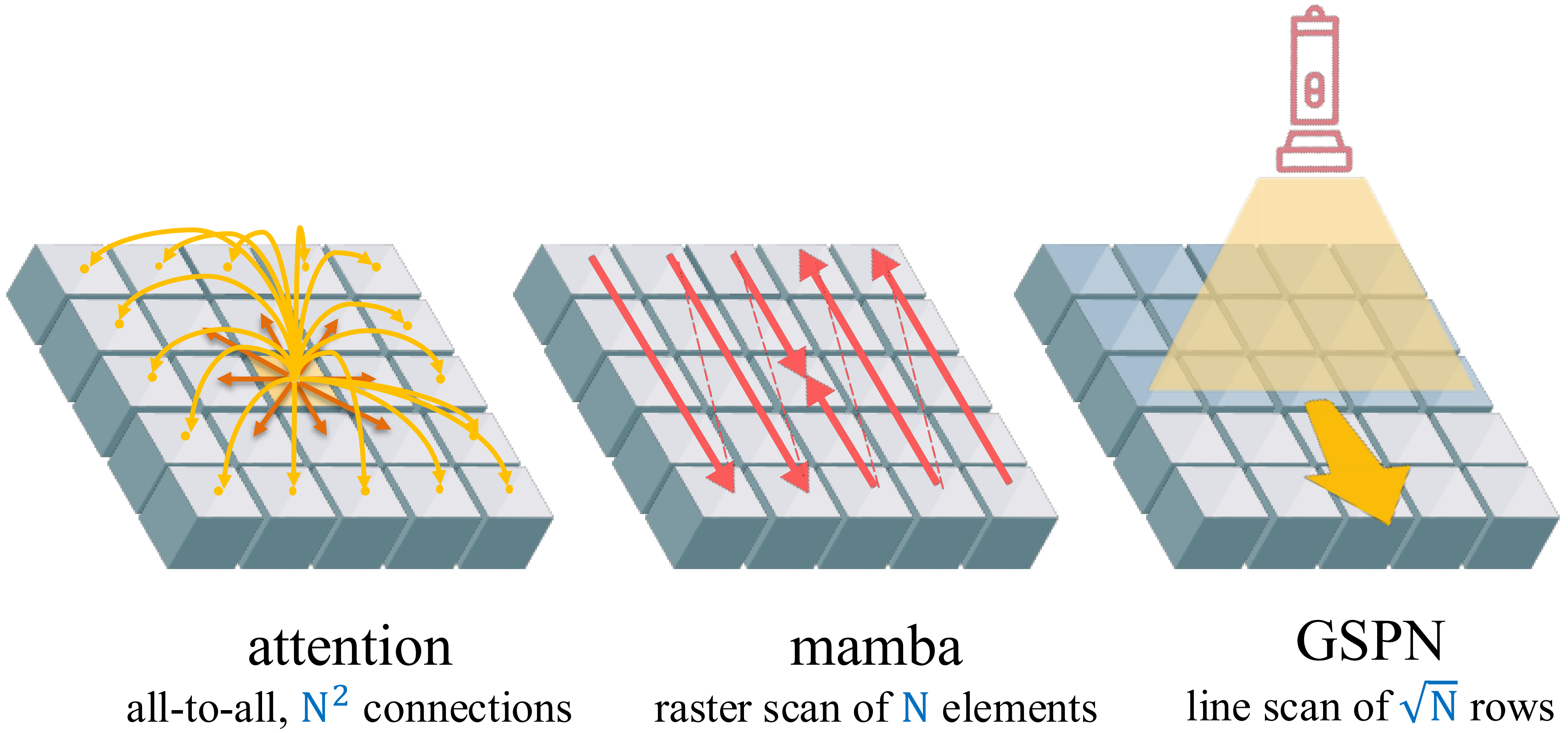
Parallel Sequence Modeling via Generalized Spatial Propagation Network

Hongjun Wang†, Wonmin Byeon, Jiarui Xu, Jinwei Gu, Ka Chun Cheung, Xiaolong Wang, Kai Han, Jan Kautz, Sifei Liu

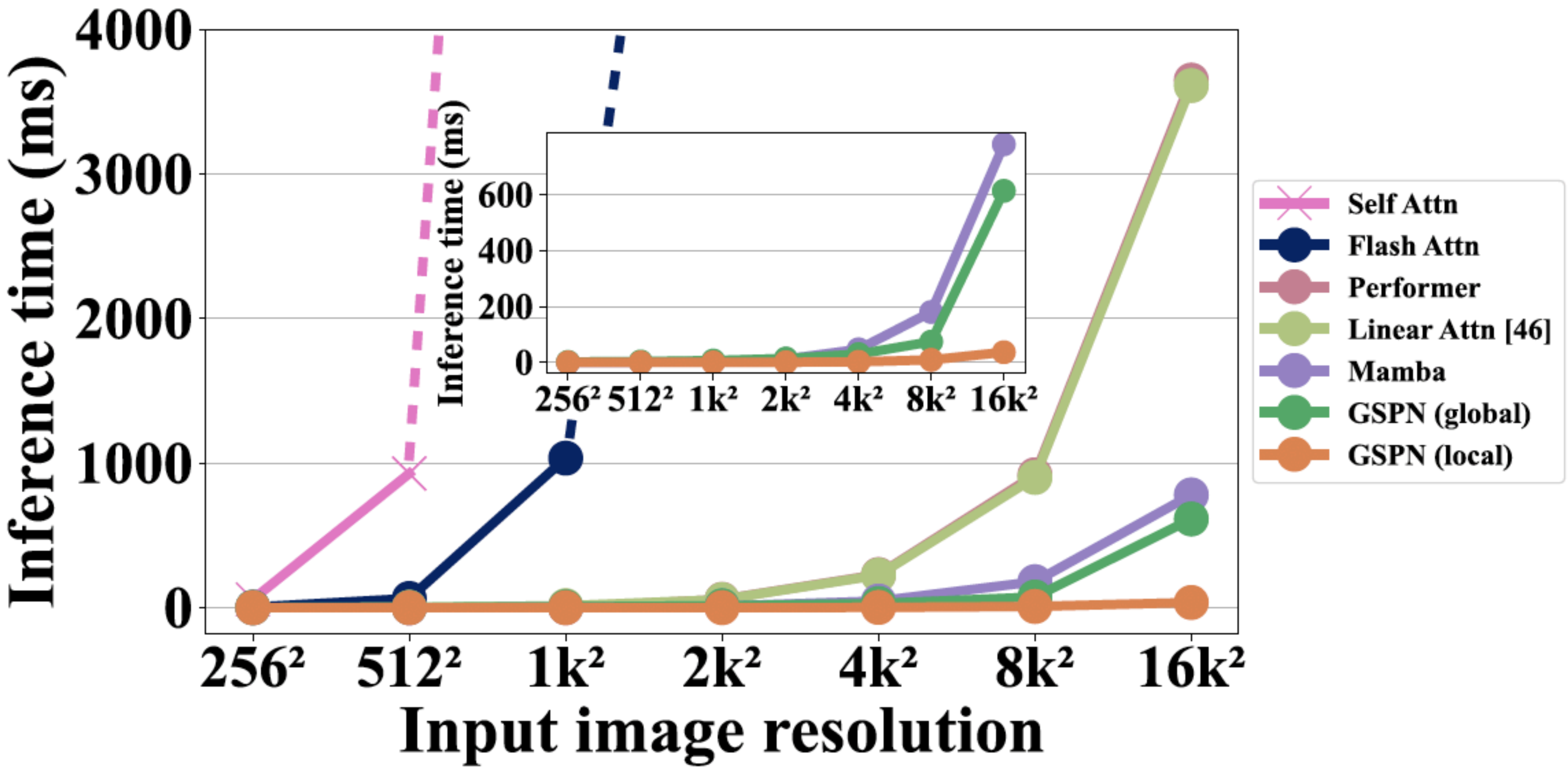
(† the work was done at an internship at NVIDIA)

