NEURAL INFORMATION PROCESSING SYSTEMS

ScaleDiff:

Higher-Resolution Image Synthesis via Efficient and Model-Agnostic Diffusion

Sungho Koh, SeungJu Cha, Hyunwoo Oh,
Kwanyoung Lee, Dong-Jin Kim

Challenge

Text-to-Image diffusion model at higher resolution

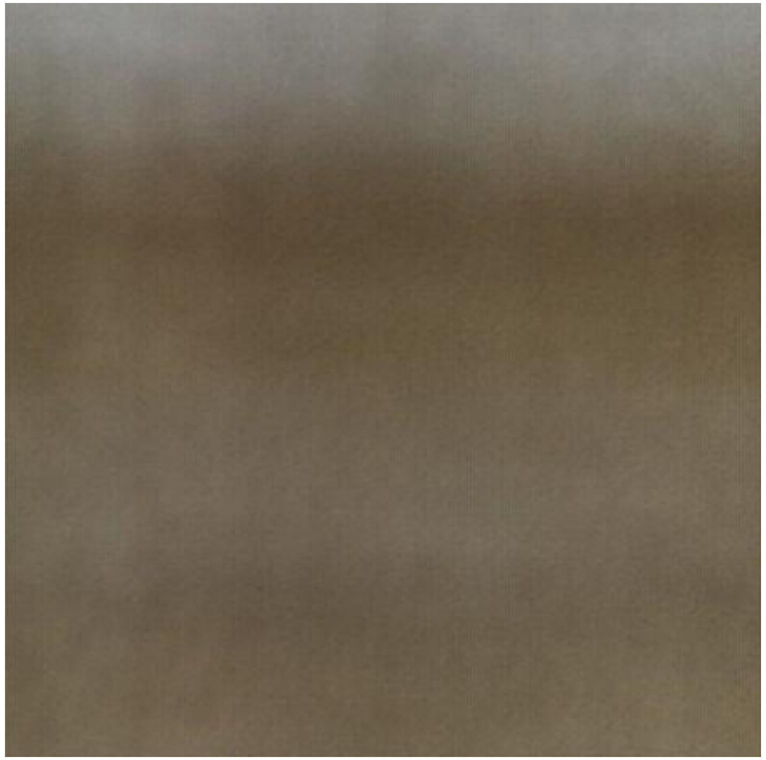
SDXL (1024x1024)

FLUX (1024x1024)

Text-to-Image diffusion model at higher resolution

- U-Net: Exhibit **repetitive artifacts** and loss of global structure yet retain the capacity to generate local details.
- DiT: Demonstrate complete failure in generating **both global structure and local detail** at higher resolutions.

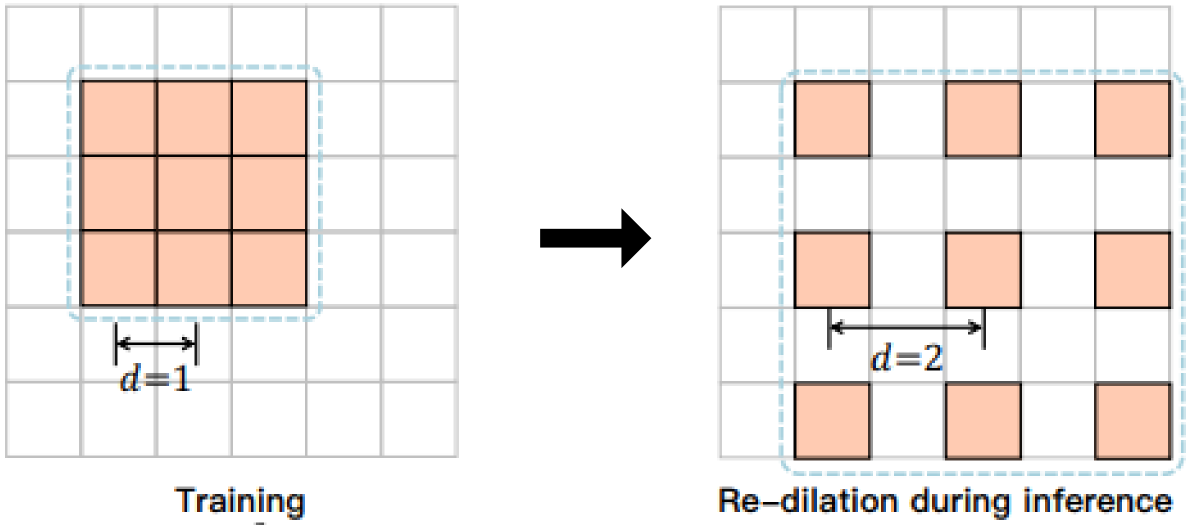
SDXL (4096x4096)



FLUX (4096x4096)

