Can We Achieve Efficient Diffusion without Self-Attention? Distilling Self-Attention into Convolutions (Supplementary Material)

Ziyi Dong[1] Chengxing Zhou[1] Weijian Deng[2] Pengxu Wei[1,3] Xiangyang Ji[4] Liang Lin[1,3]
[1]Sun Yat-sen University [2]Australian National University [3]Peng Cheng Laboratory [4]Tsinghua University
dongzy6@mail2.sysu.edu.cn, zhouchx33@mail2.sysu.edu.cn, dengwj16@gmail.com,
weipx3@mail.sysu.edu.cn, xyji@tsinghua.edu.cn, linliang@ieee.org

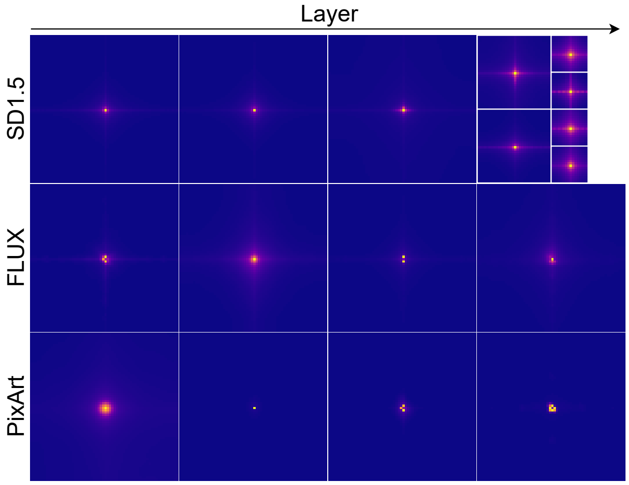

Figure I. Averaged centroid-aligned attention maps of different layers. U-Net models (*e.g.* SD1.5) has multi-scale attention maps.

A. Averaged Centroid-Aligned Attention Maps

To gain deeper insights into the localized pattern as analyze in Sec. 3, Fig. I presents averaged centroid-aligned attention maps, where each attention map is shifted such that the query point is centered, making it the anchor. This transformation enables direct comparison of attention distributions across different spatial locations, revealing a strong locality pattern.

B. Visualization Results

More visualization results of our ΔConvFusion and and original self-attention based diffusion models are shown in Fig. IV. The images produced by our ΔConvFusion exhibit enhanced visual realism and improved semantic alignment with textual prompts.

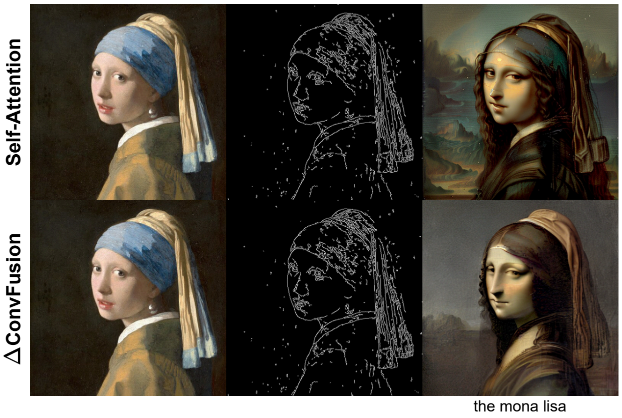

Figure II. Images generate with our ΔConvFusion and self-attention based diffusion model (SD1.5) while applying ControlNet.

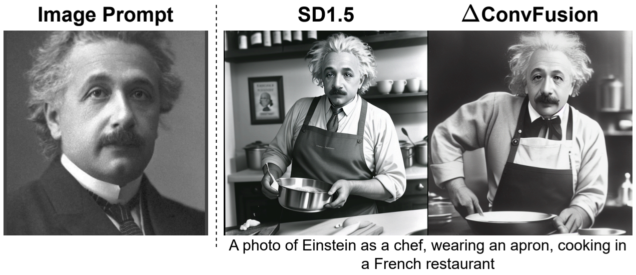
Image Prompt
SD1.5
ΔConvFusion
A photo of Einstein as a chef, wearing an apron, cooking in a French restaurant

Figure III. Images generate with our ΔConvFusion and self-attention based diffusion model (SD1.5) while applying IP-Adapter.

