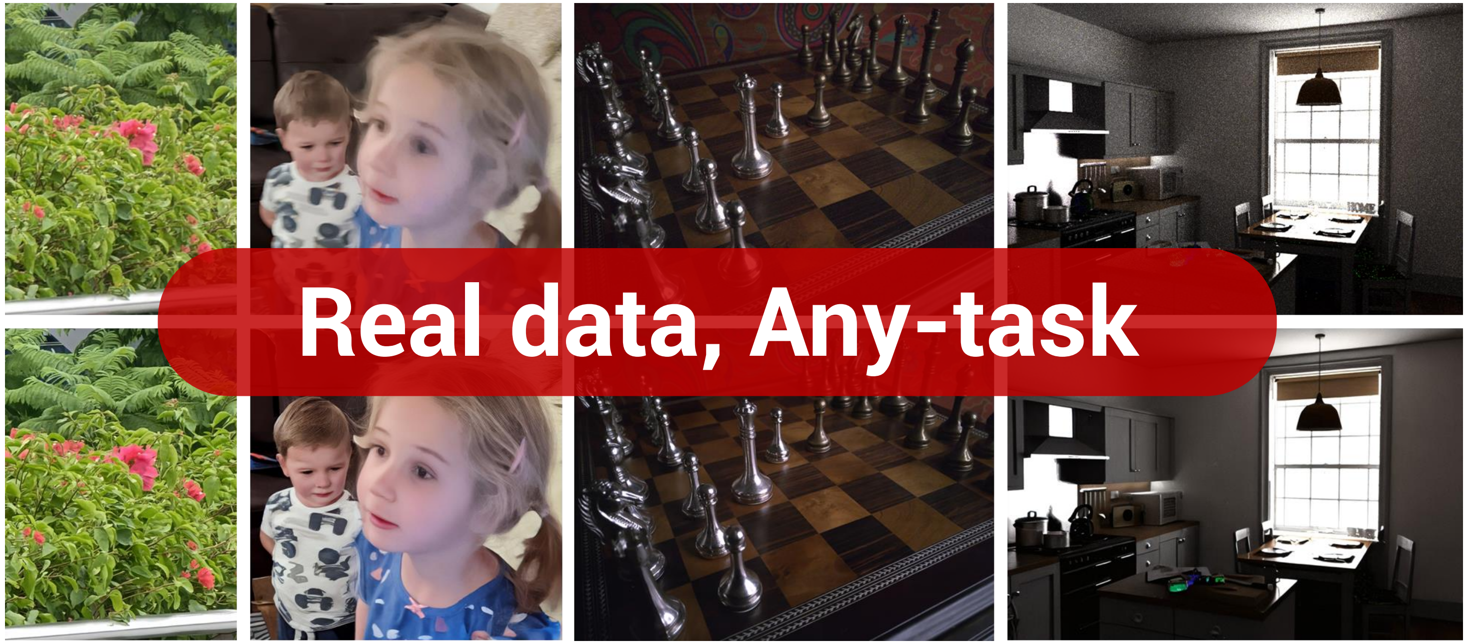
Generative Image Restoration:
from regression to generation

Zhixiang Wang

PhD student
University of Tokyo

Outline



Outline

Part I: image restoration with diffusion prior

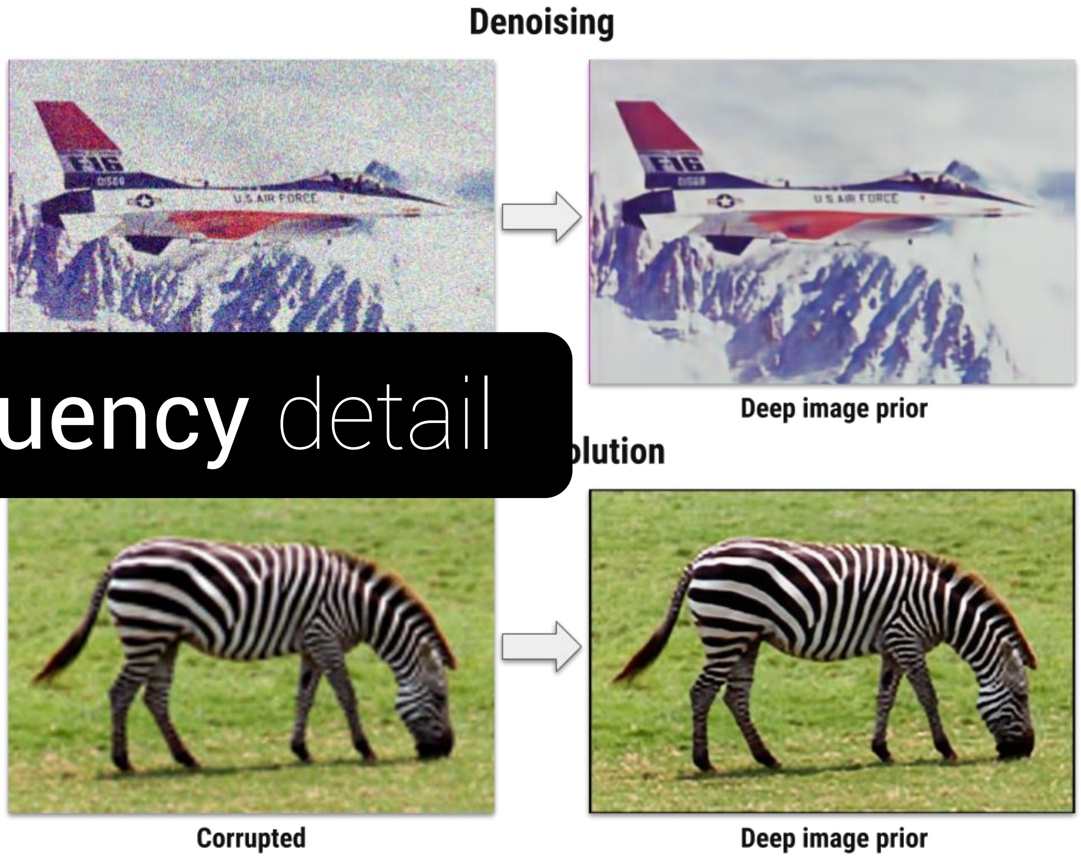
Concurrent works

StableSR, DiffBIR

Motivation: details

- **Existing works:** supervised learning and self-supervised method

missing **high-frequency** detail



Motivation: details

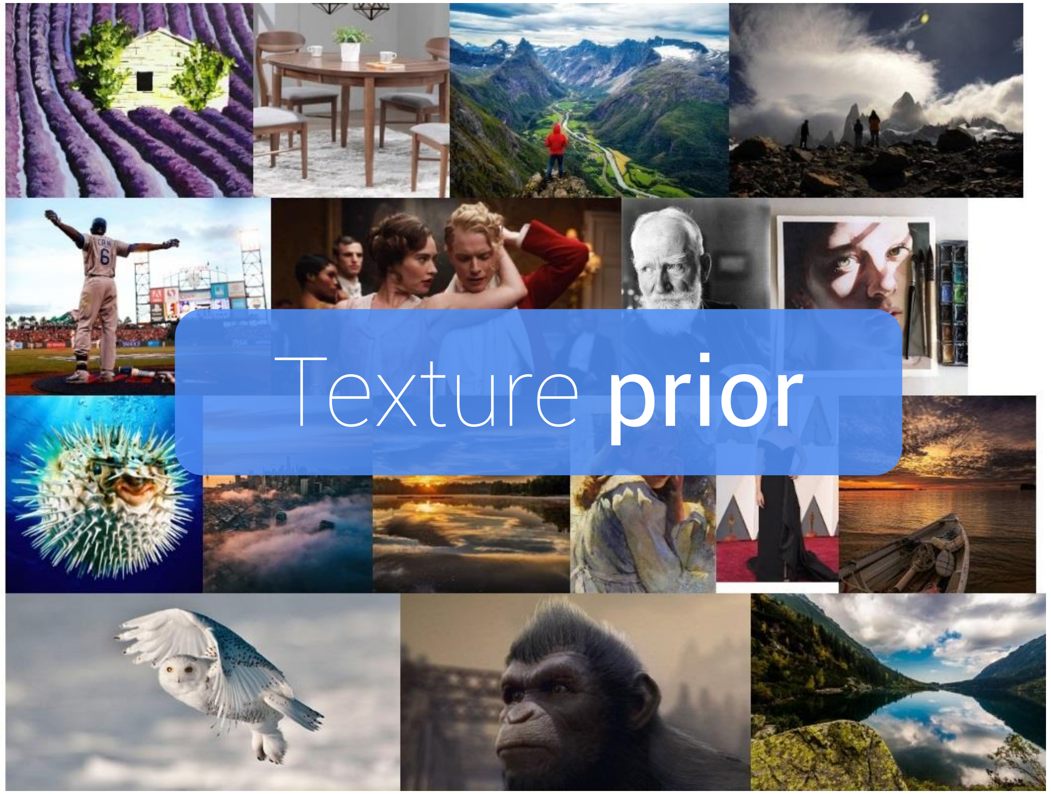
- **Existing works:** +class specific generative prior
  - e.g., GLEAN, CodeFormer (can only process specific classes)

Real-ESRGAN Face | CodeFormer | CodeFormer (fidelity) | GT

The opportunity raising by stable diffusion

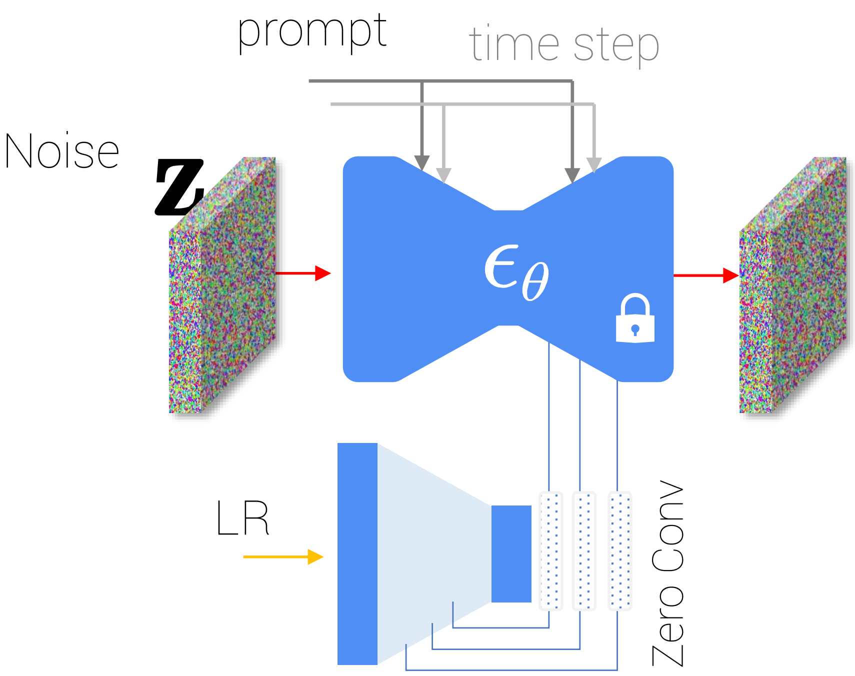
- Training data
  - Small size (700 K) ➔ Huge Size (5 B)
  - Restricted ➔ Unrestricted
    - 1 class ➔ many classes
    - Cropped ➔ uncropped