Dilated convolution-based methods

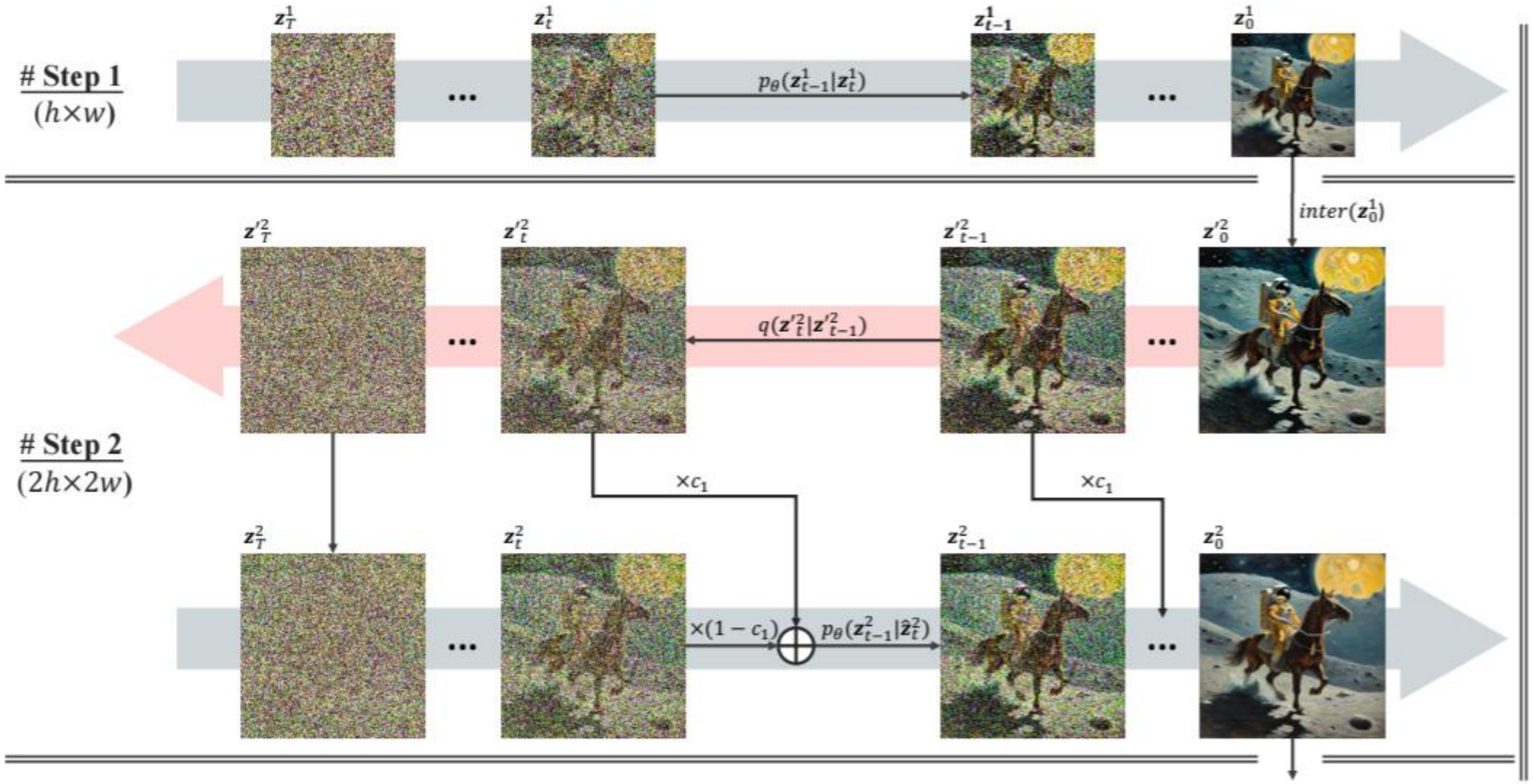
- Limited Receptive field of convolutional layer causes repetitive artifacts.
- Use dilated convolution to expand receptive field.
- **Only compatible with U-Net architectures.**



He et al., Scalecrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models, ICLR 2024.

SDEdit-based methods

- Upsample low resolution image → diffuse → denoise
- **Relies on local detail generation capability**, which pretrained DiT lacks at high resolution.