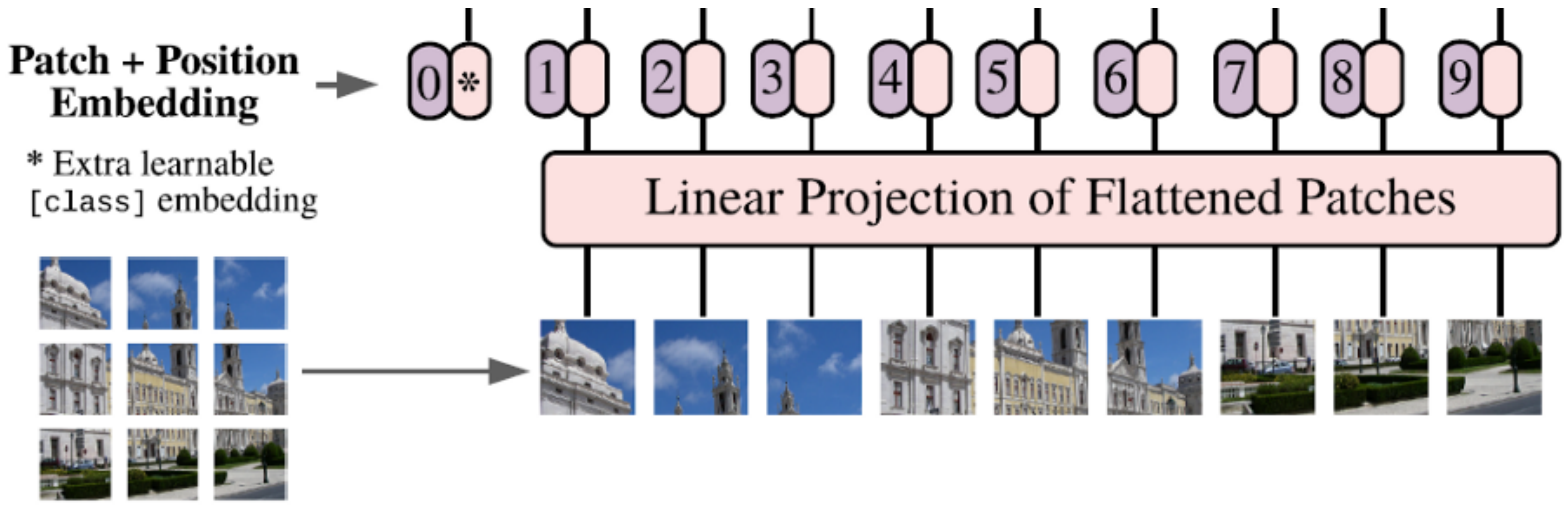


Motivation

- Quadratic computational complexity of Transformer hampers efficiency at large scales

- Transformers treat data as structure-agnostic tokens that overlook the spatial coherence (e.g. aliasing issues [1][2])





[1] Timoth ́ee Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In ICLR, 2024.
[2] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Se- ung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. Emernerf: Emergent spatial- temporal scene decomposition via self-supervision. arXiv preprint arXiv:2311.02077.