StyleGAN face



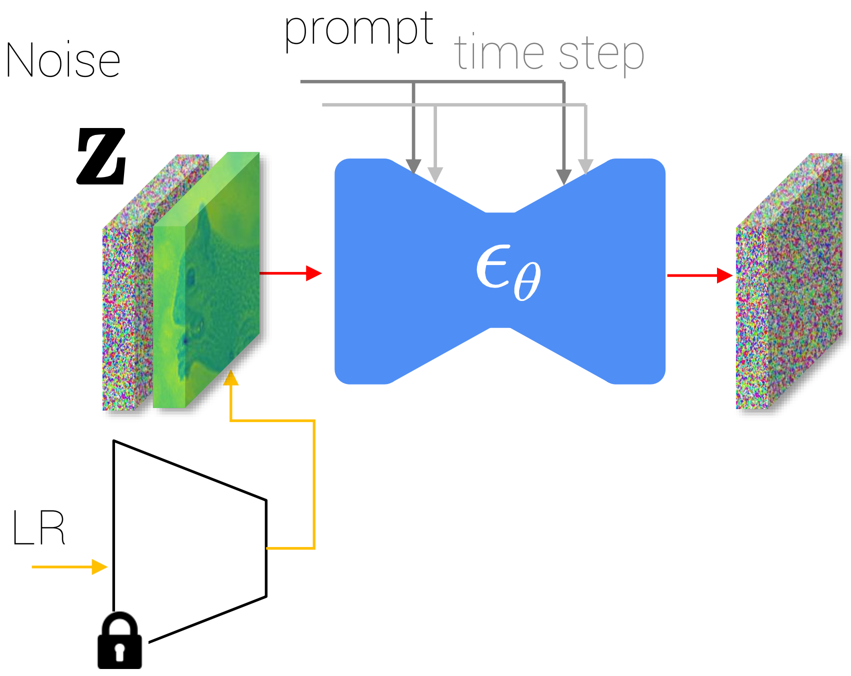
LAION-5B

How to use it? **Fine-tuning**



ControlNet Style

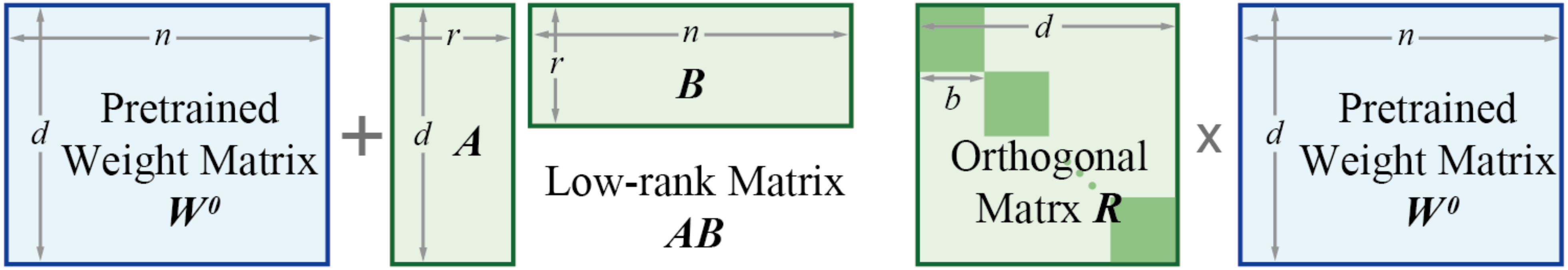
StableSR, DiffBIR



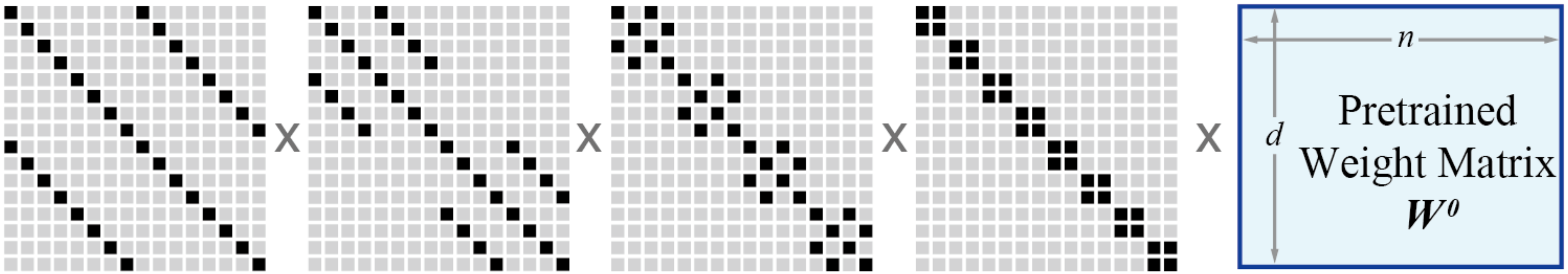
ConCat Style

Instruct-Pix2Pix, Ours

Other fine-tune methods



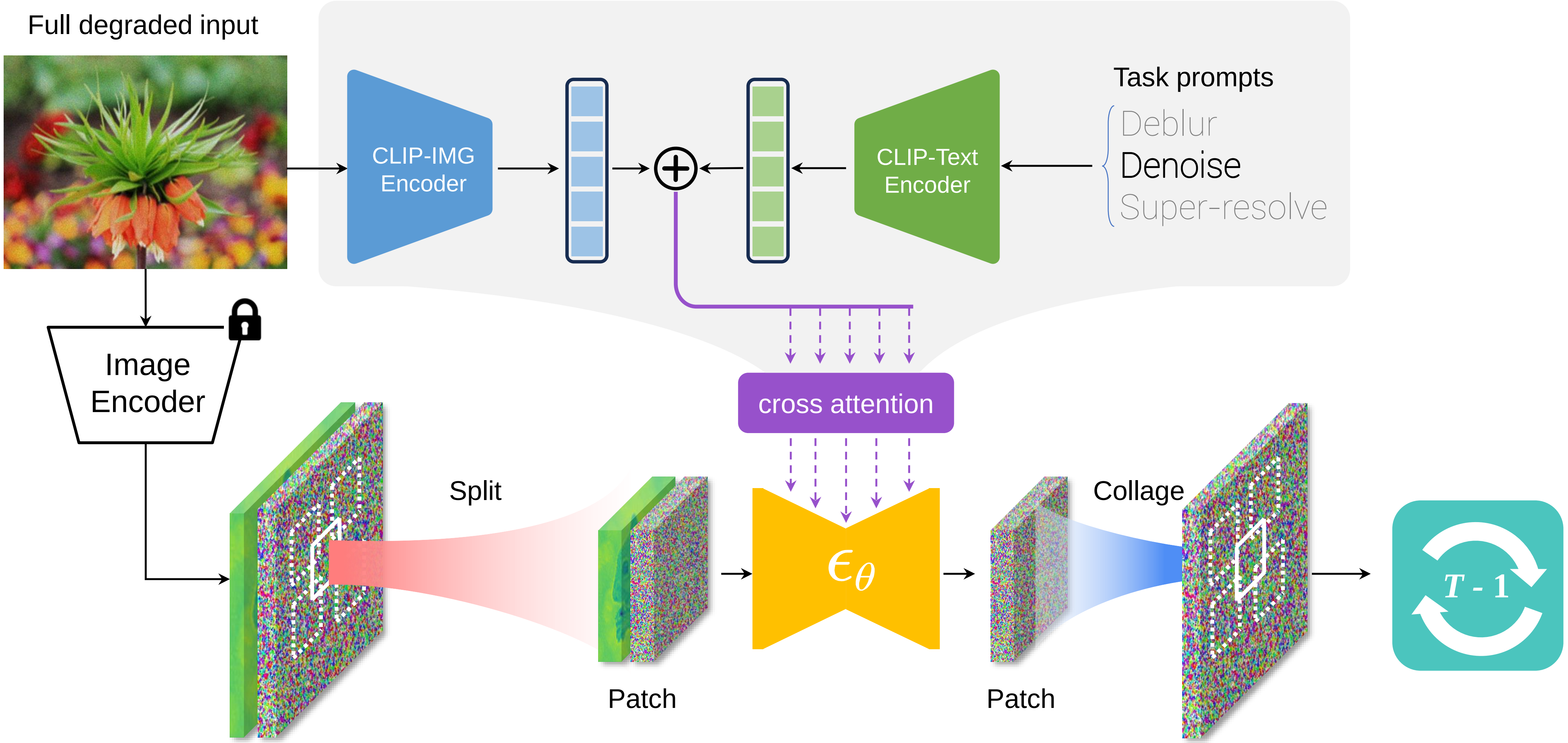
(a) Low-rank Structure in LoRA (b) Orthogonal Structure in OFT