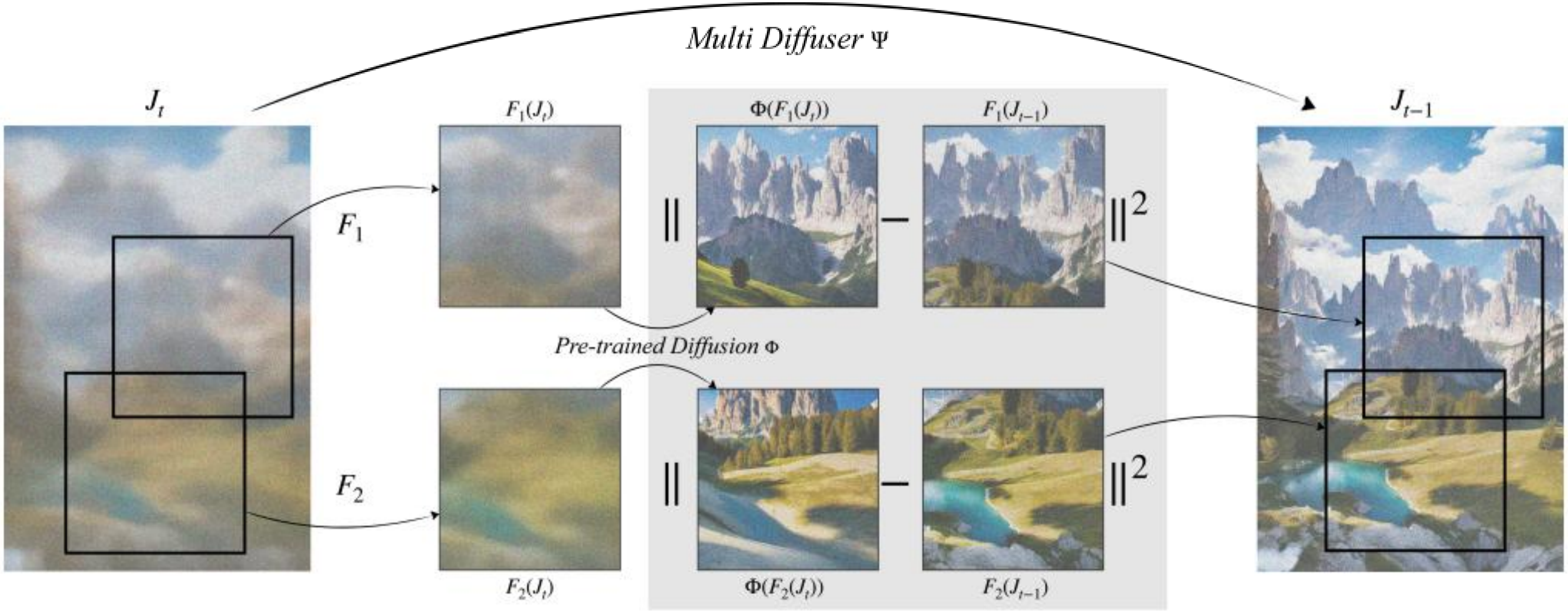
(a) Progressive upscaling with skip residual.

Meng et al., SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations, ICLR 2022.

Du et al., DemoFusion: Democratising High-Resolution Image Generation With No $$$, CVPR 2024.

Patch-based methods

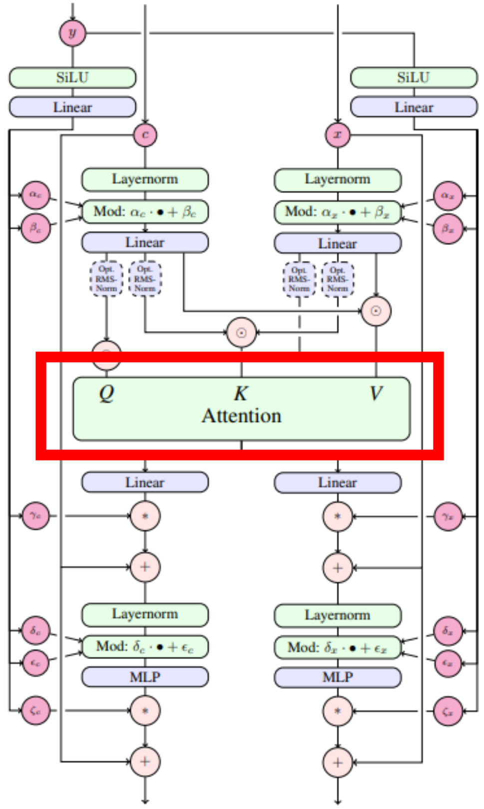
- Divide high resolution canvas into overlapping low-resolution patches.
- Independently denoise each patch and combine their outputs.
- **Massive computational redundancy** due to necessary overlap to ensure smooth transition across patch boundary.



Bar-Tal et al., Multidiffusion: Fusing Diffusion Paths for Controlled Image Generation, ICML 2023.

Neighborhood Patch Attention (NPA)

- Only self-attention (with positional encoding) is directly affected by resolution increase
- Non-self-attention layers (MLP, CNN, Cross-Attn) perform operations on individual tokens or local regions.