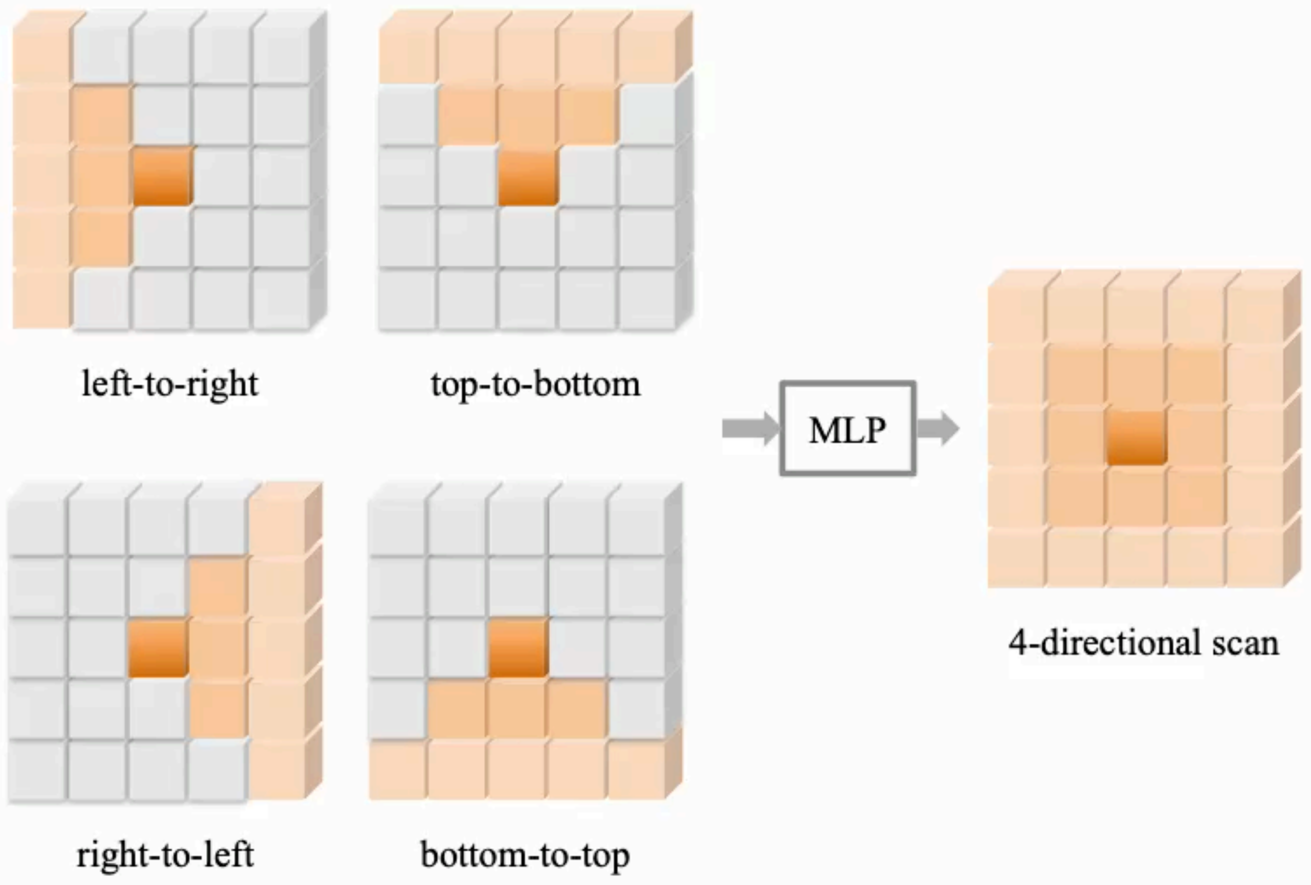
Overview

- We introduce the Generalized Spatial Propagation Network (GSPN), a linear attention mechanism optimized for multi-dimensional data such as images.
- Stability-Context Condition ensures both stability and effective long-range context propagation across 2D sequences
- GSPN parallelizes propagation across rows and columns, reducing the effective sequence length to $\sqrt{N}$, significantly enhancing the computational efficiency.

Overview

- GSPN uses a 3-way connection for parameter efficiency, while a 4-direction integration ensures full pixel connectivity, thereby forming dense pairwise connections through the line-scan manner.

- During propagation, GSPN computes a weighted sum for each pixel using pixels from its previous row or column, with weights that are learnable and input-dependent.



2D Linear Propagation

- The 2D propagation follows a linear recurrent process:

$$h_i^c = w_i^c h_{i-1}^c + \lambda_i^c \odot x_i^c, \quad i \in [1, n-1],\ c \in [0, C-1] \ (1)$$

- Vectorizing sequence of concatenated rows of hidden states and inputs, we have:

$$H = GX$$

where G is a lower triangular $N \times N$ matrix with $n \times n$ sub-matrices

- The output y_i can be represented as a weighted sum of X:

$$y_i = u_i \sum_{j=0}^{t} \prod_{\tau=j+1}^{i} w_\tau \lambda_j x_j$$

Stability-Context Condition

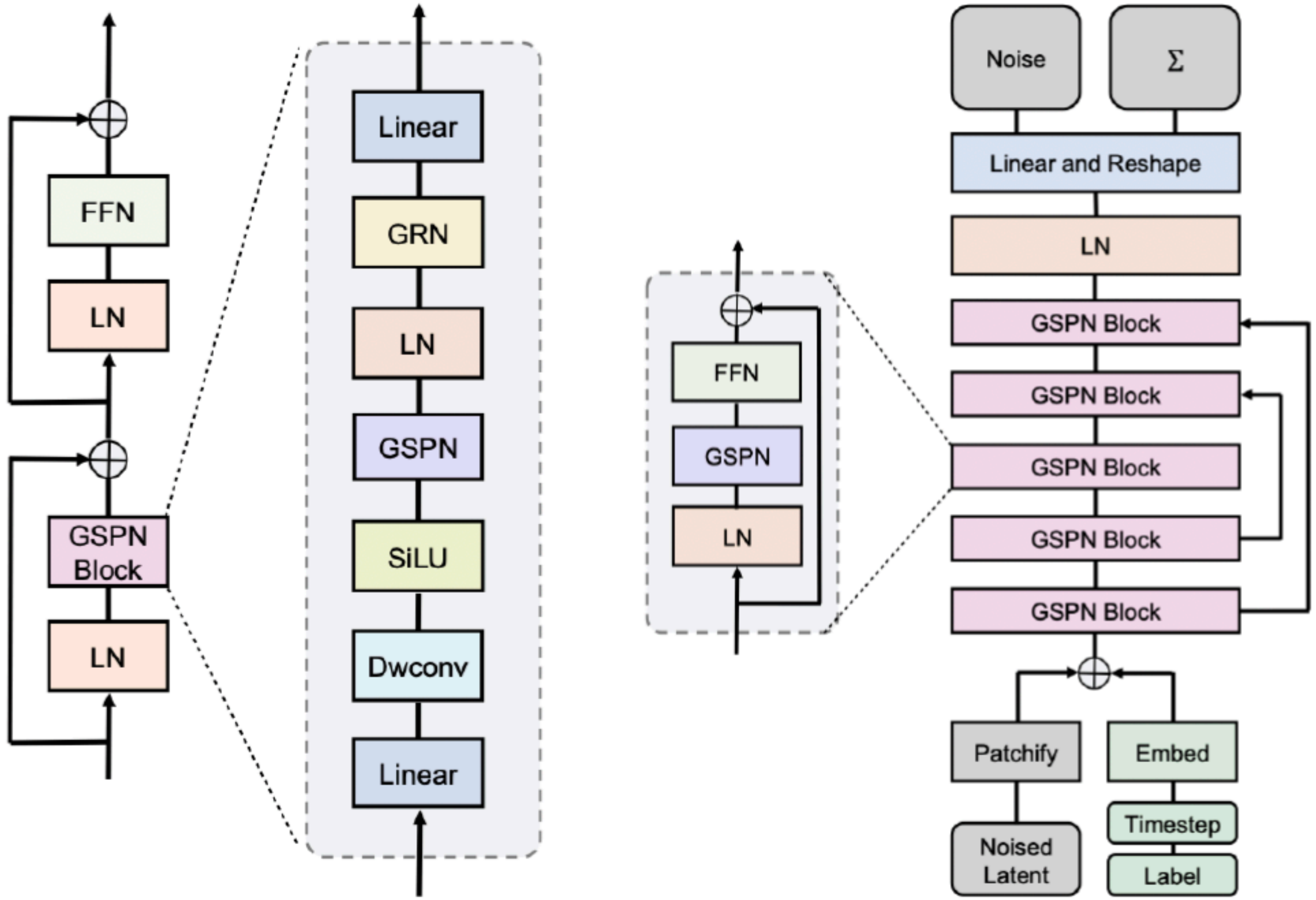
- Theorem 1 ensures effective long-range propagation

