SPIRE: Semantic Prompt-Driven Image Restoration

Chenyang Qi[1,2*], Zhengzhong Tu[1], Keren Ye[1] Mauricio Delbracio[1], Peyman Milanfar[1], Qifeng Chen[2], and Hossein Talebi[1]

[1] Google Research
[2] HKUST

ECCV 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

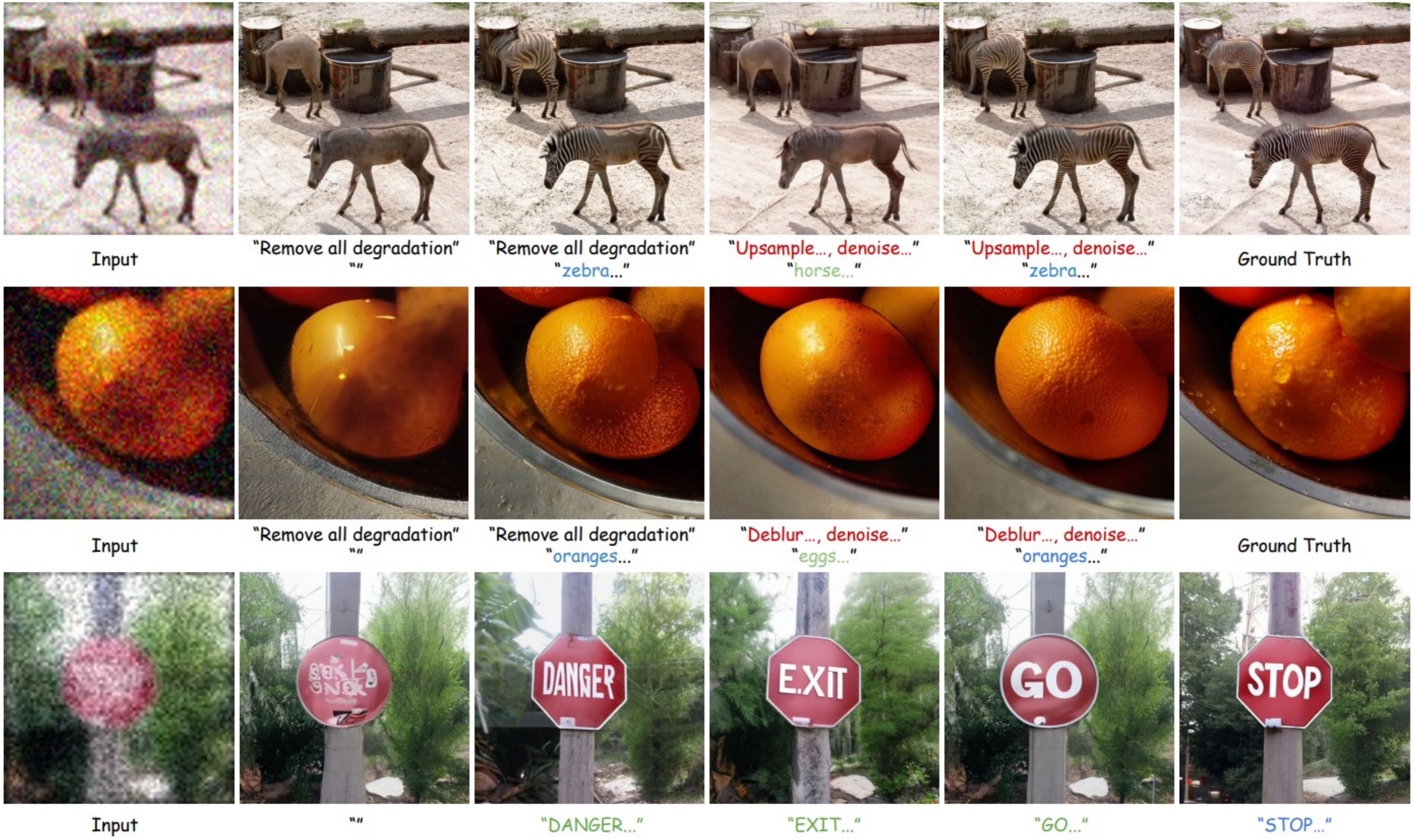
Introduction

- Introduce the **first** unified text-driven image restoration model that supports both **semantic prompts** and **restoration instructions**.