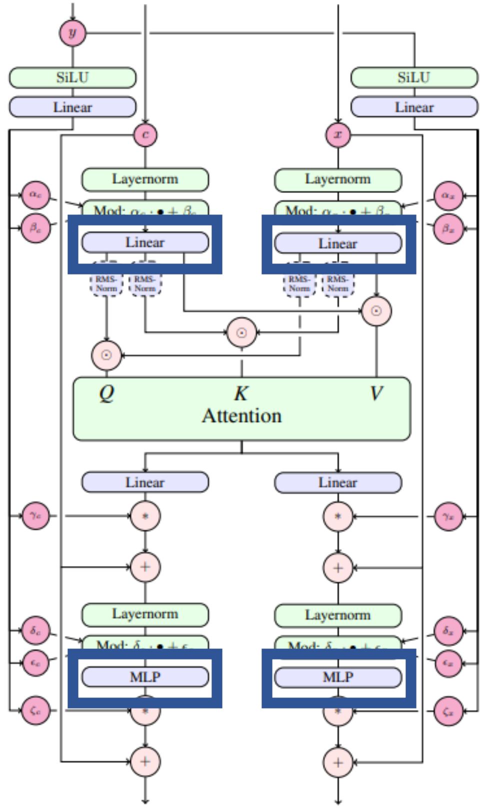
(b) One *MM-DiT* block

Patch based processing only in self-attention

Esser, Patrick, et al., Scaling rectified flow transformers for high-resolution image synthesis, ICML 2024.

Neighborhood Patch Attention (NPA)

- Only self-attention (with positional encoding) is directly affected by resolution increase
- Non-self-attention layers (MLP, CNN, Cross-Attn) perform operations on individual tokens or local regions.



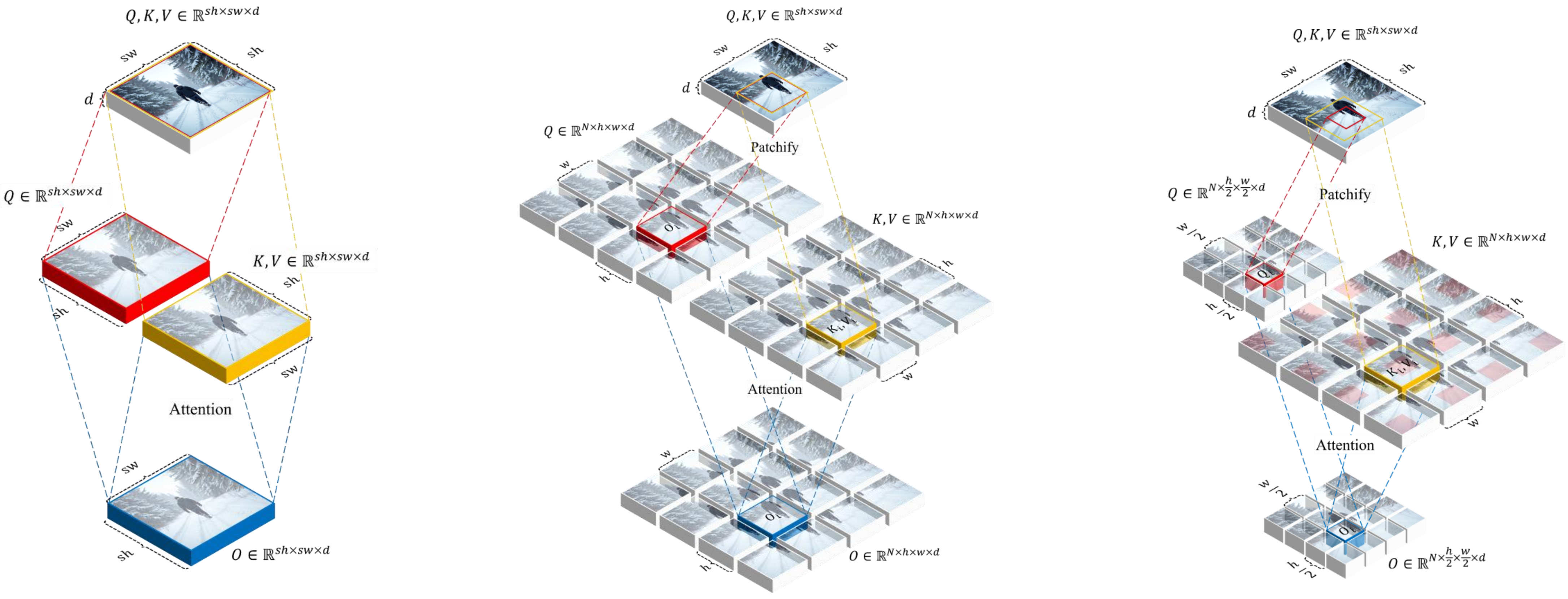
(b) One *MM-DiT* block

Process full high-resolution tensor in a single pass.

Esser, Patrick, et al., Scaling rectified flow transformers for high-resolution image synthesis, ICML 2024.

Neighborhood Patch Attention (NPA)

- **Non-overlapping** query patches → eliminate redundant computation.
- Extract **overlapping** key-value patch from query's **spatial neighborhood.**
- Overlap between these patches allows every query patch to attend to a wider context, ensuring smooth transitions.



Base **Multidiffusion** **NPA (Ours)**

Neighborhood Patch Attention (NPA)

- **Non-overlapping** query patches → eliminate redundant computation.
- Extract **overlapping** key-value patch from query's **spatial neighborhood.**
- Overlap between these patches allows every query patch to attend to a wider context, ensuring smooth transitions.