(c) Butterfly Orthogonal Structure in BOFT

https://boft.wyliu.com/

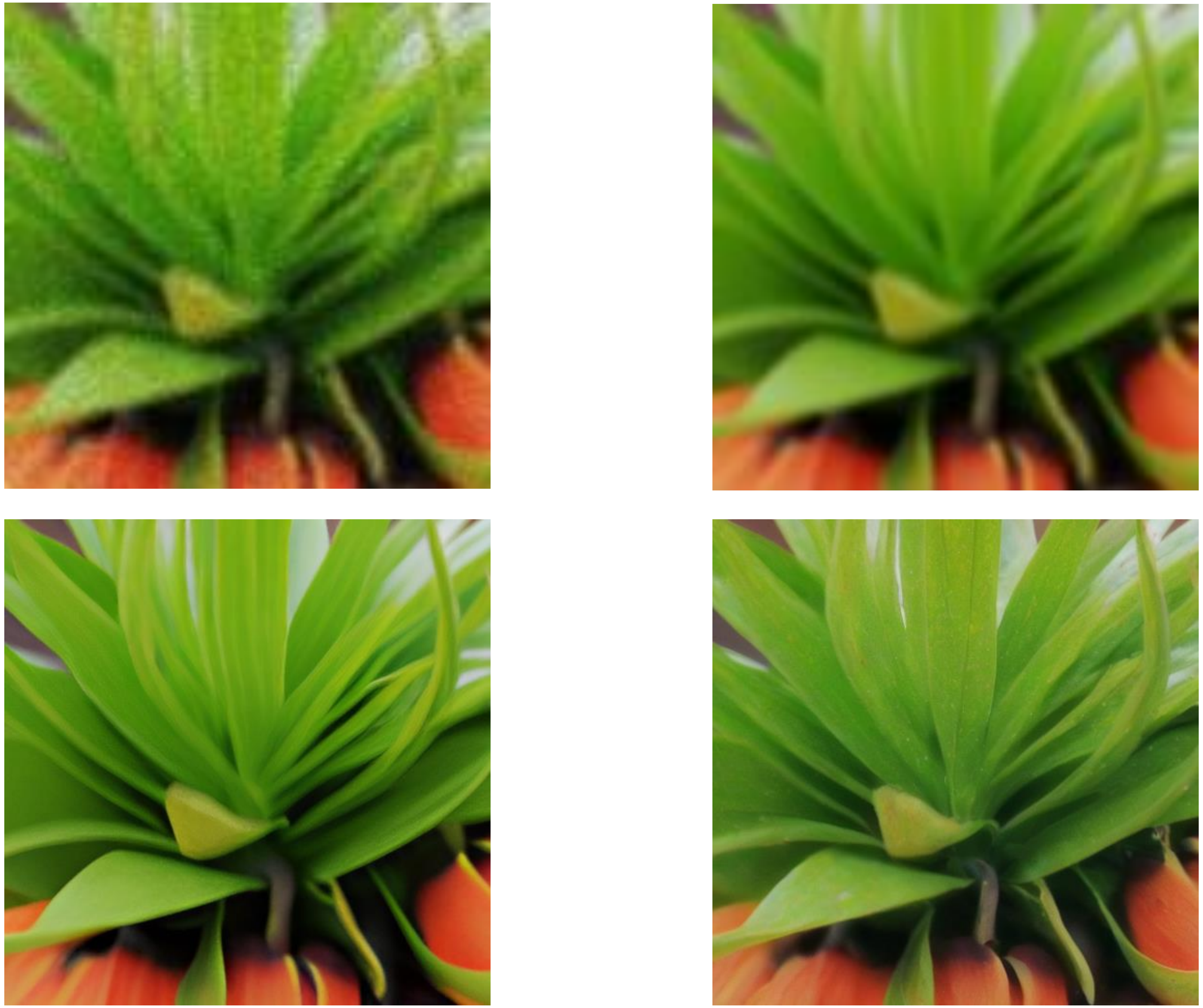
Method: fine-tune patch-based diffusion



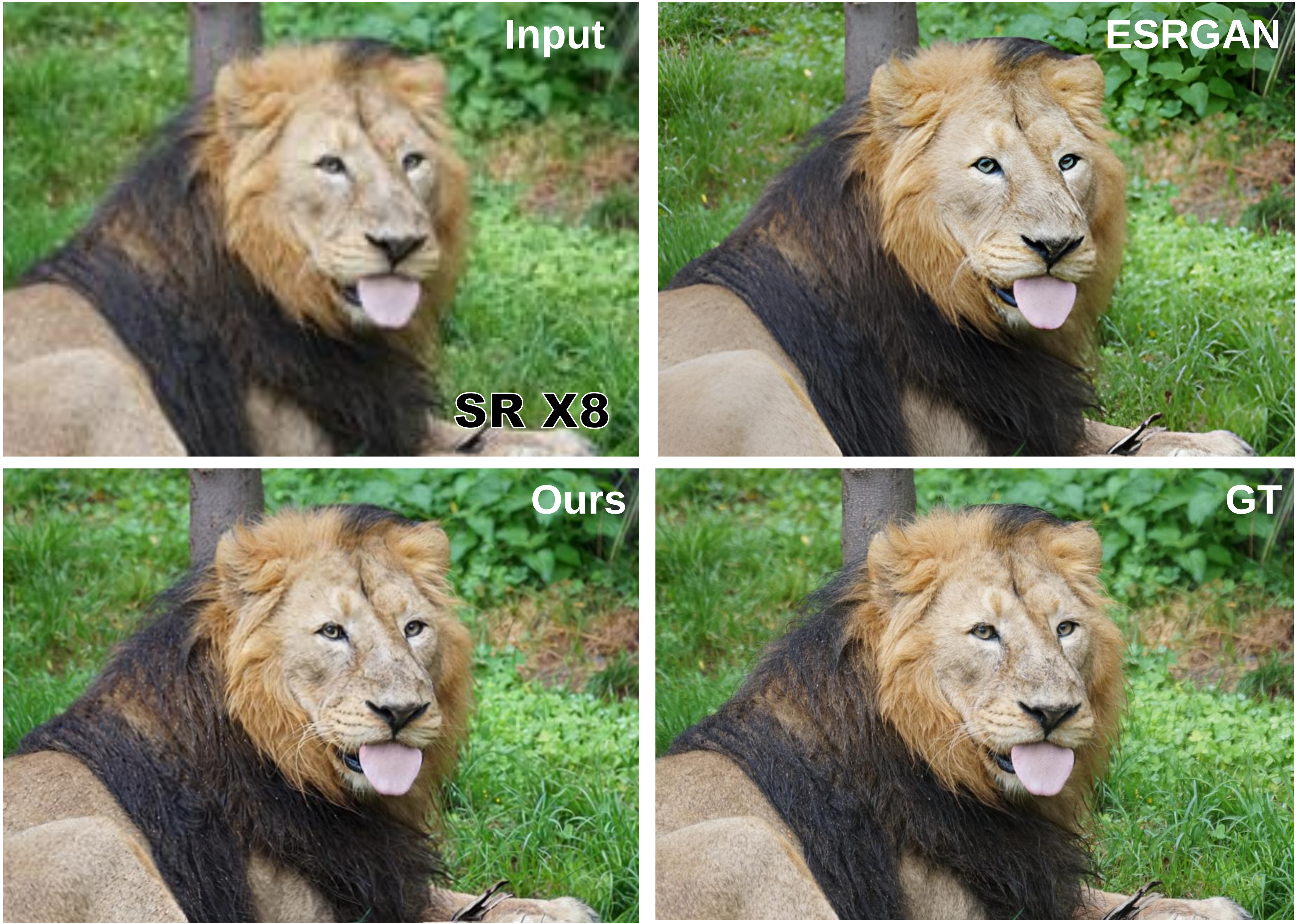
Input SCUNET

SwinIR Ours

Input SCUNET

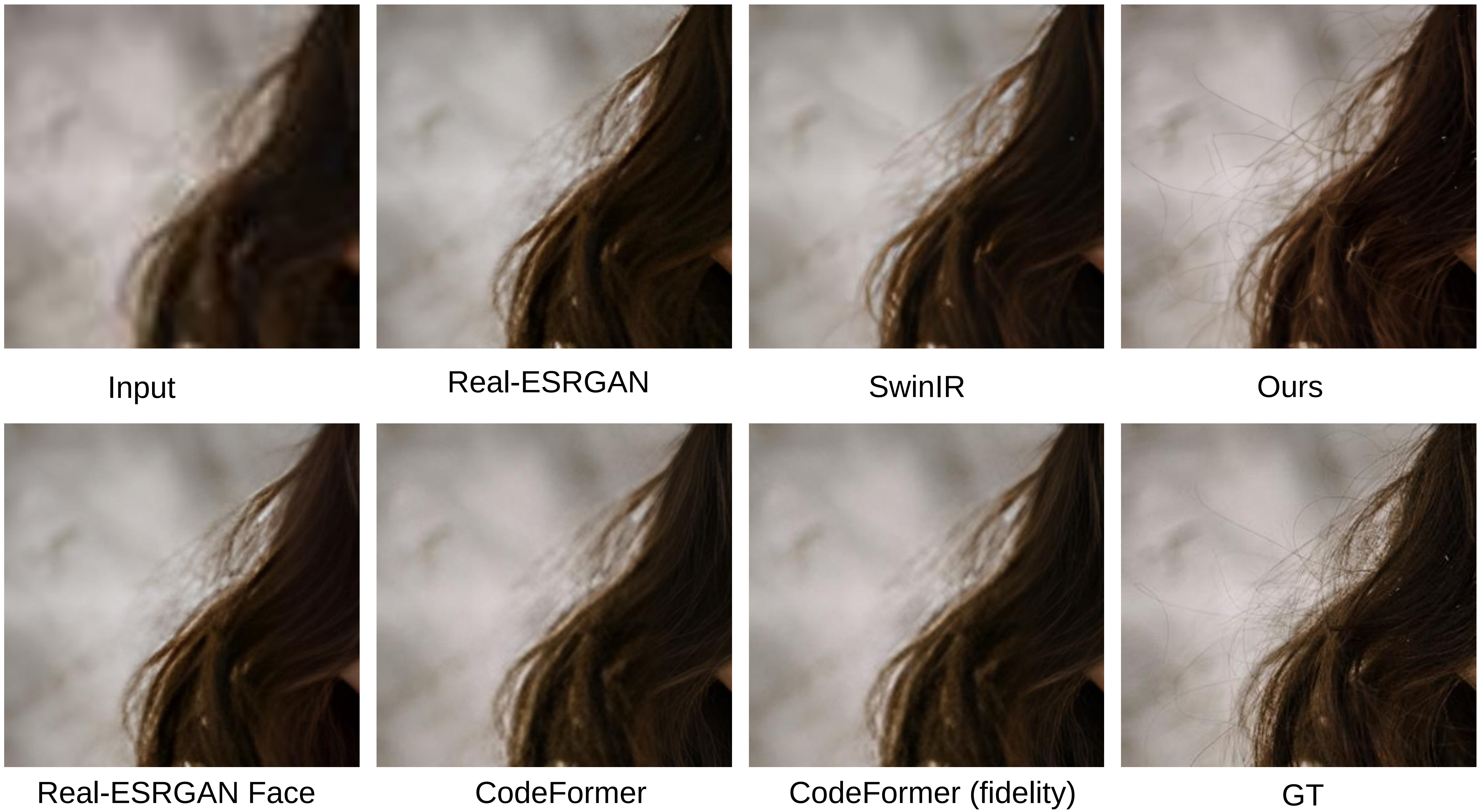


SwinIR Ours


Input
SR X8
ESRGAN
Ours
GT

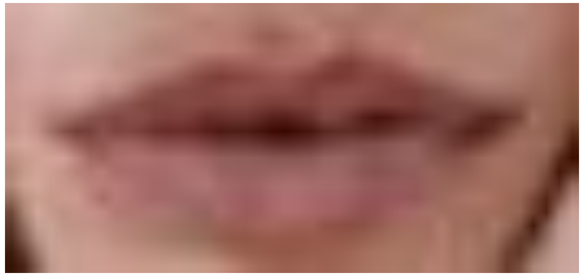
Input Real-ESRGAN SwinIR Ours

Real-ESRGAN Face CodeFormer CodeFormer (fidelity) GT


Input
Real-ESRGAN
SwinIR
Ours
Real-ESRGAN Face
CodeFormer
CodeFormer (fidelity)
GT

Input | Real-ESRGAN | SwinIR | Ours

Real-ESRGAN Face | CodeFormer | CodeFormer (fidelity) | GT



Input

Real-ESRGAN

SwinIR

Ours

Real-ESRGAN Face

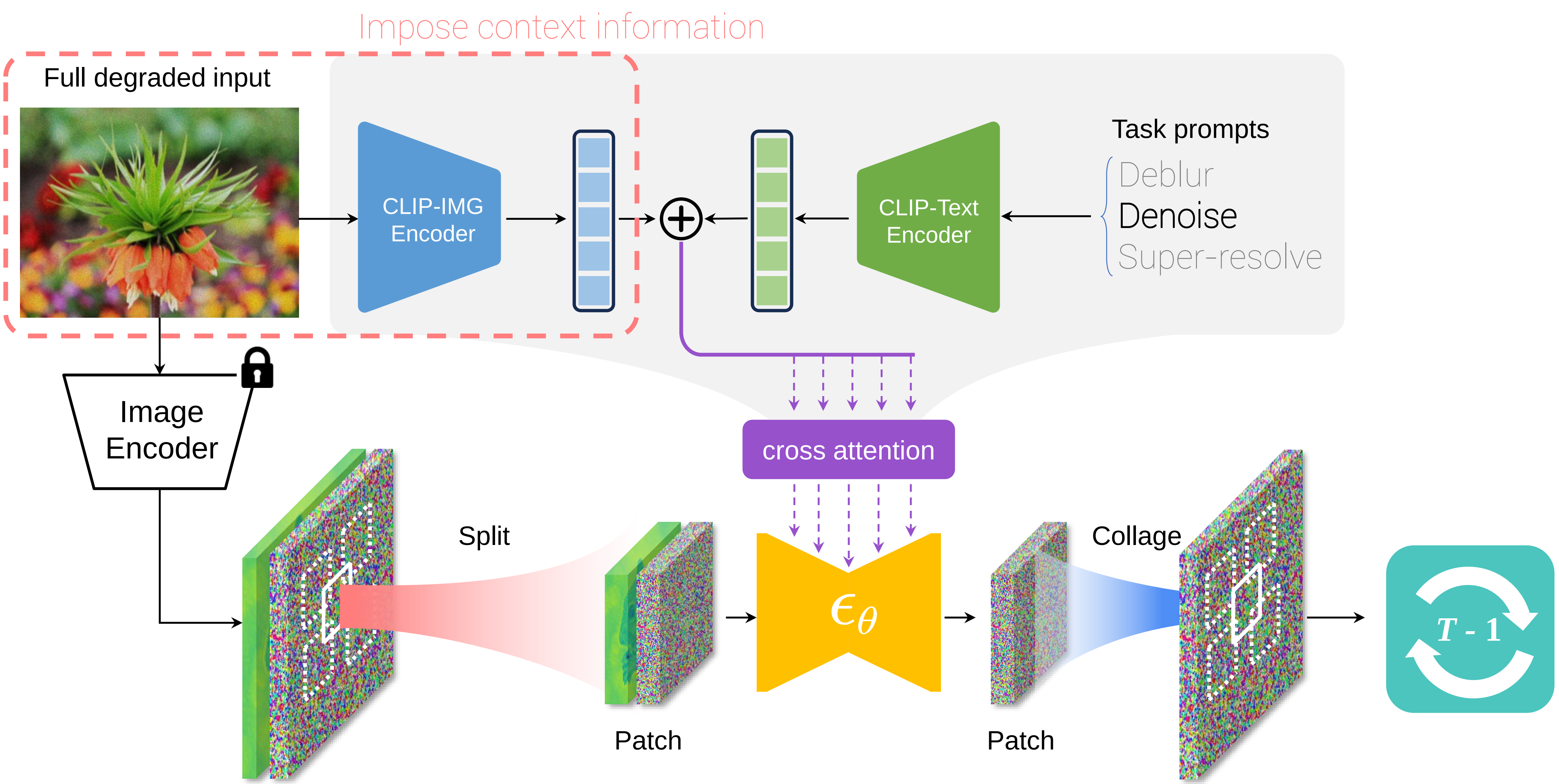
CodeFormer

CodeFormer (fidelity)