- Proposed paradigm empowers users to **fully control** the semantic outcome of the restored image using different semantic prompts during test time, providing a mechanism for users to **adjust the category and strength of the restoration effect**.

- Our experiments demonstrate that incorporating semantic prompts and restoration instructions **significantly enhances** the restoration quality, **eliminating the need for task-specific model design**.

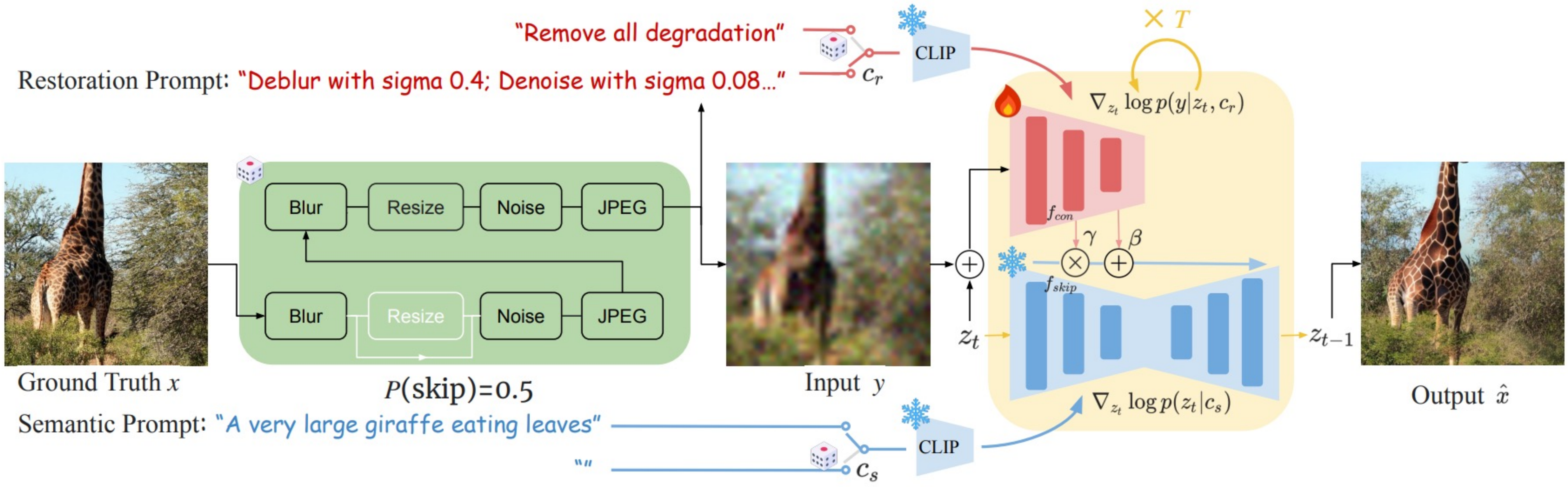
Introduction



Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Framework



"Remove all degradation"
Restoration Prompt: "Deblur with sigma 0.4; Denoise with sigma 0.08..."
c_r
CLIP
$\times T$
$\nabla_{z_t} \log p(y|z_t, c_r)$
Blur
Resize
Noise
JPEG
Blur
Resize
Noise
JPEG
f_{con}
γ
β
f_{skip}
z_t
z_{t-1}
Ground Truth x
P(skip)=0.5
Input y
Output $\hat{x}$
$\nabla_{z_t} \log p(z_t|c_s)$
Semantic Prompt: "A very large giraffe eating leaves"
""
CLIP
c_s