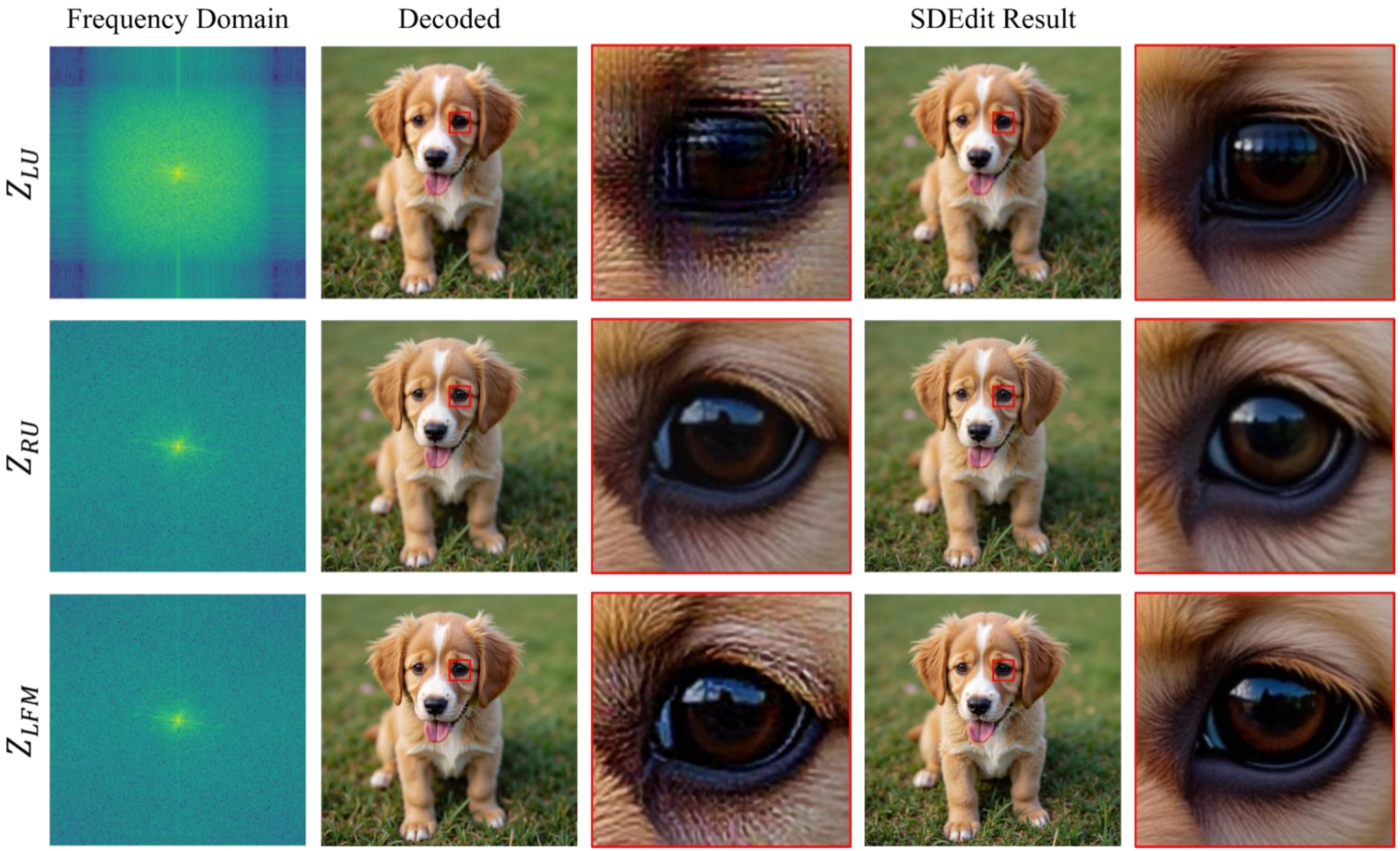
Method	Linear	Conv	Cross-Attn	Self-Attn
Base	s^2hwd^2	$s^2hwk^2d^2$	s^2hwld	$s^4h^2w^2d$
MultiDiffusion	$(2s-1)^2hwd^2$	$(2s-1)^2hwk^2d^2$	$(2s-1)^2hwld$	$(2s-1)^2h^2w^2d$
NPA(Ours)	s^2hwd^2	$s^2hwk^2d^2$	s^2hwld	$s^2h^2w^2d$

ScaleDiff Upscaling Pipeline

- Integrate NPA into **SDEdit pipeline**.
- Starting from low resolution image, Upample → Diffuse → Denoise.