Theorem 1. *If all the matrices w_τ are row stochastic, then $\sum_{j=0}^{n-1} W_{ij} = 1$ is satisfied.*

- Theorem 2 ensures stability of 2D linear propagation

Theorem 2. *The stability of Eq. (1) is ensured when all matrices w_τ are row stochastic.*

Architecture



(a) Classification (b) Generation

Experiment

- Image Classification
- Class-conditional Generation
- Text-to-image Generation

Image Classification

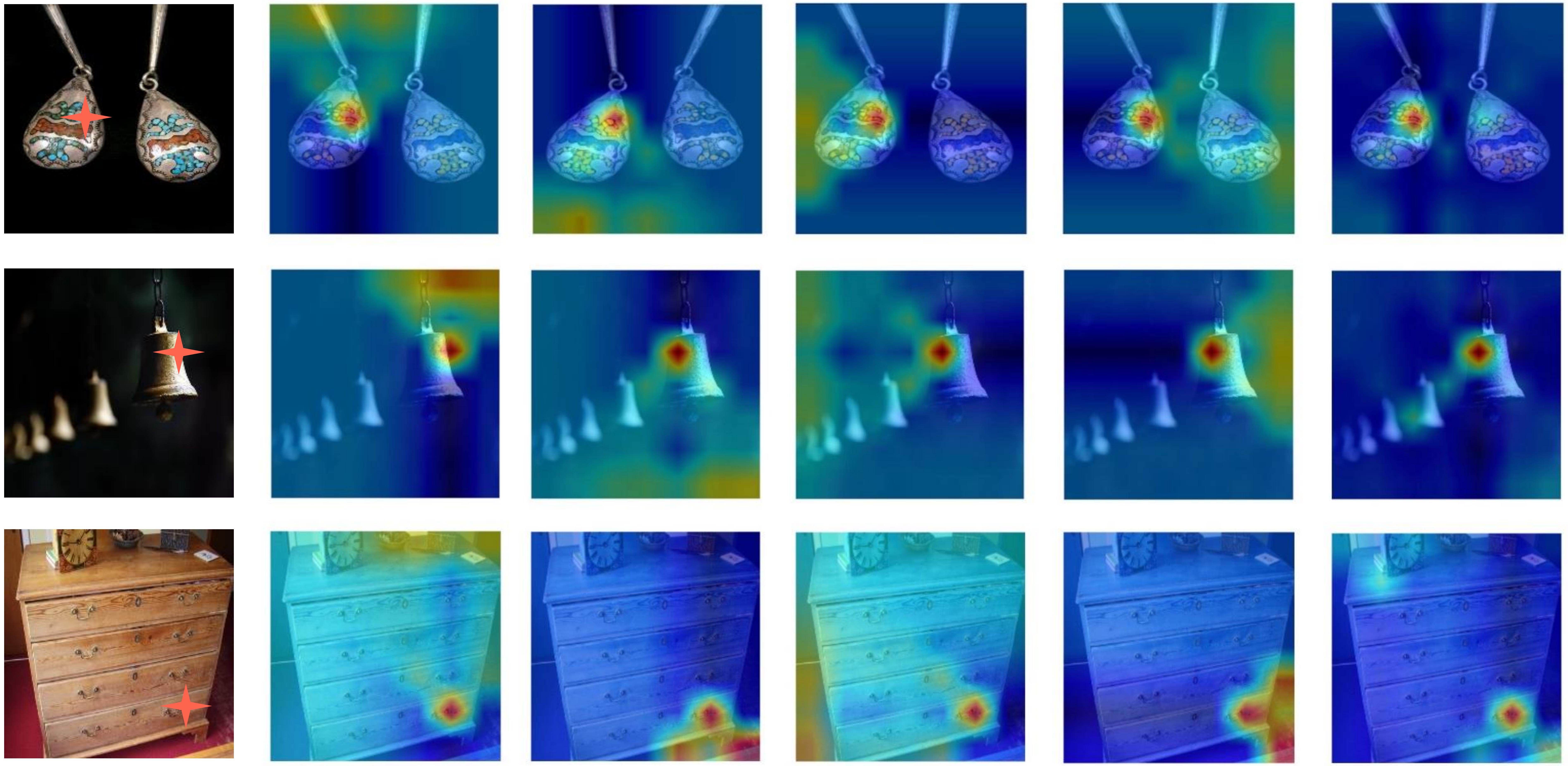
Model	Backbone	Param (M)	IN-1K MAC (G)	IN-1K Acc (%)
ConvNeXT-T [66]	ConvNet	29	4.5	82.1
MambaOut-Tiny [94]	ConvNet	27	4.5	82.7
Swin-T [65]	Transformer	29	4.5	81.3
CSWin-T [23]	Transformer	23	4.3	82.7
CoAtNet-0 [18]	Transformer	25	4.2	81.6
Vim-S [97]	Raster	26	5.1	80.5
VMamba-T [64]	Raster	22	5.6	82.2
Mamba-2D-S [57]	Raster	24	–	81.7
LocalVMamba-T [46]	Raster	26	5.7	82.7
VRWKV-S [26]	Raster	24	4.6	80.1
ViL-S [2]	Raster	23	5.1	81.5
MambaVision-T [34]	Raster	32	4.4	82.3
GSPN-T (Ours)	Line	30	5.3	**83.0**