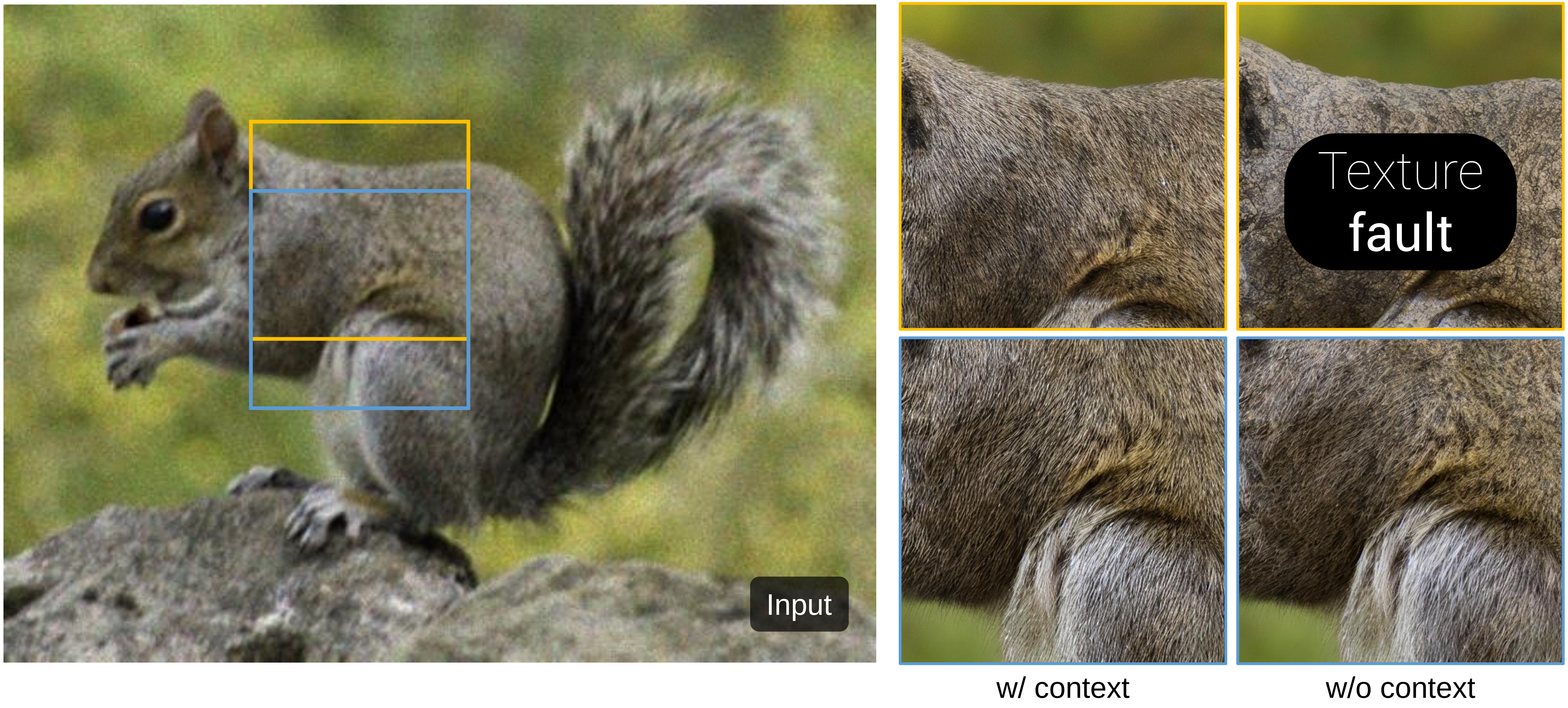
GT

Face-specific restoration

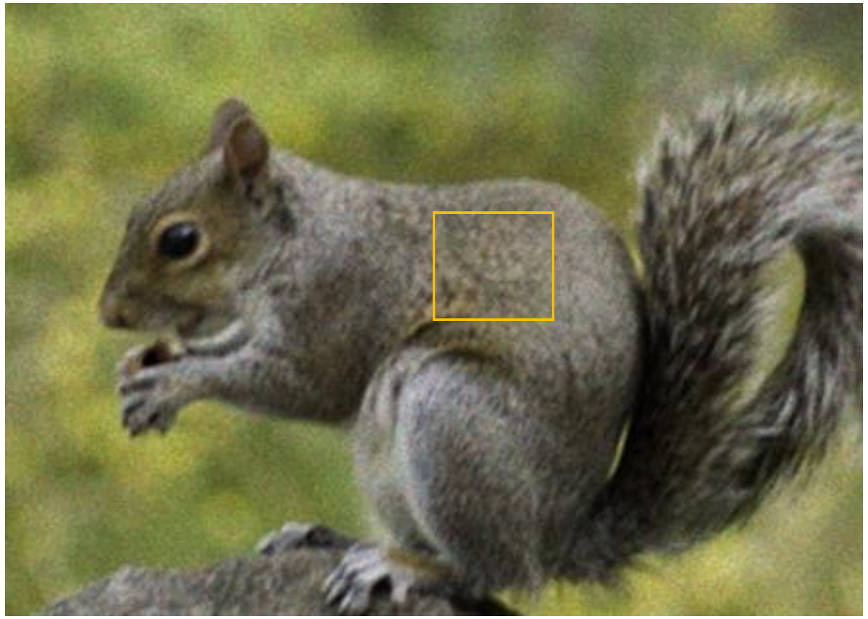
Discussion: why the design

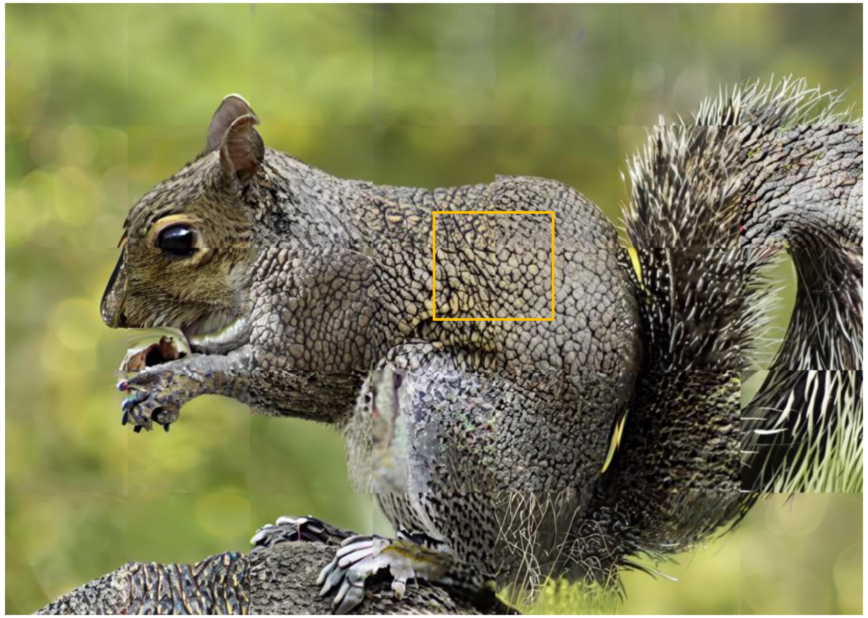


Discussion: why the design

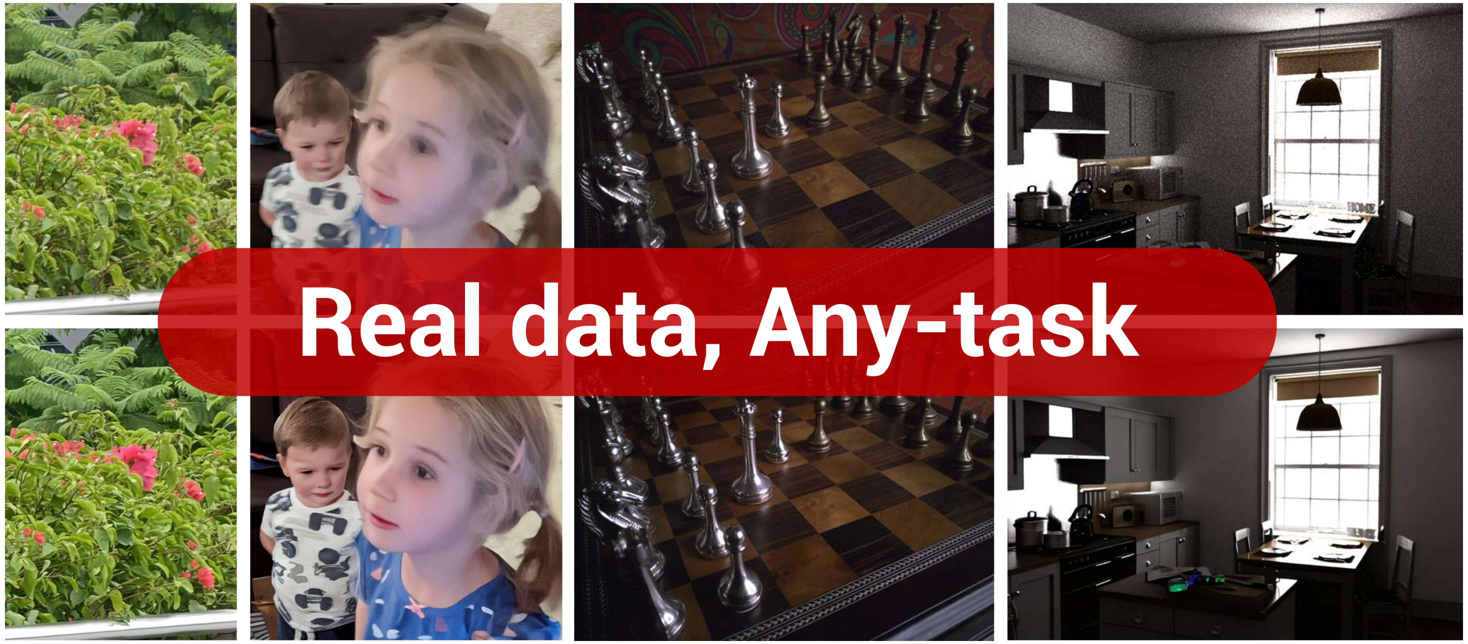


Discussion: why the design

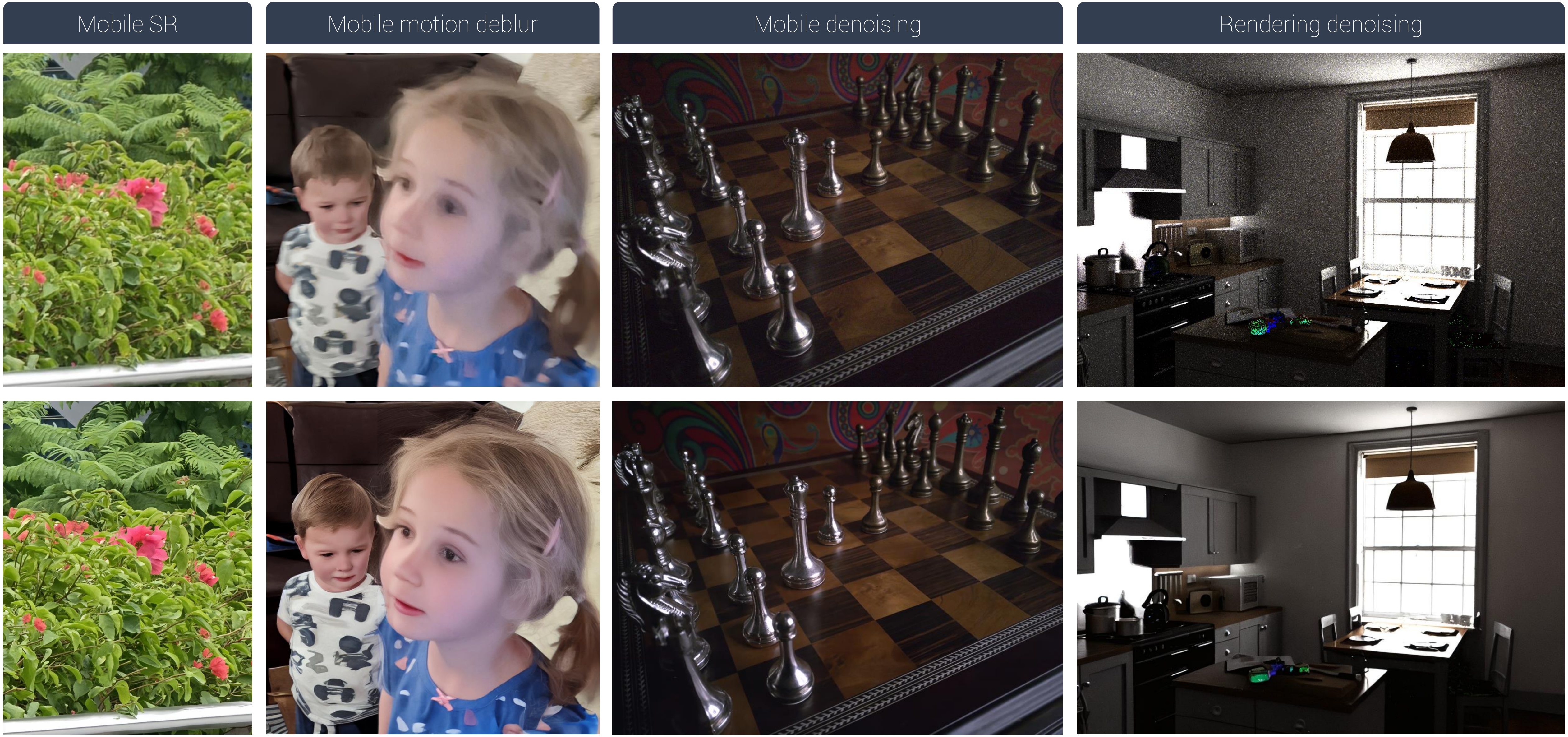




Outline



Part II: controllable restoration with text-guided diffusion


Mobile SR
Mobile motion deblur
Mobile denoising
Rendering denoising

Motivation: generalization

- Existing methods: poor generalizability

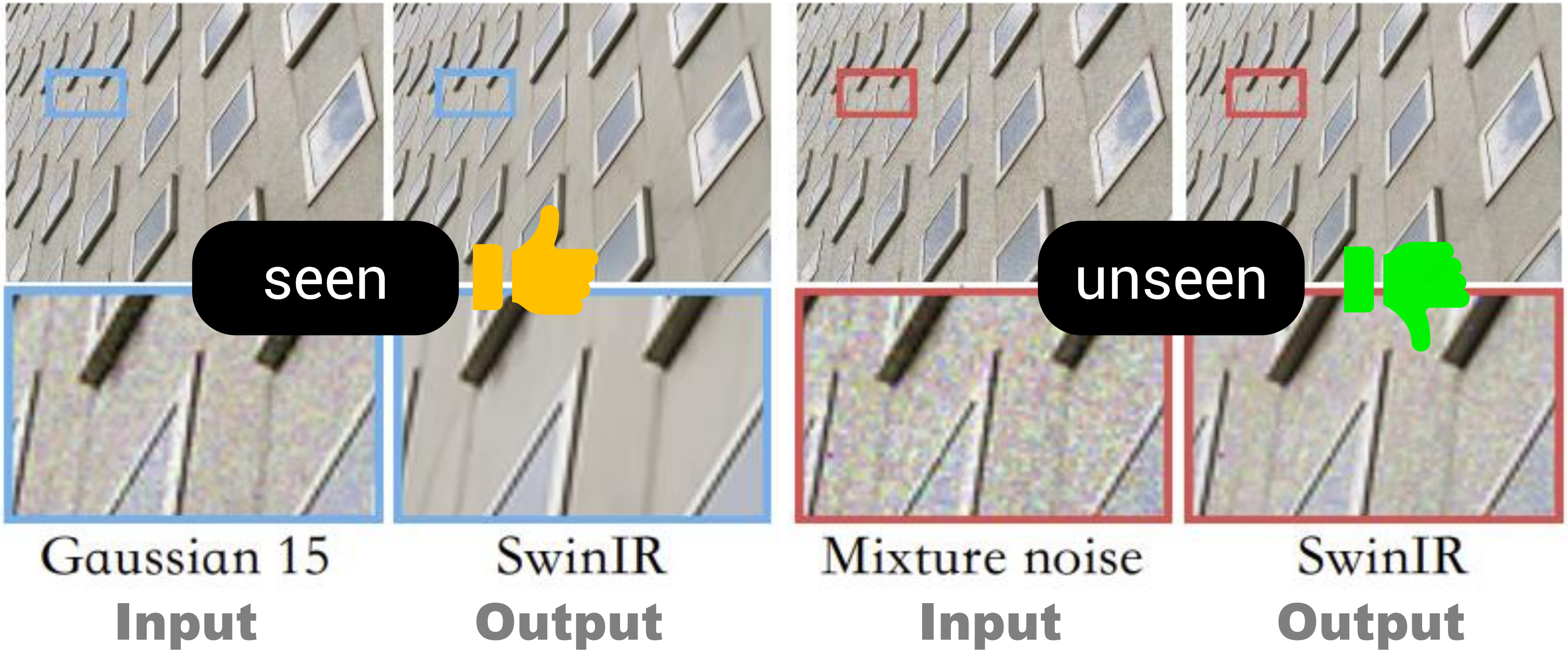
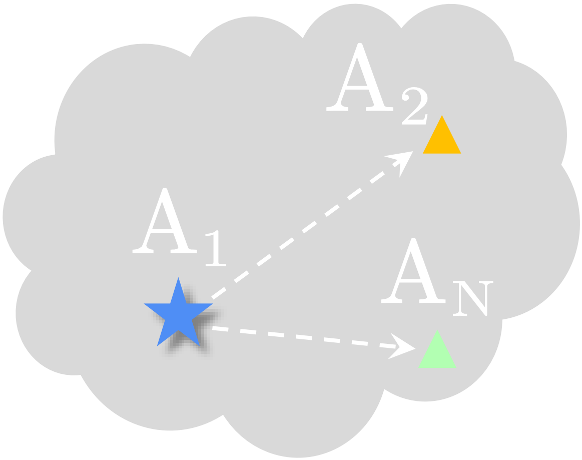


Image credit: Masked Image Training for Generalizable Deep Image Denoising, CVPR'23