C. Extend Tasks

Our ΔConvFusion remains compatible with existing methods and adapters that are trained with self-attention diffusion models, such as ControlNet, which incorporates spatial conditions, the IP-Adapter, which uses images as prompts, and inpainting operations.

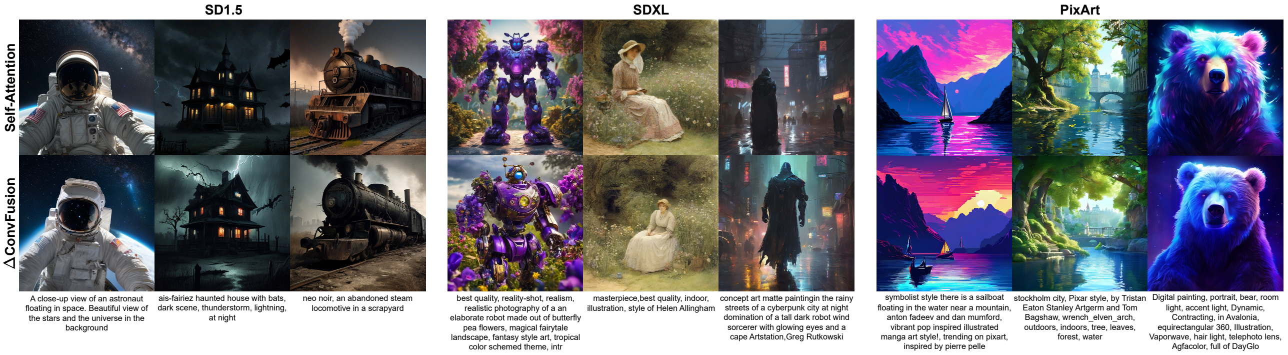
SD1.5
SDXL
PixArt
Self-Attention
ΔConvFusion
A close-up view of an astronaut floating in space. Beautiful view of the stars and the universe in the background
ais-fairiez haunted house with bats, dark scene, thunderstorm, lightning, at night
neo noir, an abandoned steam locomotive in a scrapyard
best quality, reality-shot, realism, realistic photography of a an elaborate robot made out of butterfly pea flowers, magical fairytale landscape, fantasy style art, tropical color schemed theme, intr
masterpiece,best quality, indoor, illustration, style of Helen Allingham
concept art matte paintingin the rainy streets of a cyberpunk city at night domination of a tall dark robot wind sorcerer with glowing eyes and a cape Artstation,Greg Rutkowski
symbolist style there is a sailboat floating in the water near a mountain, anton fadeev and dan mumford, vibrant pop inspired illustrated manga art style!, trending on pixart, inspired by pierre pelle
stockholm city, Pixar style, by Tristan Eaton Stanley Artgerm and Tom Bagshaw, wrench_elven_arch, outdoors, indoors, tree, leaves, forest, water
Digital painting, portrait, bear, room light, accent light, Dynamic, Contracting, in Avalonia, equirectangular 360, Illustration, Vaporwave, hair light, telephoto lens, Agfacolor, full of DayGlo

Figure IV. Generated images of our ΔConvFusion and original self-attention based diffusion models across SD1.5, SDXL and PixArt.

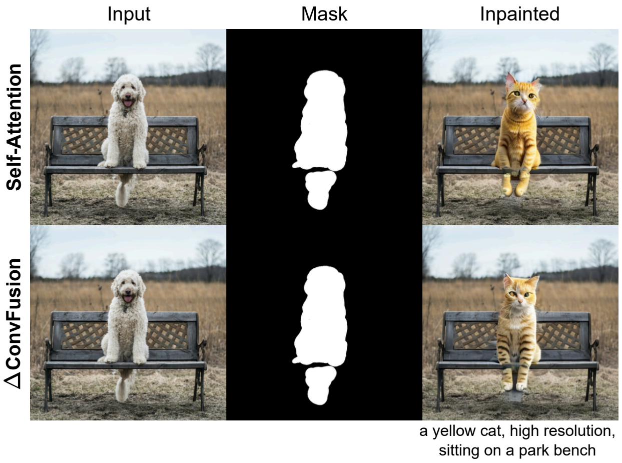

Figure V. Result of inpainting with our ΔConvFusion and self-attention based diffusion model (SD1.5).