Model	Backbone	Param (M)	IN-1K MAC (G)	IN-1K Acc (%)
ConvNeXT-S [66]	ConvNet	50	8.7	83.1
MambaOut-Small [94]	ConvNet	48	9.0	**84.1**
T2T-ViT-19 [95]	Transformer	39	8.5	81.9
Focal-Small [92]	Transformer	51	9.1	83.5
NextViT-B [53]	Transformer	45	8.3	83.2
Twins-B [16]	Transformer	56	8.3	83.1
Swin-S [65]	Transformer	50	8.7	83.0
CoAtNet-1 [18]	Transformer	42	8.4	83.3
UniFormer-B [55]	Transformer	50	8.3	83.9
VMamba-S [64]	Raster	44	11.2	83.5
LocalVMamba-S [46]	Raster	50	11.4	83.7
MambaVision-S [34]	Raster	50	7.5	83.3
GSPN-S (Ours)	Line	50	9.0	83.8

Model	Backbone	Param (M)	IN-1K MAC (G)	IN-1K Acc (%)
ConvNeXT-B [66]	ConvNet	89	15.4	83.8
MambaOut-Base [94]	ConvNet	85	15.8	84.2
DeiT-B [82]	Transformer	86	17.5	81.8
Swin-B [65]	Transformer	88	15.4	83.5
CSwin-B [23]	Transformer	78	15.0	84.2
CoAtNet-2 [18]	Transformer	75	15.7	84.1
Vim-B [97]	Raster	98	17.5	81.9
VMamba-B [64]	Raster	89	15.4	83.9
Mamba-2D-B [57]	Raster	92	–	83.0
VRWKV-B [26]	Raster	94	18.2	82.0
ViL-B [2]	Raster	89	18.6	82.4
MambaVision-B [34]	Raster	98	15.0	84.2
GSPN-B (Ours)	Line	89	15.9	**84.3**

Heatmaps

- Anisotropic behavior across four distinct directional scans



Input top-to-bottom bottom-to-top left-to-right right-to-left Avg

Class-conditional Generation

- For 400K iterations, GSPN-XL/2 establishes new SoTA performance
- GSPN-L/2 achieves superior performance with merely 65.6% of the parameters compared to prior models

Class-Conditional ImageNet 256×256						
Model	# Params (M)	FID↓	sFID↓	IS↑	Precision↑	Recall↑
DiT-XL/2 [74]	675	20.05	6.87	64.74	0.621	0.609
U-ViT-H/2 [5]	641	21.71	7.24	62.76	0.608	0.584
PixArt-α-XL/2 [12]	650	24.81	<u>6.38</u>	51.76	0.603	0.615
SiT-XL/2 [68]	675	18.04	**5.17**	73.90	0.630	**0.640**
GSPN-B/2 (Ours)	137	28.70	6.87	50.12	0.585	0.609
GSPN-L/2 (Ours)	443	<u>17.25</u>	8.78	<u>77.37</u>	<u>0.657</u>	0.417
GSPN-XL/2 (Ours)	690	**15.26**	6.51	**85.99**	**0.670**	<u>0.624</u>