Motivation: generalization

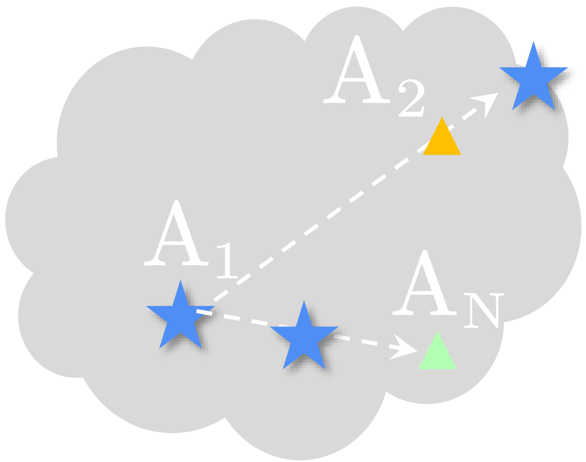
$$y = A \otimes x$$

observation		degradation		true signal



Testing on many unseen degradations

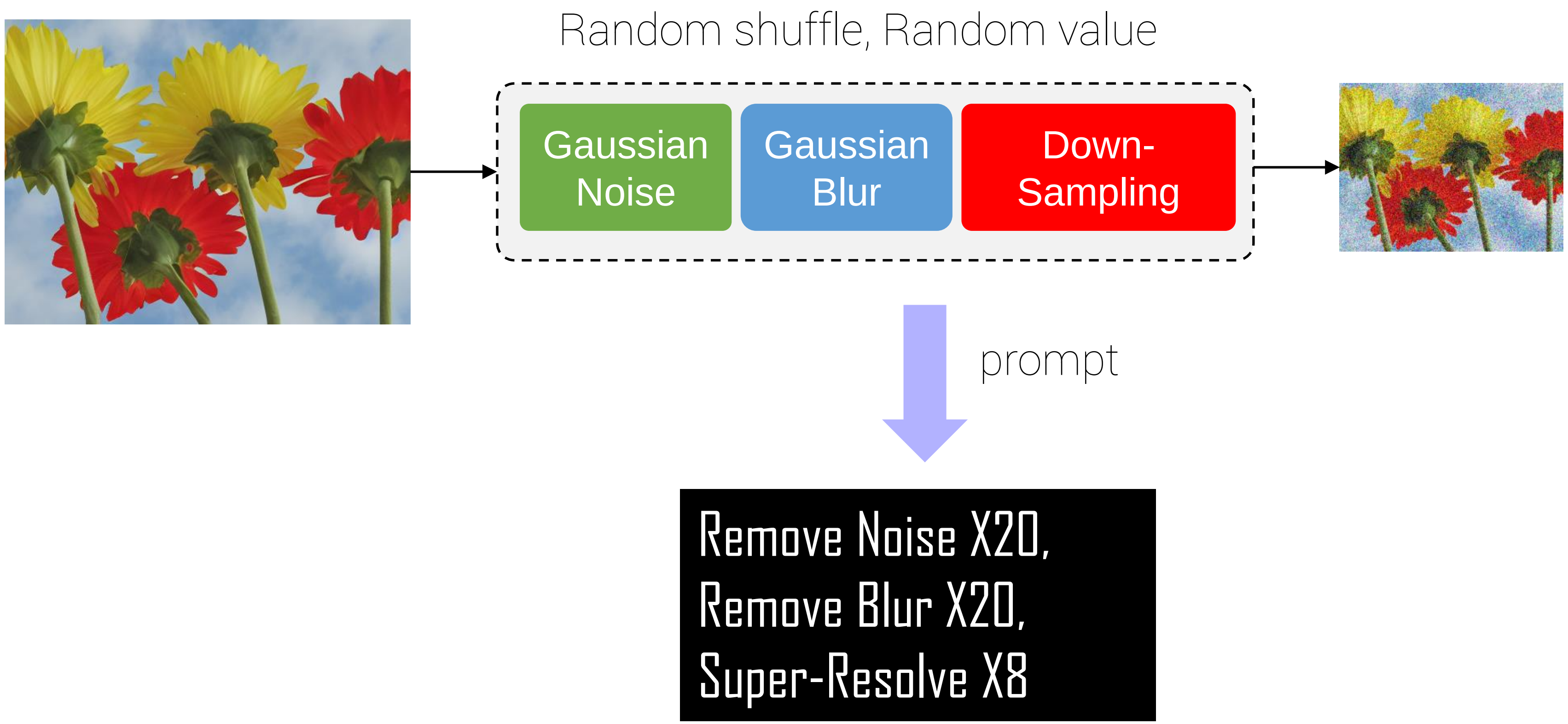
Naiive ideas



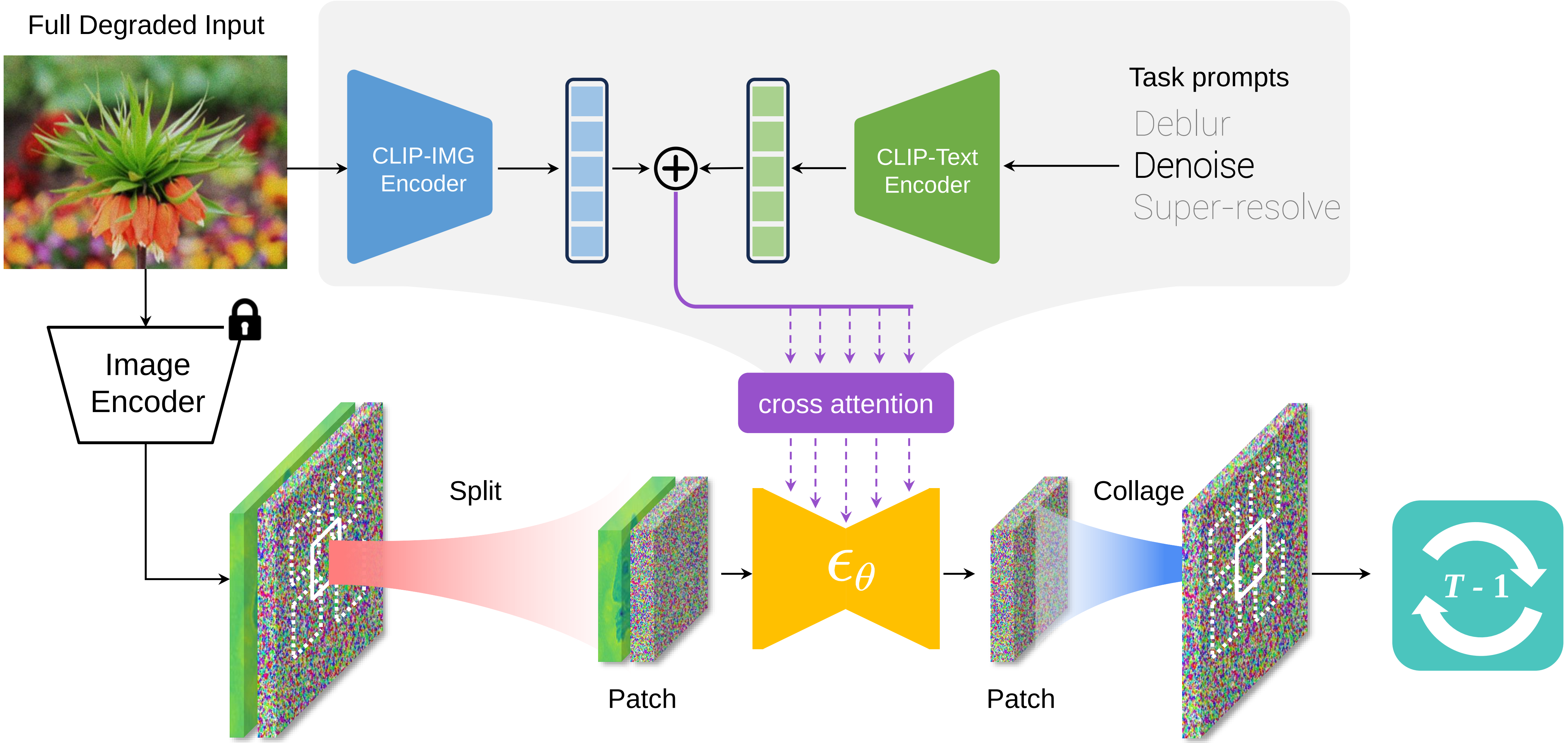
- Degradation augmentation/randomization
  - Like, Real-ESRGAN
- Controllable
- Degradation-invariant representation learning
  - Content prior
  - Masked image modeling