Outline

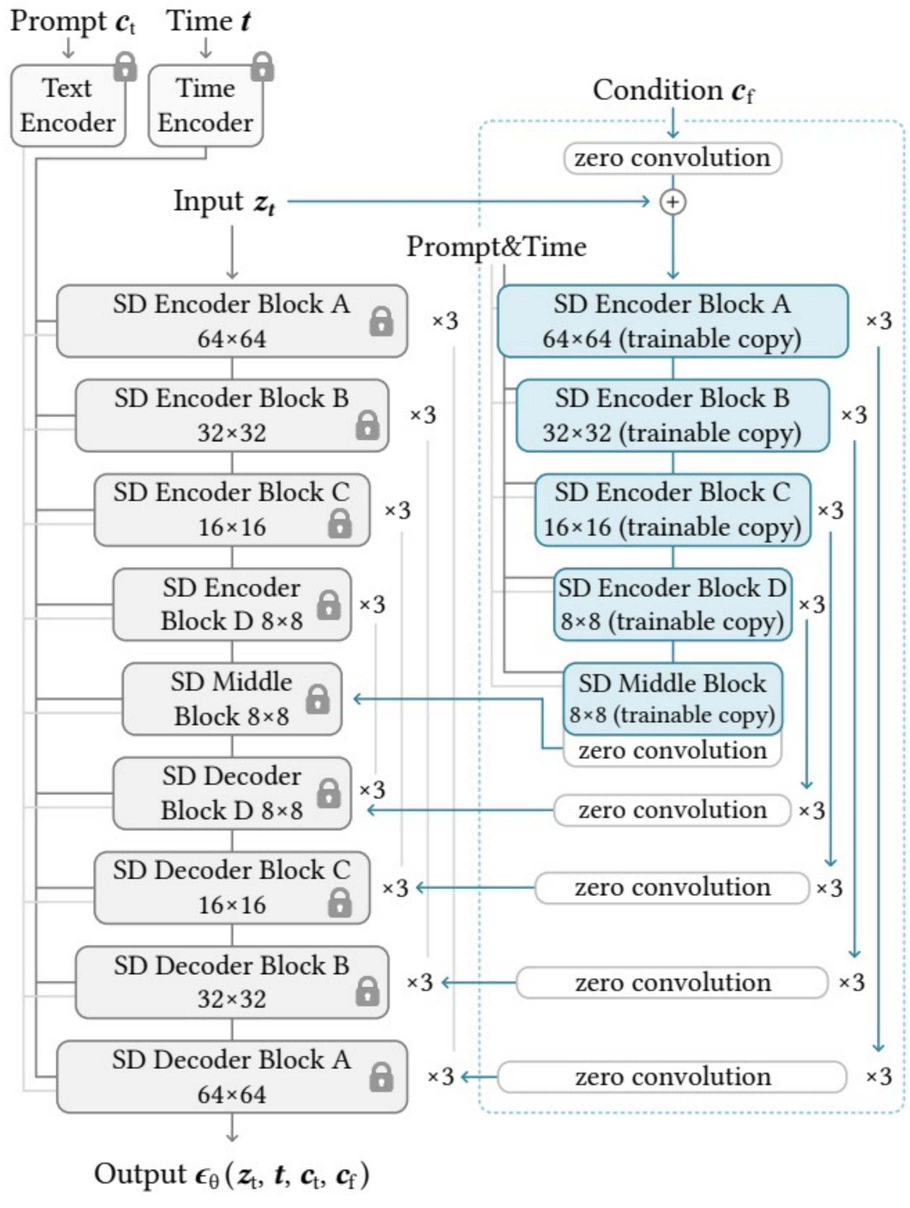
- Introduction
- Framework
- Method
- Experiment
- Conclusion

Decoupling Semantic and Restoration Prompts

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{z}_0, \boldsymbol{y}) \sim p_{\text{data}}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \left(\boldsymbol{z}_t, t, \boldsymbol{y} \right) \right\|_2^2, \tag{1}$$

$$\nabla_{z_t} \log p(\boldsymbol{z}_t | \boldsymbol{y}, \boldsymbol{c}_s, \boldsymbol{c}_r) \approx \underbrace{\nabla_{z_t} \log p(\boldsymbol{z}_t | \boldsymbol{c}_s)}_{\text{Semantic-aware (frozen)}} + \underbrace{\nabla_{z_t} \log p(\boldsymbol{y} | \boldsymbol{z}_t, \boldsymbol{c}_r)}_{\text{Restoration-aware (learnable)}}. \tag{2}$$

$$\hat{\boldsymbol{f}}_{\text{skip}} = (1 + \boldsymbol{\gamma}) \boldsymbol{f}_{\text{skip}} + \boldsymbol{\beta}; \;\; \boldsymbol{\gamma}, \boldsymbol{\beta} = \mathcal{M}(\boldsymbol{f}_{\text{con}})$$