Text-to-image Generation

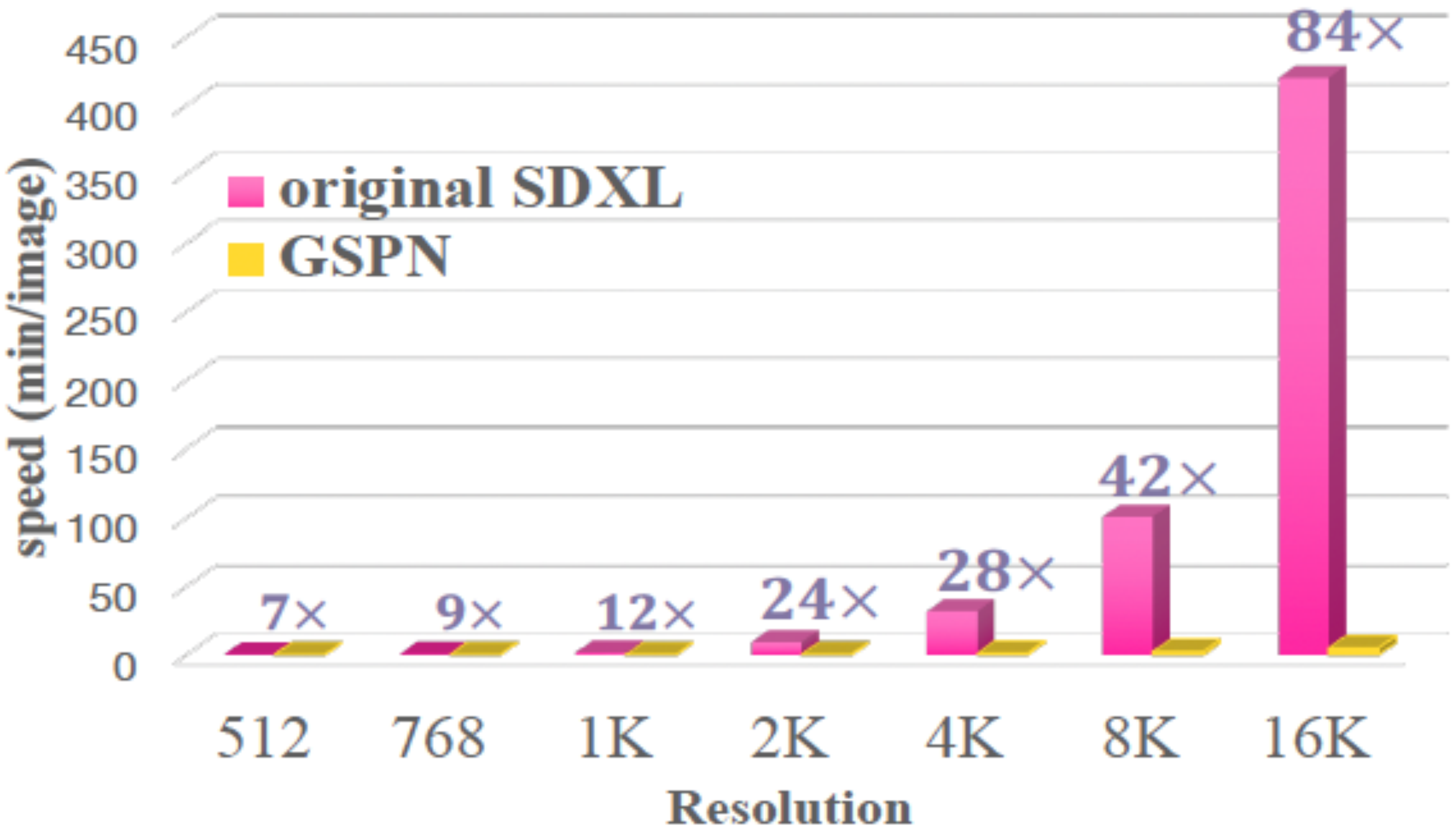
- Train on the COCO benchmark using 512 X 512 resolution
- Infer at 1024 X 1024 resolution (unseen during training)

Model	FID(↓)	CLIP-T(↑)
SD-v1.5 (baseline)	32.71	0.290
Mamba [31] (w/ norm)	50.30	0.263
Mamba2 [19] (w/ norm)	37.02	0.273
Linfusion [63] (w/ norm)	36.33	0.285
SD-v1.5-GSPN w/o init (Ours)	36.89	0.278
SD-v1.5-GSPN (Ours)	**30.86**	**0.307**

SD-v1.5

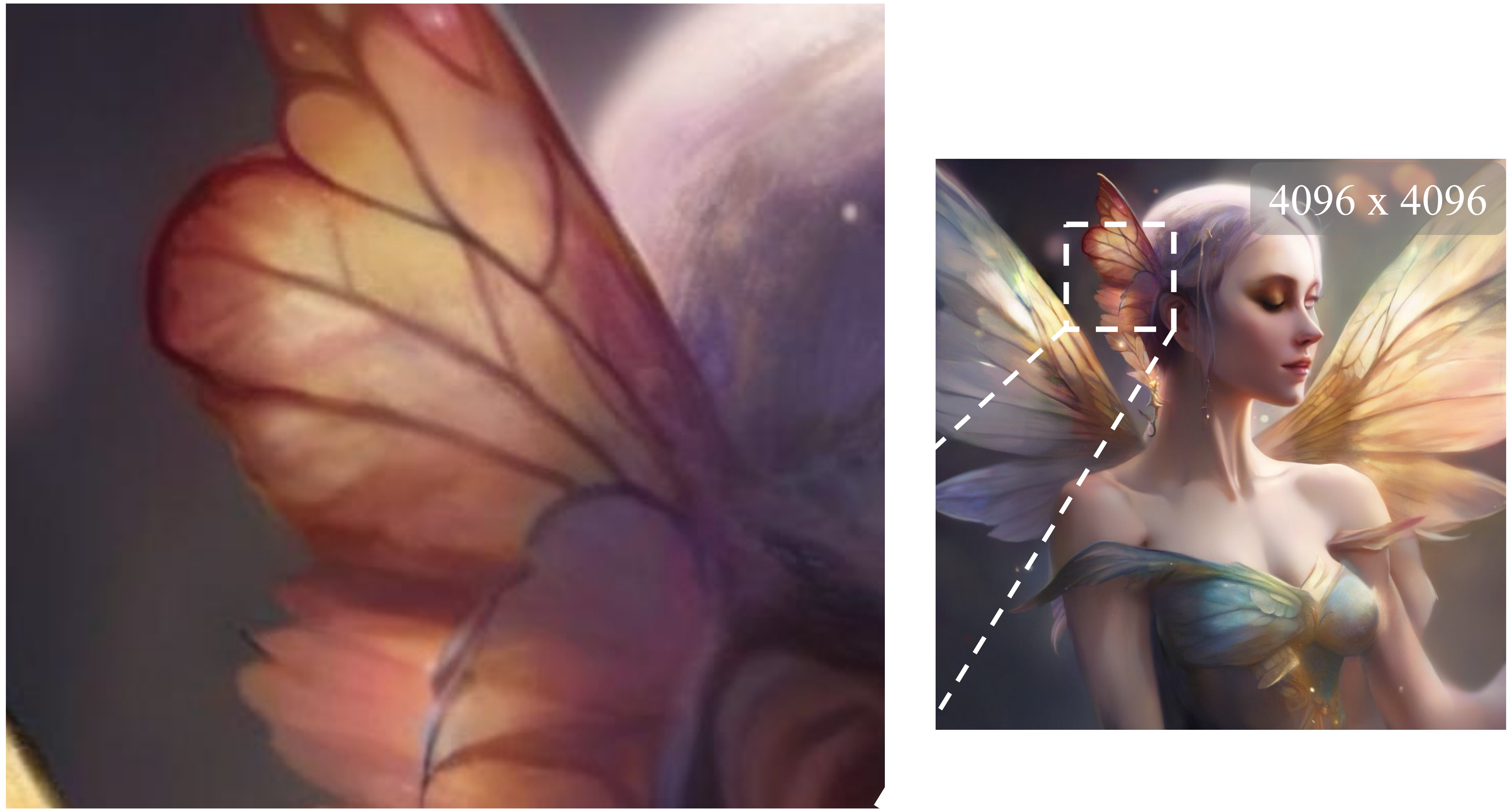
SD-XL

- Speed comparison from 512×512 to $16K \times 8K$ resolution
- GSPN achieves ~84× speedup at $16K \times 8K$ resolution