Method: degradation augmentation

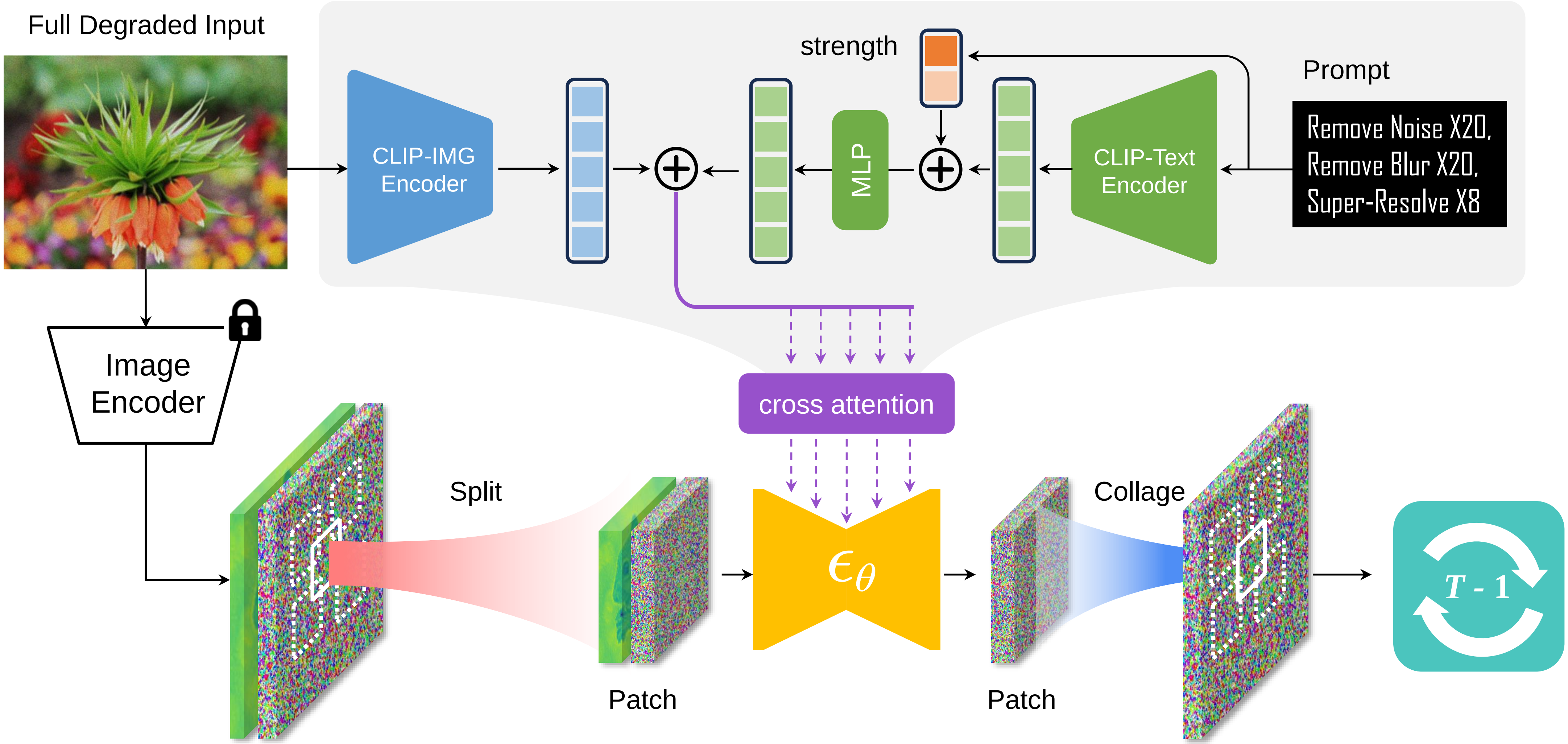
- Synthesize image and text prompt



Method: fine tune diffusion model



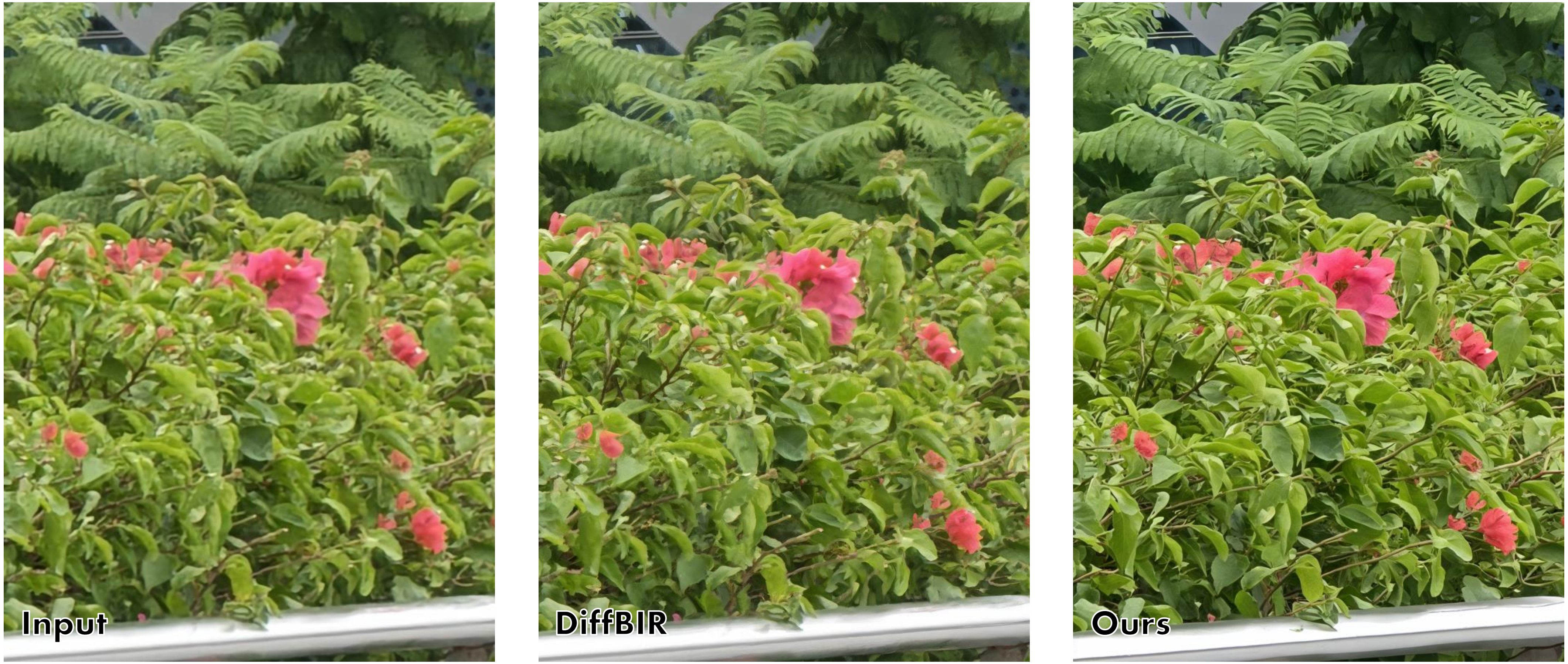
Method: degradation prompt



Results

- Model trained on **synthetic** data but testing on **real data**

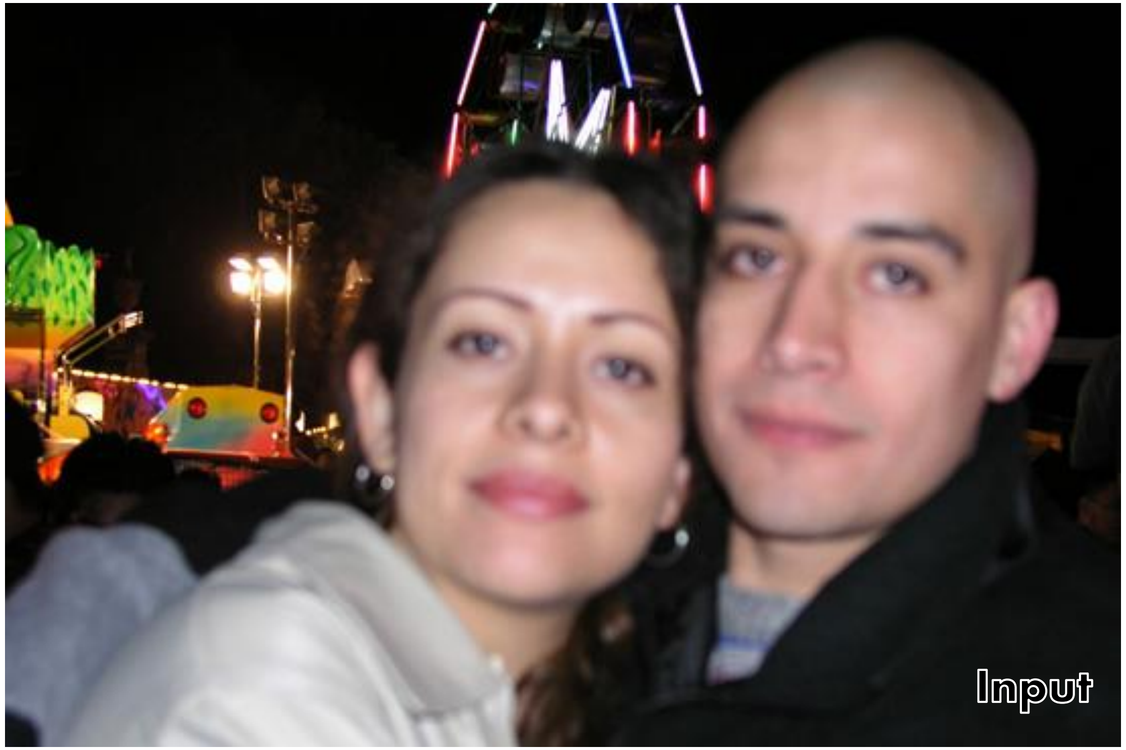
Results – mobile phone SR

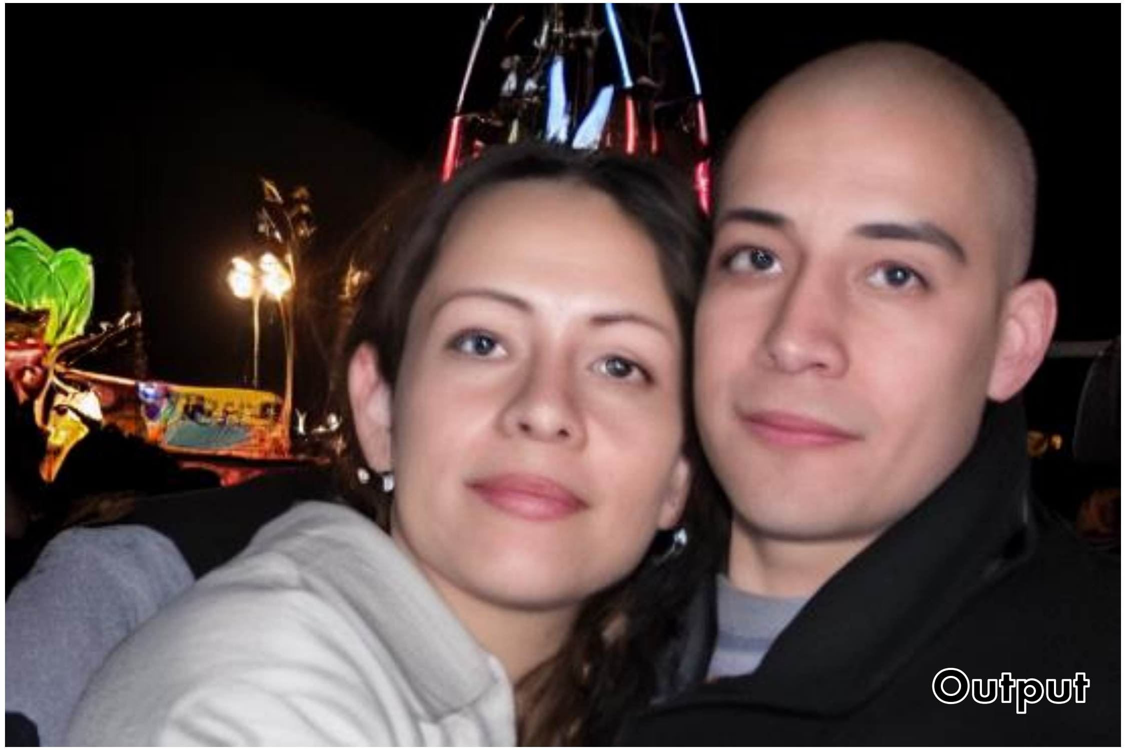


Results – mobile phone denoising (SIDD)

Results – rendering denoising

Results – out-of-focus deblurring





Results – motion deblurring

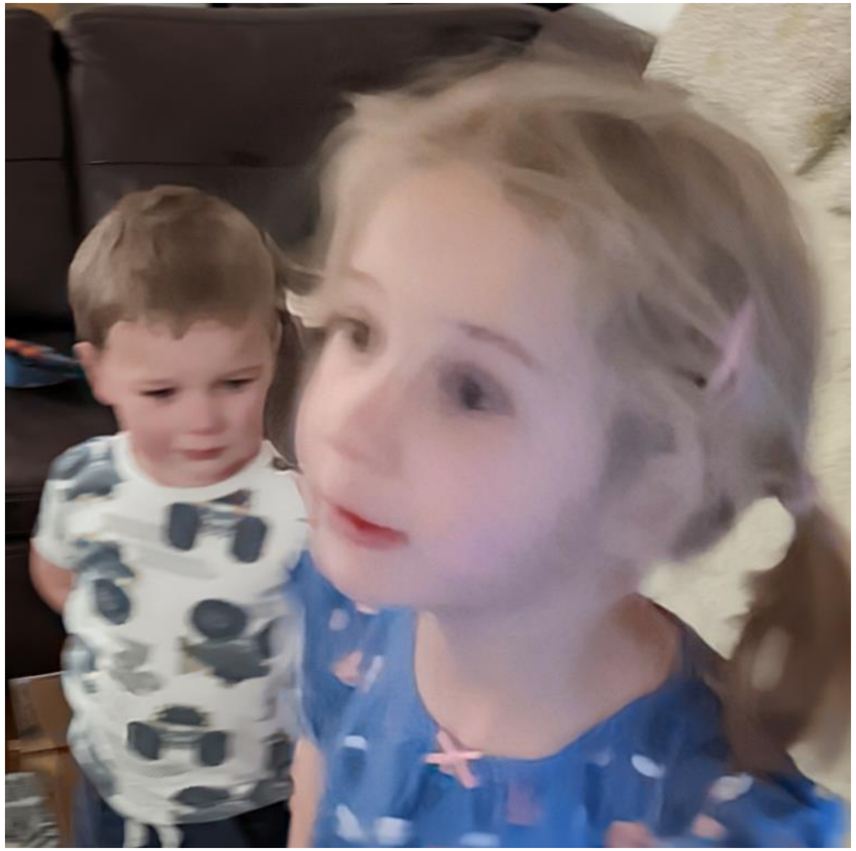
Input | DeblurGANv2 | MPRNet | Burst | Lai, SIGGRAPH'22