(a) Stable Diffusion (b) ControlNet

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Training degradation

Parameterized			Real-ESRGAN		
Degradation Process	p(choose)	Restoration Prompt	Degradation Process	p(choose)	Restoration Prompt
Gaussian Blur	0.5	Deblur with {sigma$\in [0.2, 3.0]$} or Deblur	Blur	1.0	Remove all degradation
Downsample	0.5	Upsample to {resizing factor$\in [1.0, 7.0]$} or Upsample	Resize	1.0	
Gaussian Noise	0.5	Denoise with {sigma$\in [0.0, 0.12]$} or Denoise	Noise	1.0	
JPEG	0.5	Dejpeg with quality {quality factor$\in [30, 92]$} or Dejpeg	JPEG	1.0	

Degradation ambiguities

Input "Remove all degradation" "Deblur with sigma 3.0" Ground Truth

Results

Method	Prompts		Parameterized Degradation with synthesized $\boldsymbol{c}_r$						Real-ESRGAN Degradation without $\boldsymbol{c}_r$					
	Sem	Res	FID↓	LPIPS↓	PSNR↑	SSIM↑	CLIP-I↑	CLIP-T↑	FID↓	LPIPS↓	PSNR↑	SSIM↑	CLIP-I↑	CLIP-T↑
SwinIR [30]	✗	✗	43.22	0.423	**24.40**	**0.717**	0.856	0.285	48.37	0.449	**23.45**	**0.699**	0.842	0.284
StableSR [67]	✗	✗	20.55	0.313	21.03	0.613	0.886	0.298	25.75	0.364	20.42	0.581	0.864	0.298
DiffBIR [34]	✗	✗	17.26	0.302	22.16	0.604	0.912	0.297	19.17	0.330	21.48	0.587	0.898	0.298
ControlNet-SR [84]	✗	✗	13.65	0.222	23.75	0.669	0.938	0.300	16.99	0.269	22.95	0.628	0.924	0.299
Ours w/o text	✗	✗	12.70	0.221	23.84	0.671	0.939	0.299	16.25	**0.262**	23.15	0.636	0.929	0.300
DiffBIR [34] + SDEdit [37]	✓	✗	19.36	0.362	19.39	0.527	0.891	0.305	17.51	0.375	19.15	0.521	0.887	**0.308**
DiffBIR [34] + CLIP [46]	✓	✗	18.46	0.365	20.50	0.526	0.896	**0.308**	20.31	0.374	20.45	0.539	0.885	0.307
ControlNet-SR + CLIP [46]	✓	✗	13.00	0.241	23.18	0.648	0.937	0.307	15.16	0.286	22.45	0.610	0.926	**0.308**
Ours	✓	✓	**11.34**	**0.219**	23.61	0.665	**0.943**	0.306	**14.42**	**0.262**	23.14	0.633	**0.935**	**0.308**