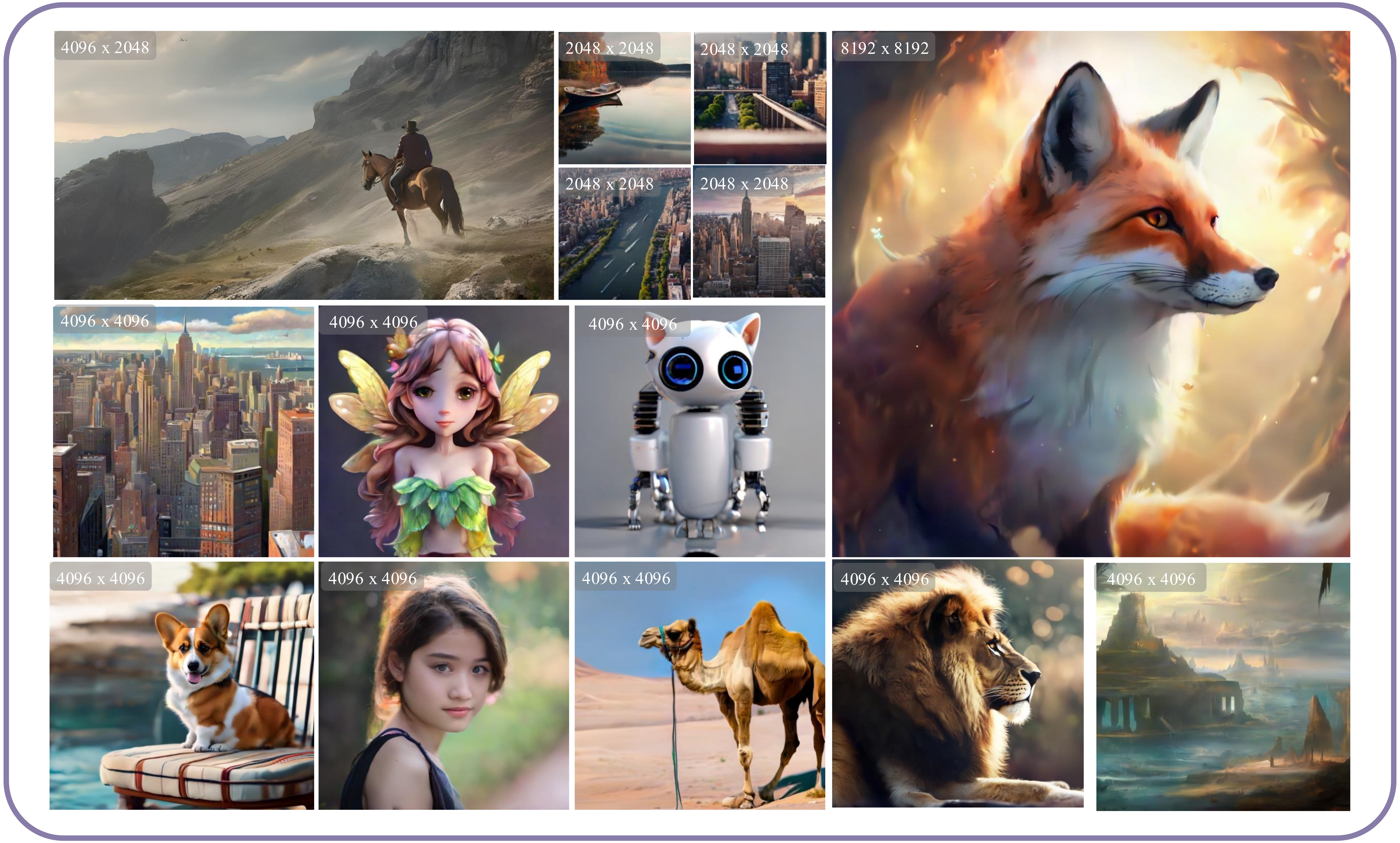


SD-XL

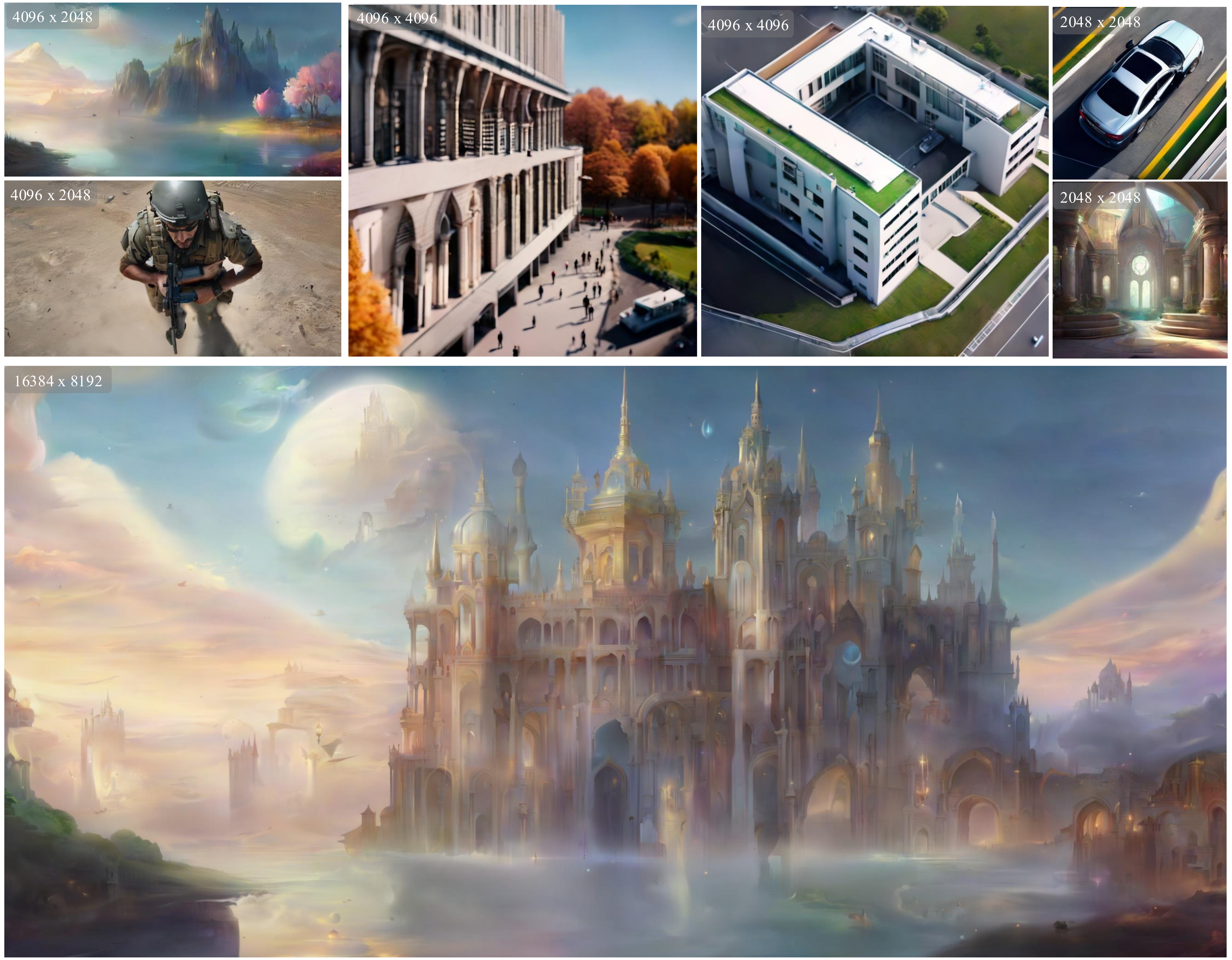
SD-XL