SwinIR | Codeformer | StableSR | Ours - SR | Ours - Deblur

Results – controllability

Input Deblur X20 Deblur X40

Results – controllability

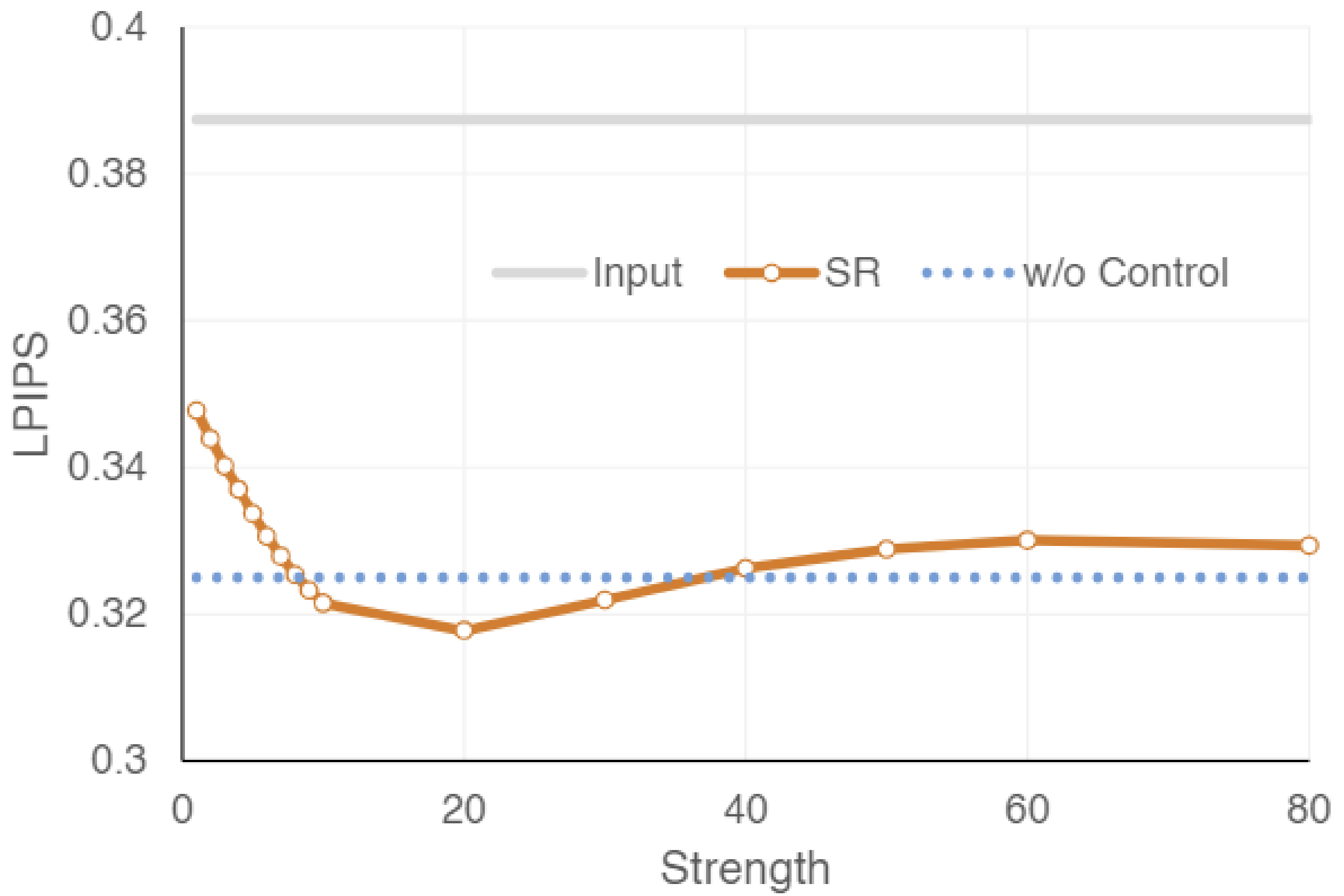


Input

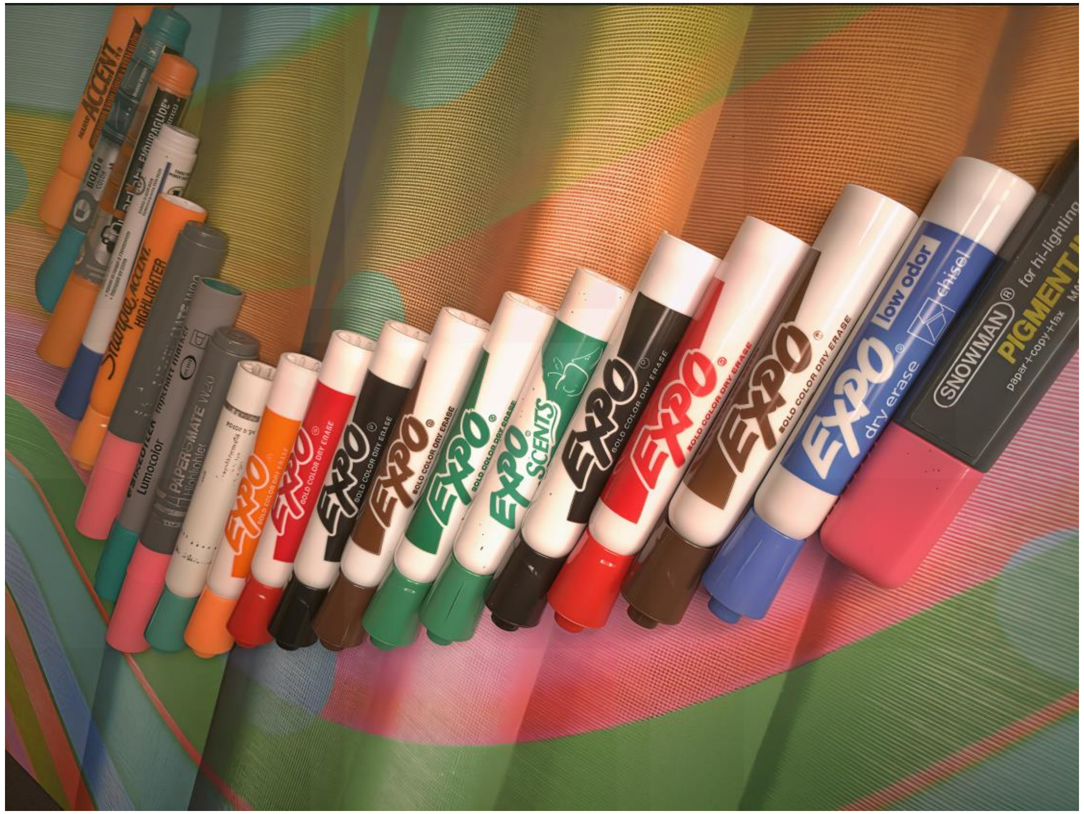
SR X3

SR X16

Results – controllability



Results – different input imaging conditions



StableSR

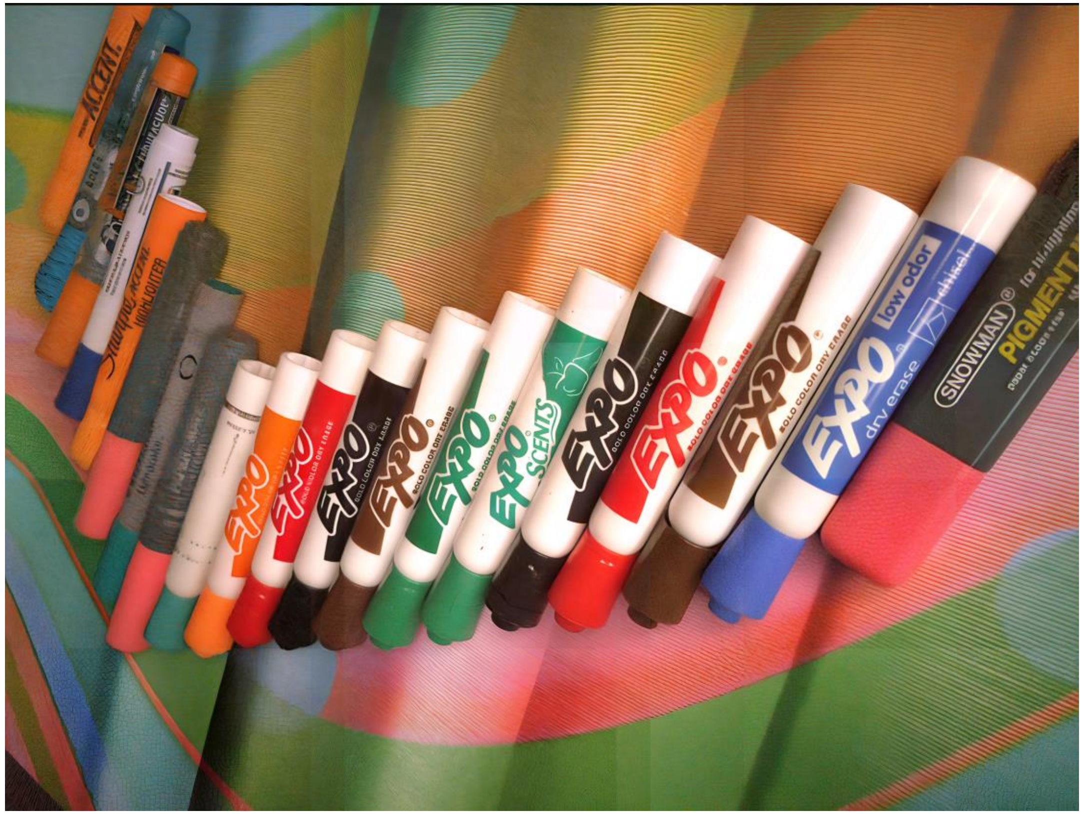
Ours

Results – different input imaging conditions

StableSR

Ours

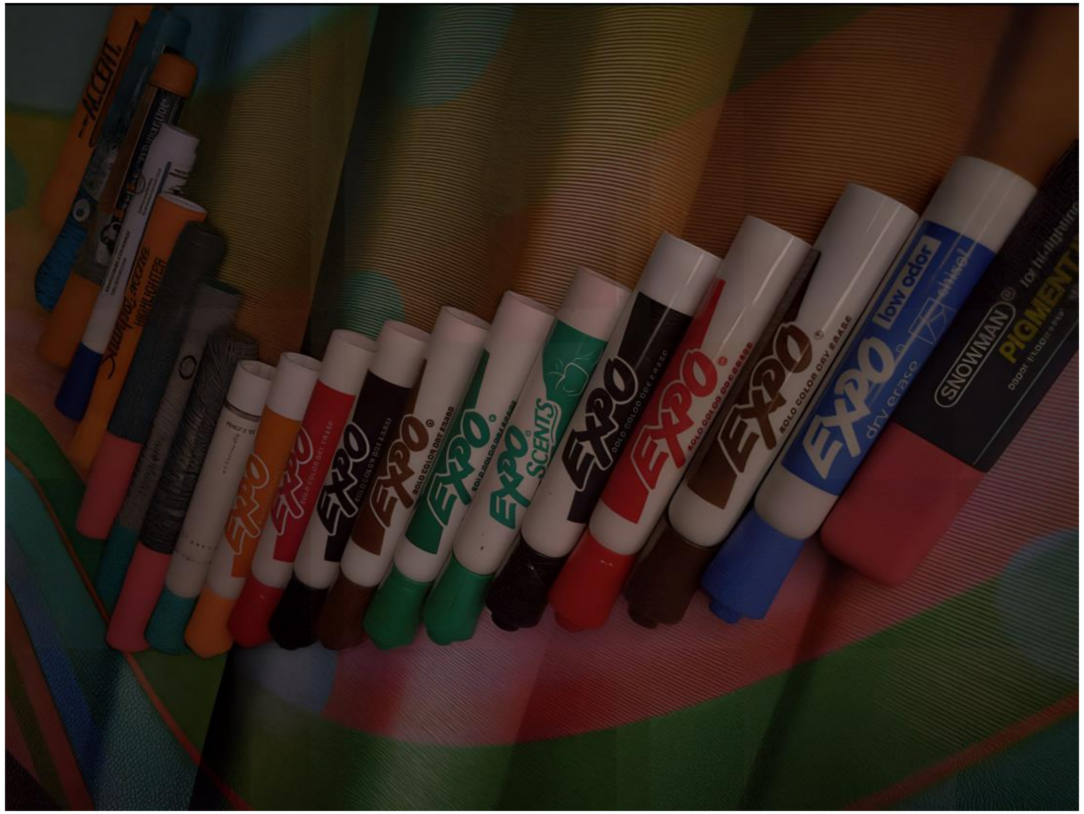
Results – different input imaging conditions