Results

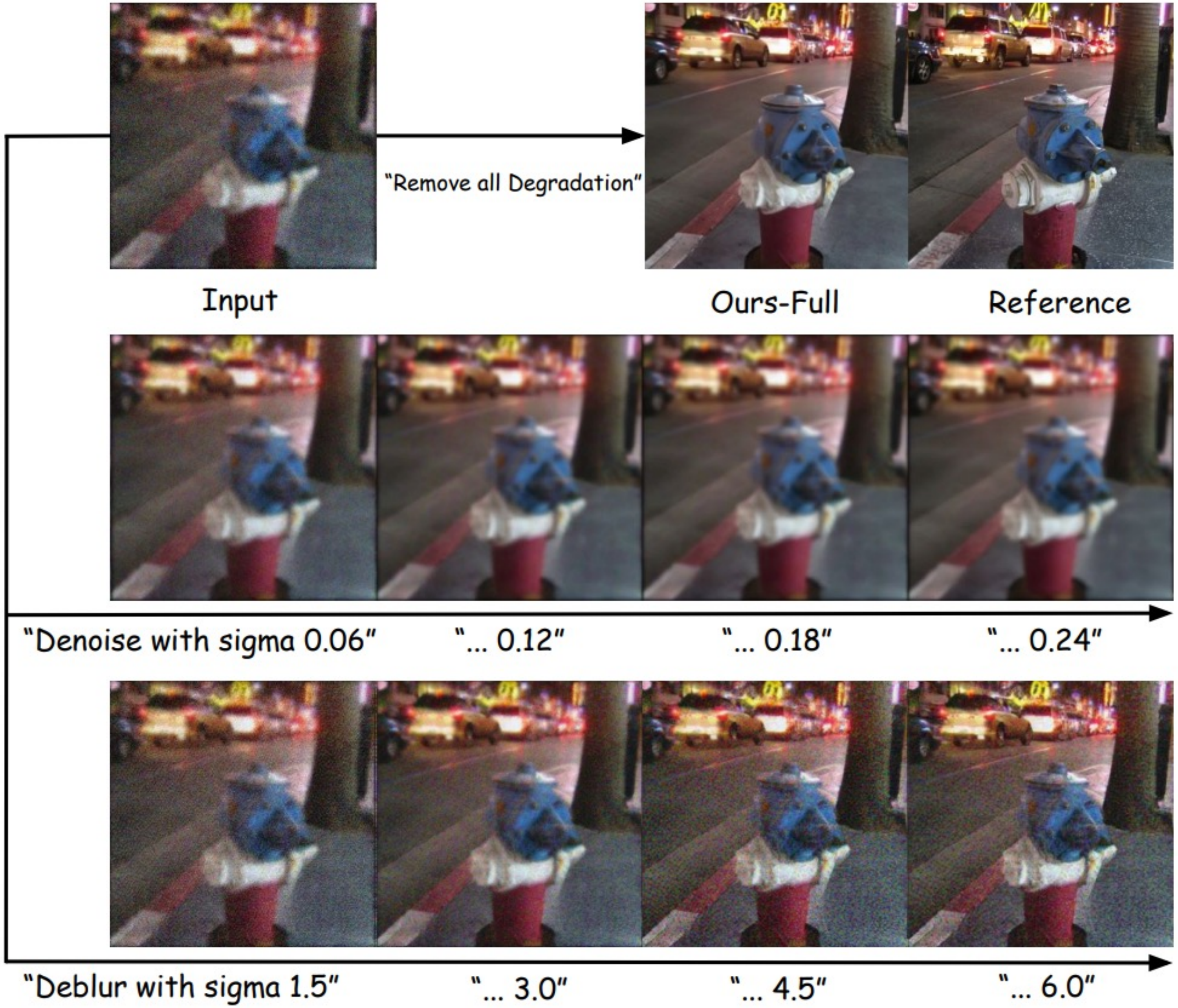
Input
DiffBIR [34]
DiffBIR + SDEdit [37]
DiffBIR + CLIP [46]
ControlNet-SR [84]
Our SPIRE Model
Ground-Truth
c_s = "A blue, white and red fire hydrant sitting on a sidewalk."
c_s = "Four young giraffes in a zoo, with one of them being fed leaves by a person."
c_s = "Two hands holding and dialing a cellular phone."
c_s = "An old boat sitting in the middle of a field."

Results

Method	FID↓	LPIPS↓	PSNR↑	SSIM↑	CLIP-I↑
Real-ESRGAN [72]	32.37	0.312	**22.52**	**0.646**	0.683
DiffBIR [34] (zero-shot)	30.71	0.354	22.01	0.526	0.921
StableSR	24.44	**0.311**	21.62	0.533	0.928
Ours w/o text (zero-shot)	28.80	0.352	21.68	0.549	0.927
Ours w/o text (finetuned)	**22.45**	0.321	21.38	0.532	**0.932**

Table 3: Numerical results on the DIV2K testset without any prompt.

Results



"Remove all Degradation"
Input
Ours-Full
Reference
"Denoise with sigma 0.06"
"... 0.12"
"... 0.18"
"... 0.24"
"Deblur with sigma 1.5"
"... 3.0"
"... 4.5"
"... 6.0"

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- First framework to support both semantic and parameter-embedded restoration instructions simultaneously.

- Decoupled way to better preserve the **semantic text-to-image generative prior** while efficiently learning to **control both the restoration direction and its strength**.

- Extensive experiments have shown that this method significantly outperforms prior works in terms of both quantitative and qualitative results.