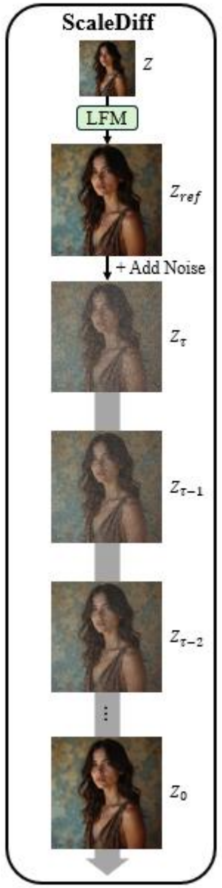
Latent Frequency Mixing (LFM)

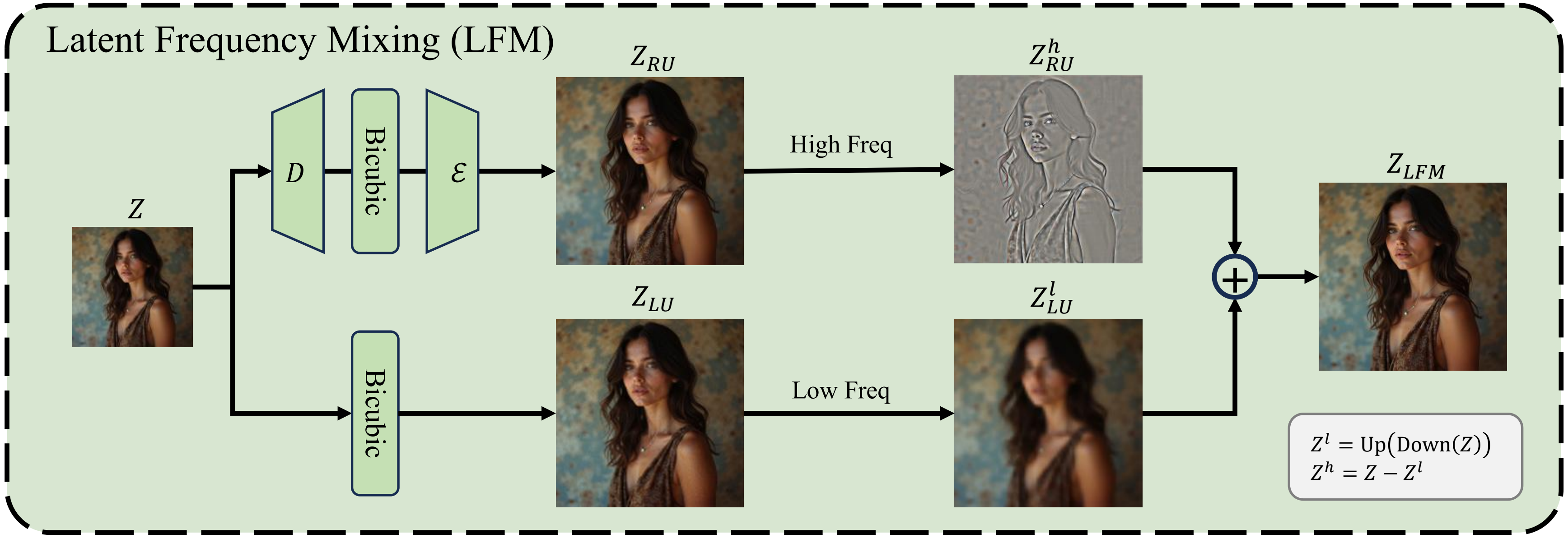
- Upsample in Latent-Space (Z_{LU}) : Lack of high frequencies → **decoding artifacts**. No oversmoothing bias.
- Upsample in RGB-Space (Z_{RU}) : Biases the model toward **oversmoothed** results. No decoding artifacts



Latent Frequency Mixing (LFM)

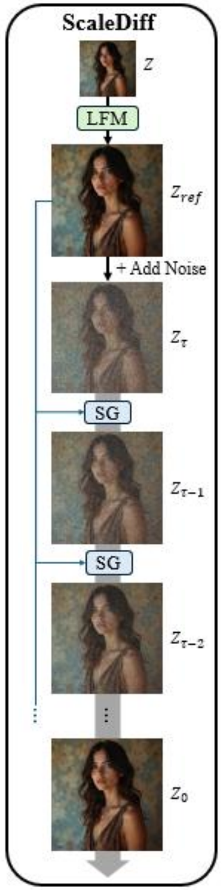
- $Z_{ref} = Z_{LU}^{l} + Z_{RU}^{h}$
- **Low-frequency** components are obtained from Z_{LU} to alleviate oversmoothing bias.
- While **high-frequency** components are utilized from Z_{RU} to avoid decoding artifacts..