StableSR

Ours

Results – different input imaging conditions



StableSR

Ours

Results – different input imaging conditions

StableSR

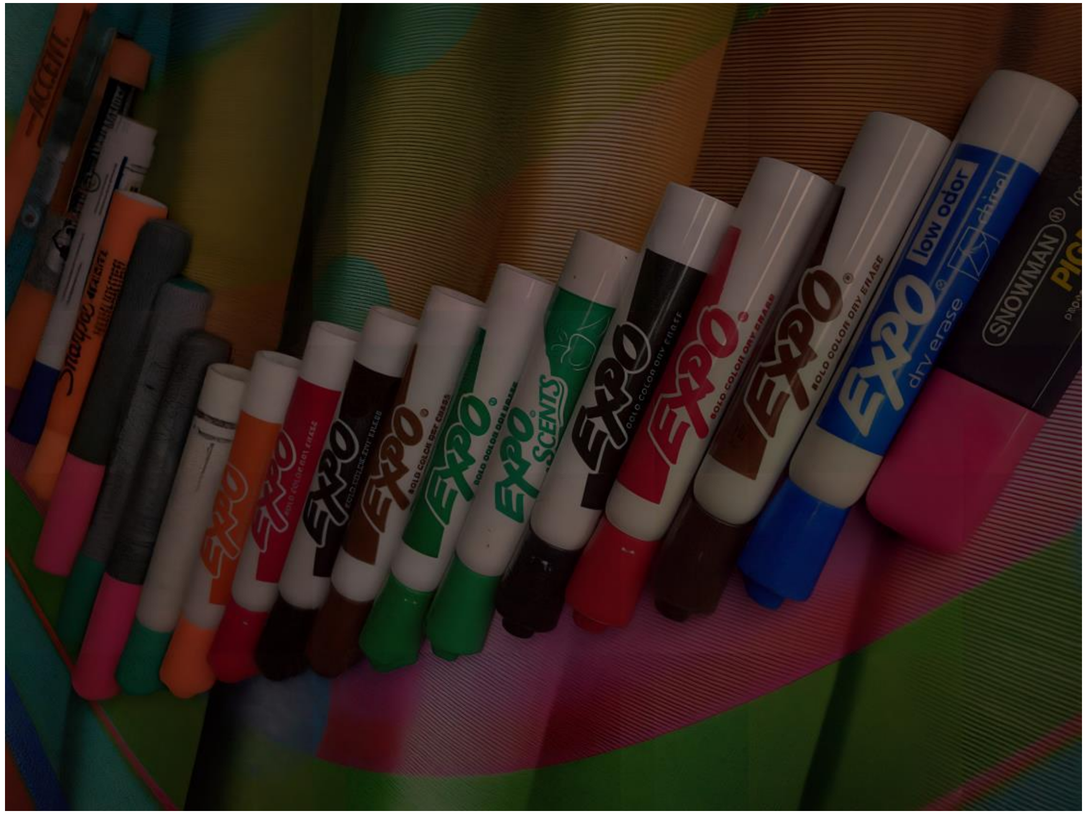
Ours

Results – different input imaging conditions

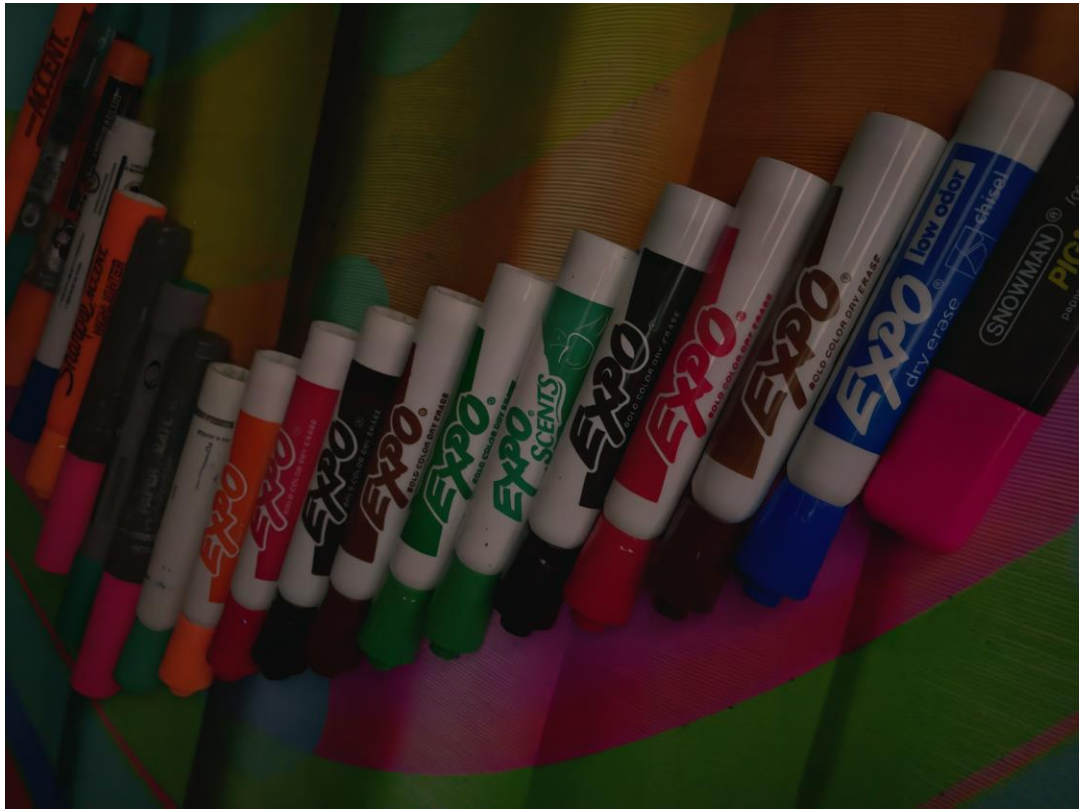
StableSR

Ours

Results – different input imaging conditions

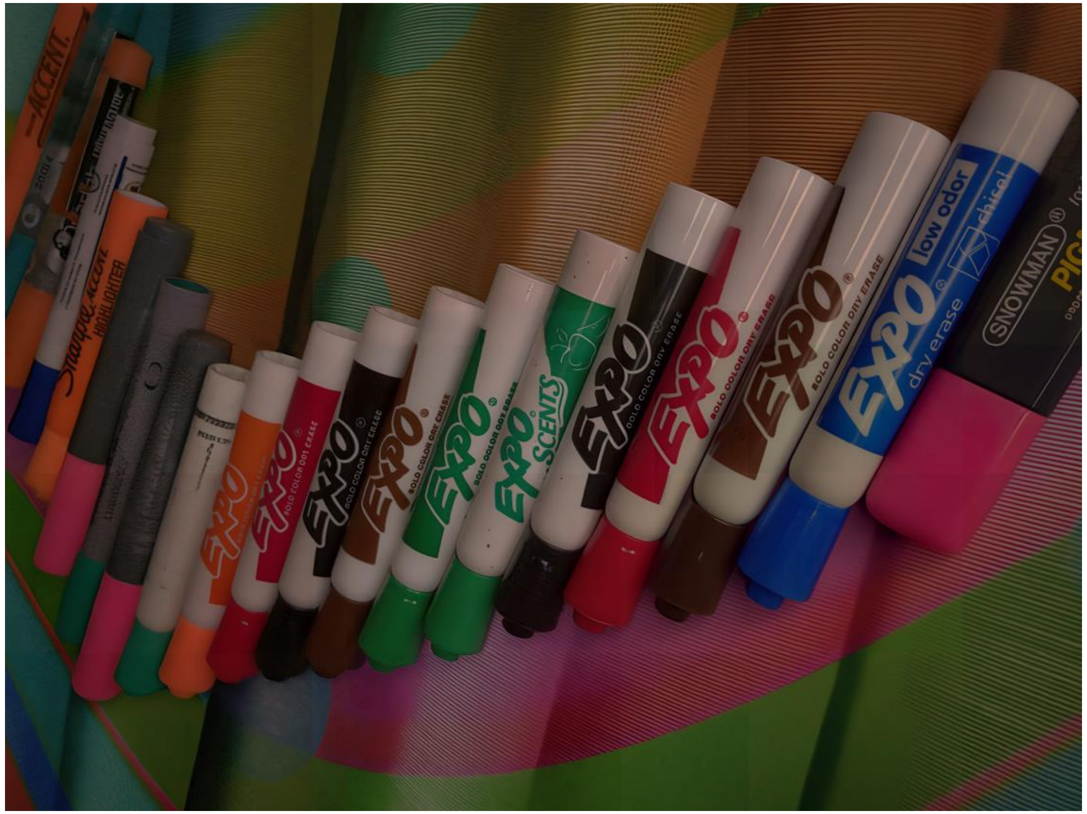


StableSR



Ours

Results – different input imaging conditions



StableSR

Ours

Results – different input imaging conditions

StableSR

Ours

Results – different input imaging conditions

StableSR

Ours

Results – different input imaging conditions

StableSR

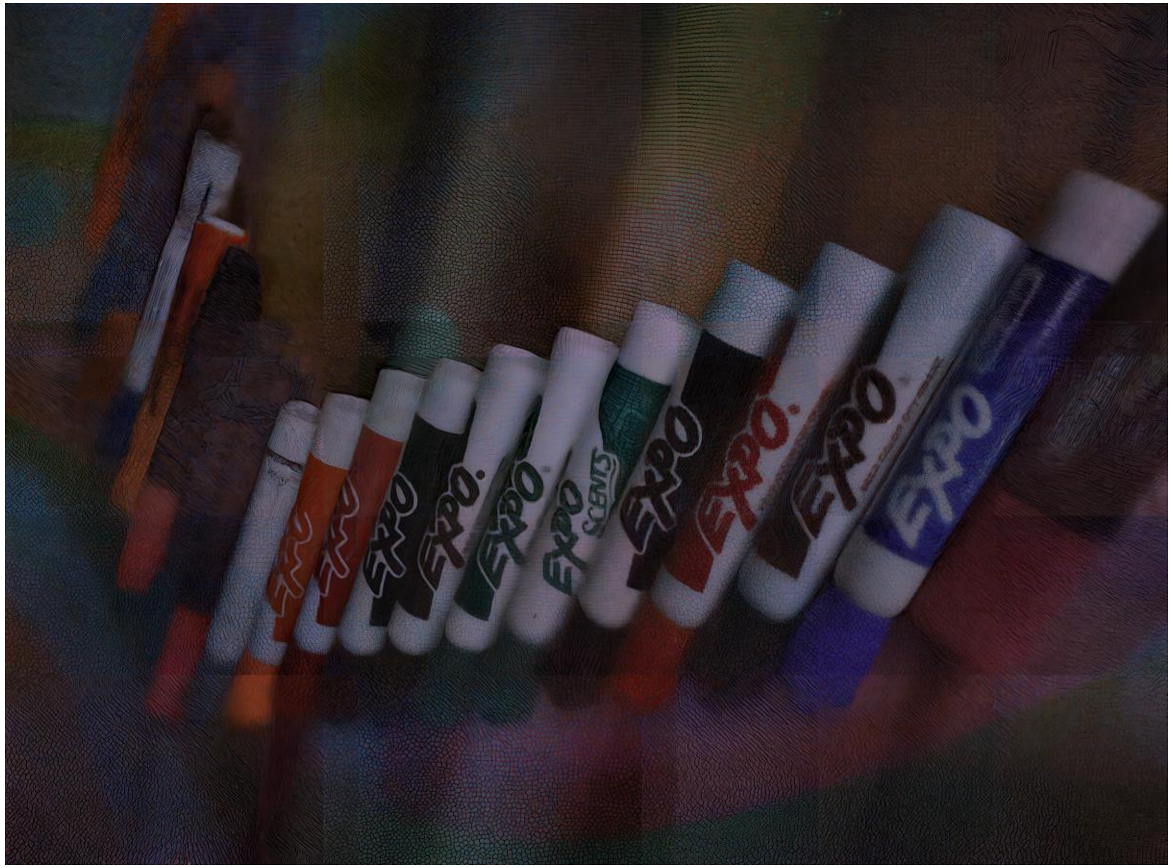
Ours

Results – different input imaging conditions

StableSR

Ours

Results – different input imaging conditions



StableSR

Ours

Take home messages

- Challenge but opportunities
  - ~~Inconsistency caused by patch processing~~
  - When the noise not totally removed, noise → inaccurate texture
    - Ongoing: increase the synthesis noise level

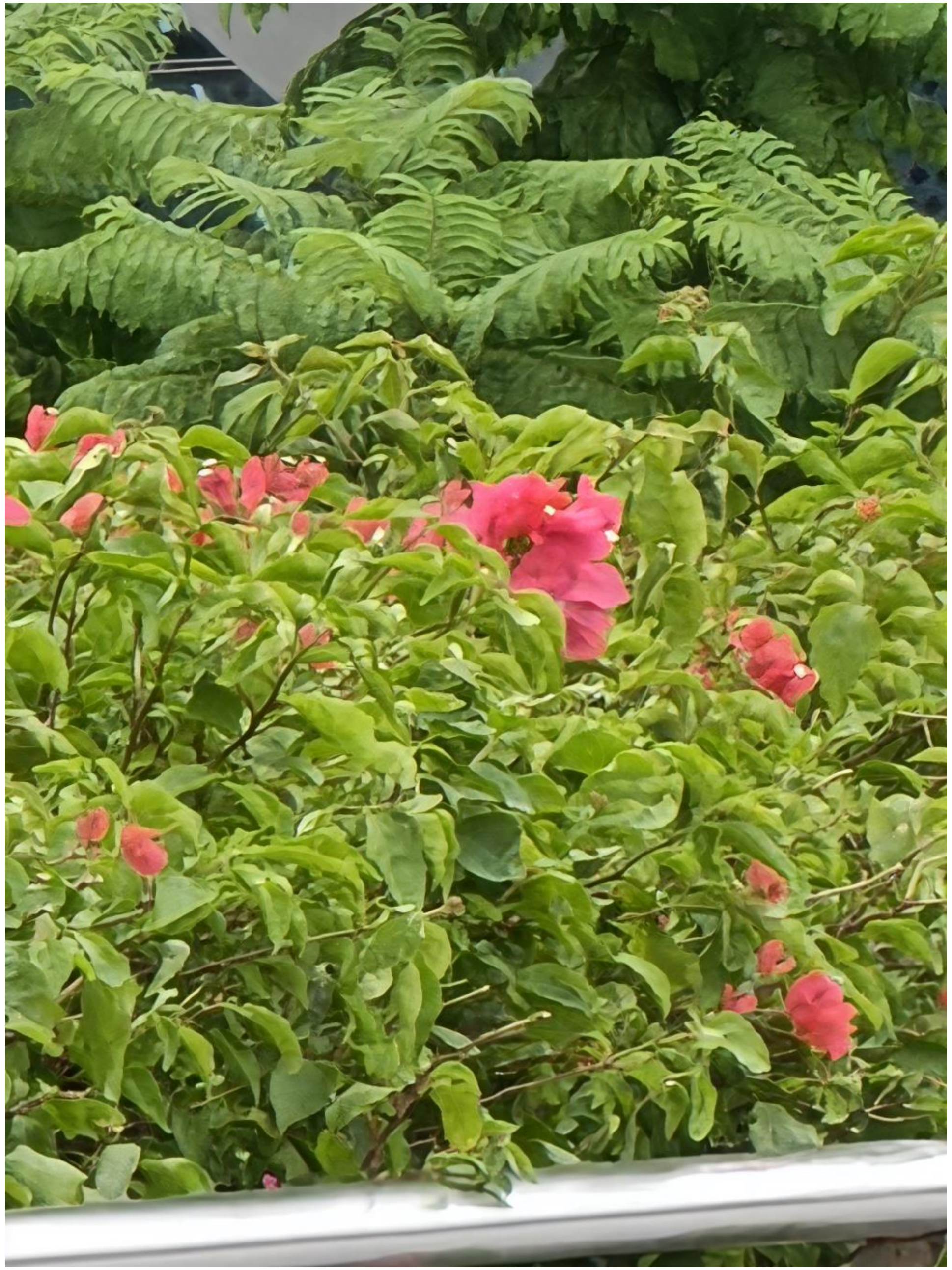

Input


Output

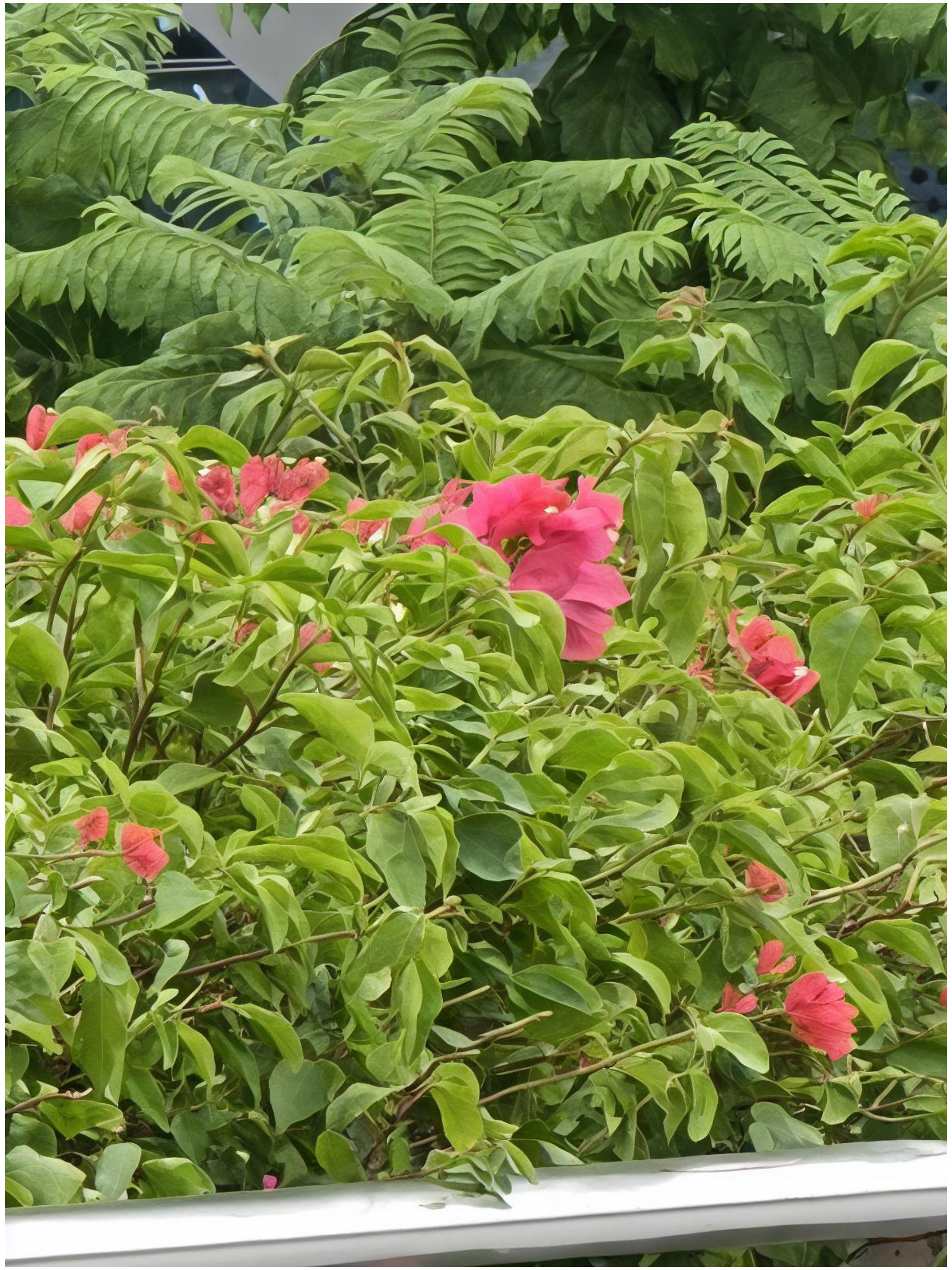
Take home messages

- Challenge but opportunities
  - ~~Inconsistency caused by patch processing~~
  - When the noise not totally removed, noise → inaccurate texture
    - Ongoing: increase the synthesis noise level
  - Frequency control

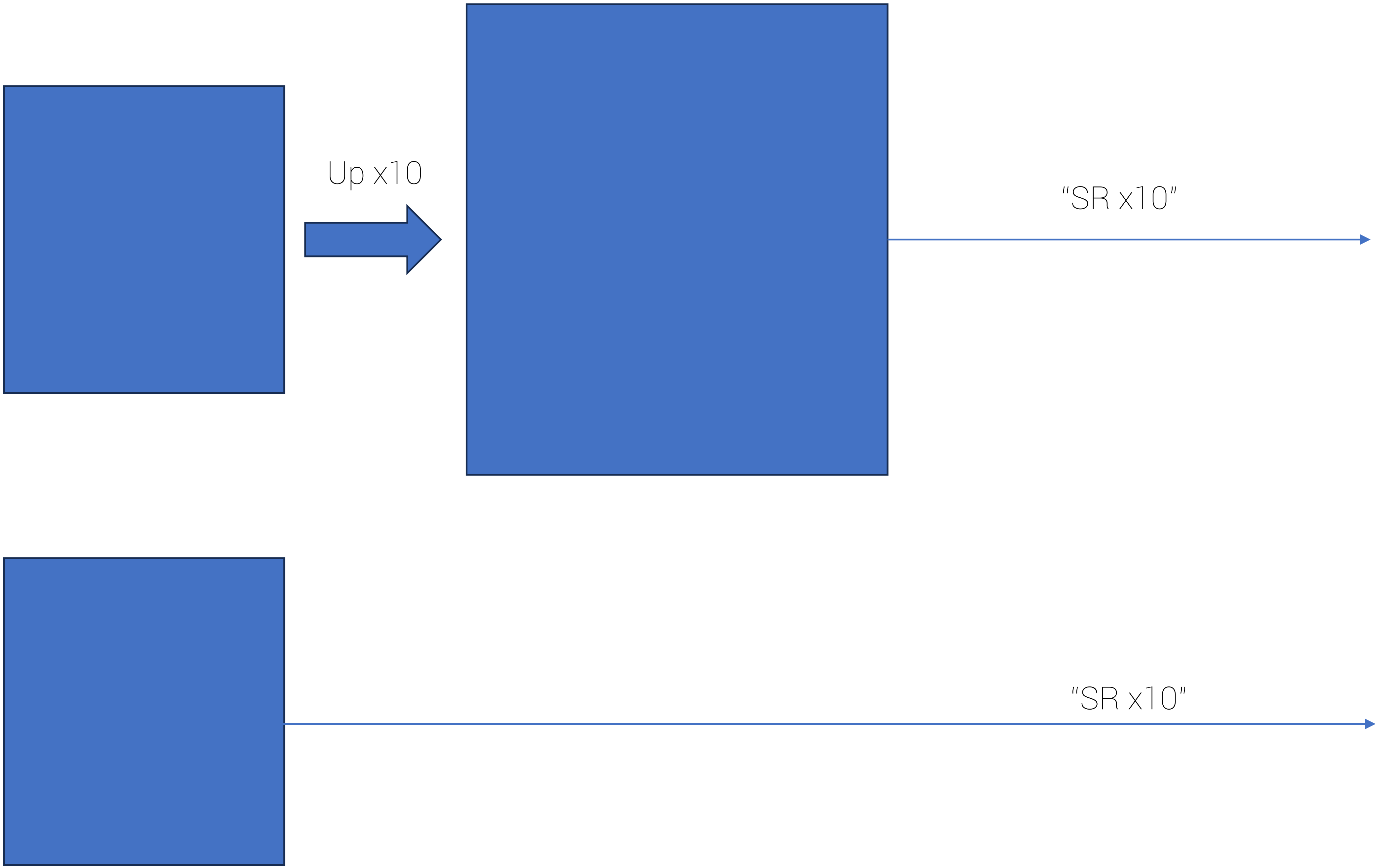
Input



sr60-up4-cfg11



sr60-up2-cfg11


Up x10
"SR x10"
"SR x10"

Take home messages

- Challenge but opportunities
  - ~~Inconsistency caused by patch processing~~
  - When the noise not totally removed, noise → inaccurate texture
    - Ongoing: increase the synthesis noise level
  - Frequency control
  - Uniform → Non-uniform restoration
    - Synthetic gaussian is hard for real deblur
    - Need different manipulation level, like dehaze
      - Regional controllable
      - Guidance
      - Layered
  - Processing time of patch-based method
    - TODO:
  - Structural content, like text
  - Continuous representation