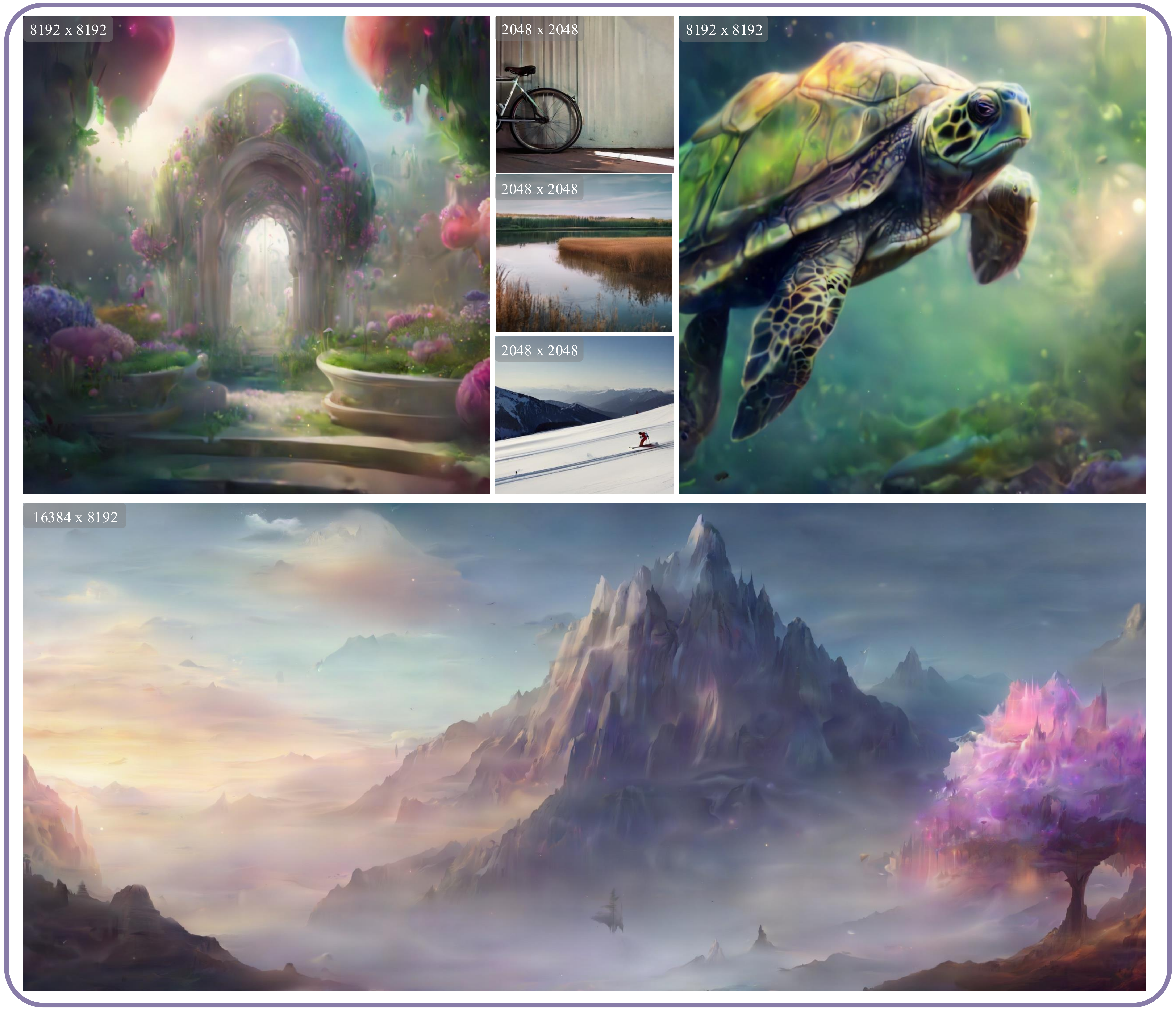
More results



More results



More results



Thank you!