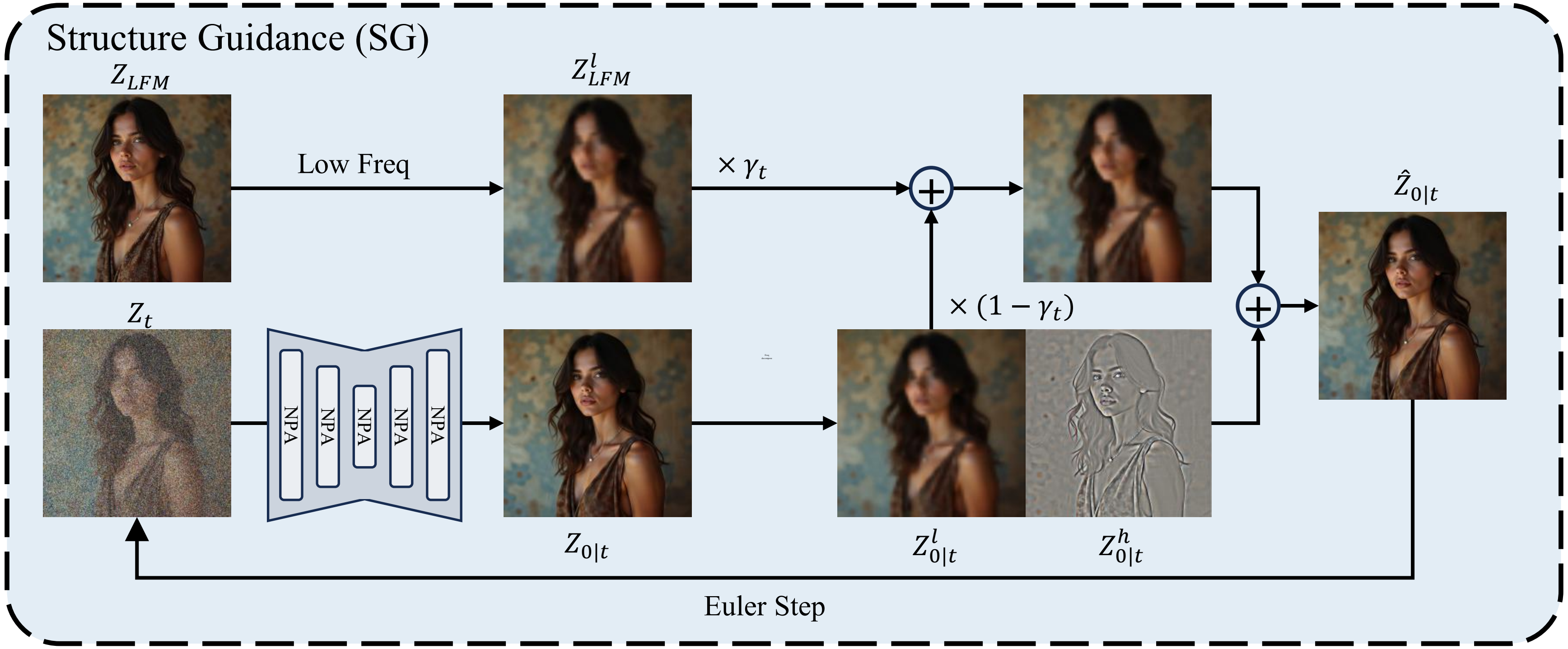




Structure Guidance (SG)

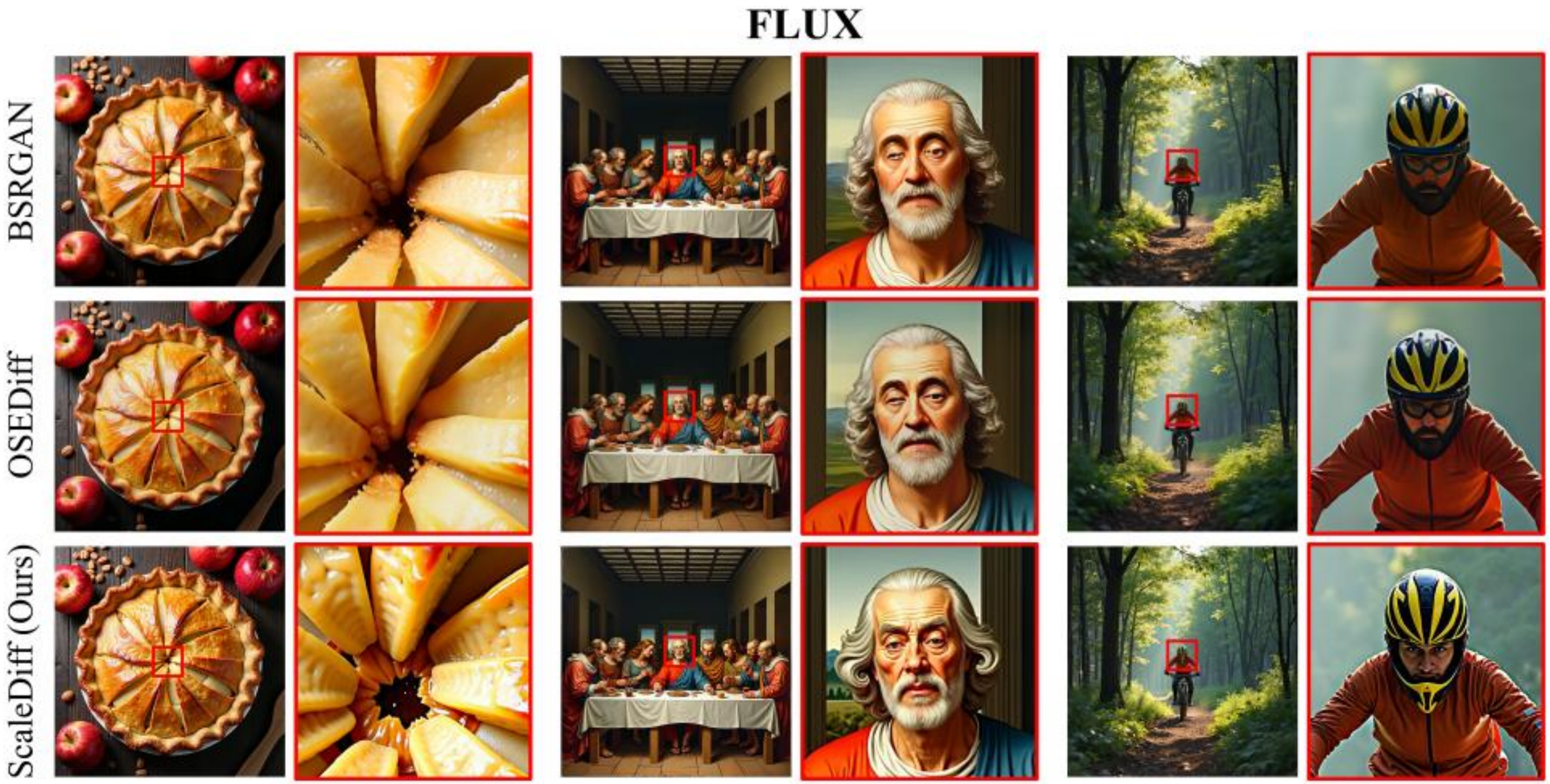
- $\hat{Z}_{\mathbf{0|t}} = Z^h_{0|t} + (1-\gamma_t)Z^l_{0|t} + \gamma_t Z^l_{ref}$
- Guide **low-frequency components** to the reference to enforce global structural consistency.