C.1. Generate with ControlNet

As shown in Fig. II, our ΔConvFusion is compatible with existing ControlNet models, enabling the model to generate images with spatial conditions.

C.2. Generate with IP-Adapter

As shown in Fig. III, ΔConvFusion is compatible with existing IP-Adapters, which can enable using image as prompt.

C.3. Inpainting

As shown in Fig. V, our ΔConvFusion effectively integrates with self-attention based diffusion inpainting methods, successfully replacing the dog in the image with a cat.

D. Block-Wise ERF

In Fig. VI, we present the ERF of blocks at scales 64×64 and 32×32 in SD1.5. The figure demonstrates that the ERF distribution of ΔConvBlock closely resembles the pattern of the self-attention maps shown in Fig. I and Fig. 4b. This similarity further verifies that our design effectively captures the self-attention characteristics within the diffusion model.

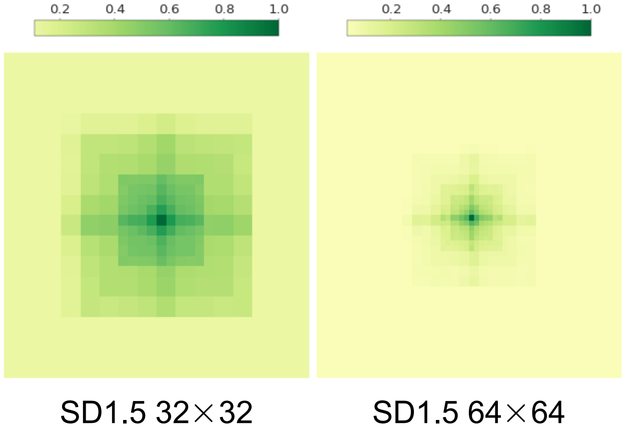

Figure VI. ERF of blocks at scales 64×64 and 32×32 in SD1.5

Table I. VRAM usage of self-attention and our ΔConvBlock with batch size of 32. The self-attention is running with FlashAttention.

	Self-Attention	ΔConvBlock
Memory (1024^2)	1.77GB	1.46GB
Memory ($4k$)	14.68GB	12.17GB

E. Memory Efficiency

Our method not only achieves faster computation but also requires less memory compared to self-attention. Tab. I presents the memory usage of our ΔConvBlock and self-attention. Our ΔConvBlock substantially reduces memory usage at both 1024×1024 and 4K resolutions.

F. Naïve Baselines

Table II. FLOPs and performance of our ΔConvBlock and local attention (NA).

	Metrics			FLOPs ↓			
Method	**DS ↑**	**FDD ↓**	**CLIP ↑**	**512^2**	**1024^2**	**4K**	**16K**
NA (K=13)	42.55	232.31	30.57	23.35	93.39	787.98	11819.72
NA (K=17)	43.18	222.35	30.66	24.48	97.92	826.20	12393.03
Ours (SD1.5)	44.72	200.15	30.73	8.43	34.86	294.10	4411.40

We include the local attention (NA) [12] in the Tab. II. While it has linear complexity, local attention incurs higher memory costs than convolution, making it less efficient than our ΔConvBlock. Moreover, our method shows clear advantages over NA in image generation.

G. Choice of Pyramid Kernel Size

Table III. Ablation analysis of pyramid kernel size.

K	DS ↑	FDD ↓	CLIP ↑
7	42.67	186.50	30.76
13	42.56	181.06	30.84
17	43.11	182.87	30.78

According to the analysis in Tab. III, a larger kernel size does not necessarily lead to better performance. Therefore, we select $K = 13$ (PixArt) and $K = 9$ (SD1.5) for our experiments.