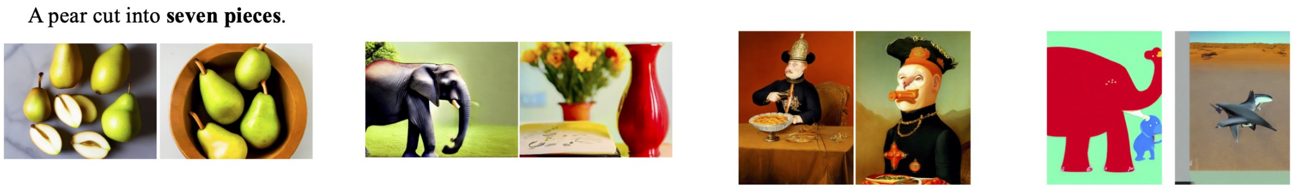

A painting by Grant Wood of an astronaut couple, american gothic style.
Colouring page of large cats climbing the eifel tower in a cyberpunk future.
An old photograph of a 1920s airship shaped like a pig, floating over a wheat field.
Ours
StableDiffusion-XL

Figure 11. **Additional qualitative comparison with SDXL on Full-HD vertical images**. We use randomly sampled DrawBench prompts from the Reddit Category. Similar to horizontal resolutions, *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 12. **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.