Qualitative Comparison

Qualitative Comparison



Quantitative Comparison

Model	Resolution	Method	FID ↓	KID ↓	IS ↑	FID_p ↓	KID_p ↓	IS_p ↑	CLIP ↑	Time ↓
SDXL	2048^2	SDXL Direct	88.56	0.0124	13.25	58.73	0.0137	20.79	31.57	47
		SDXL + BSRGAN	64.60	0.0041	18.40	41.40	0.0092	23.19	33.03	**13**
		SDXL + OSEDiff	64.79	0.0046	18.89	41.76	0.0094	23.58	32.79	29
		UltraPixel	64.61	0.0056	18.58	42.44	0.0093	25.15	32.61	71
		ScaleCrafter	68.68	0.0033	16.56	43.46	0.0064	23.52	32.07	64
		HiDiffusion	69.52	0.0040	18.22	42.92	0.0067	24.01	31.50	33
		DiffuseHigh	63.27	0.0033	19.10	38.15	0.0062	24.95	32.77	45
		FreeScale	63.50	**0.0031**	19.06	38.27	0.0062	24.25	32.62	69
		AccDiffusion v2	64.86	0.0039	18.37	38.24	0.0068	25.66	32.62	199
		Demofusion	63.36	0.0032	19.15	**35.98**	**0.0050**	**26.42**	32.72	125
		ScaleDiff (Ours)	**62.98**	0.0032	**19.54**	38.03	0.0067	25.70	**33.11**	31
	4096^2	SDXL Direct	182.05	0.0717	7.99	80.80	0.0250	17.68	27.82	328
		SDXL + BSRGAN	64.88	0.0044	18.16	48.97	0.0160	17.04	33.02	**14**
		SDXL + OSEDiff	65.35	0.0045	18.69	45.67	0.0118	17.61	32.88	122
		UltraPixel	65.39	0.0055	19.08	47.09	0.0112	**20.64**	32.33	386
		ScaleCrafter	86.66	0.0110	15.14	79.39	0.0217	14.47	30.25	932
		HiDiffusion	105.37	0.0216	13.87	112.30	0.0494	12.22	27.21	124
		DiffuseHigh	63.91	0.0034	18.99	42.30	0.0079	19.54	32.68	325
		FreeScale	64.33	0.0036	19.18	39.56	**0.0079**	18.91	32.56	517
		AccDiffusion v2	64.64	0.0037	18.56	40.92	0.0083	18.42	32.34	1599
		Demofusion	65.06	0.0041	19.13	41.29	0.0079	19.59	32.61	1005
		ScaleDiff (Ours)	**61.87**	**0.0025**	**19.56**	**38.89**	0.0080	20.41	**33.04**	113

ScaleDiff:

Higher-Resolution Image Synthesis via Efficient and Model-Agnostic Diffusion

Sungho Koh, SeungJu Cha, Hyunwoo Oh,
Kwanyoung Lee, Dong-Jin Kim

Code Available