Improving Image Restoration through Removing Degradations in Textual Representations

Jingbo Lin[1], Zhilu Zhang[1], Yuxiang Wei[1], Dongwei Ren[1], Dongsheng Jiang[2], Wangmeng Zuo[1,*]
[1]Harbin Institute of Technology [2]Huawei Cloud Computing Co., Ltd.
jblincs1996@gmail.com, cszlzhang@outlook.com, yuxiang.wei.cs@gmail.com,
rendongweihit@gmail.com, dongsheng_jiang@outlook.com, cswmzuo@gmail.com

CVPR 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

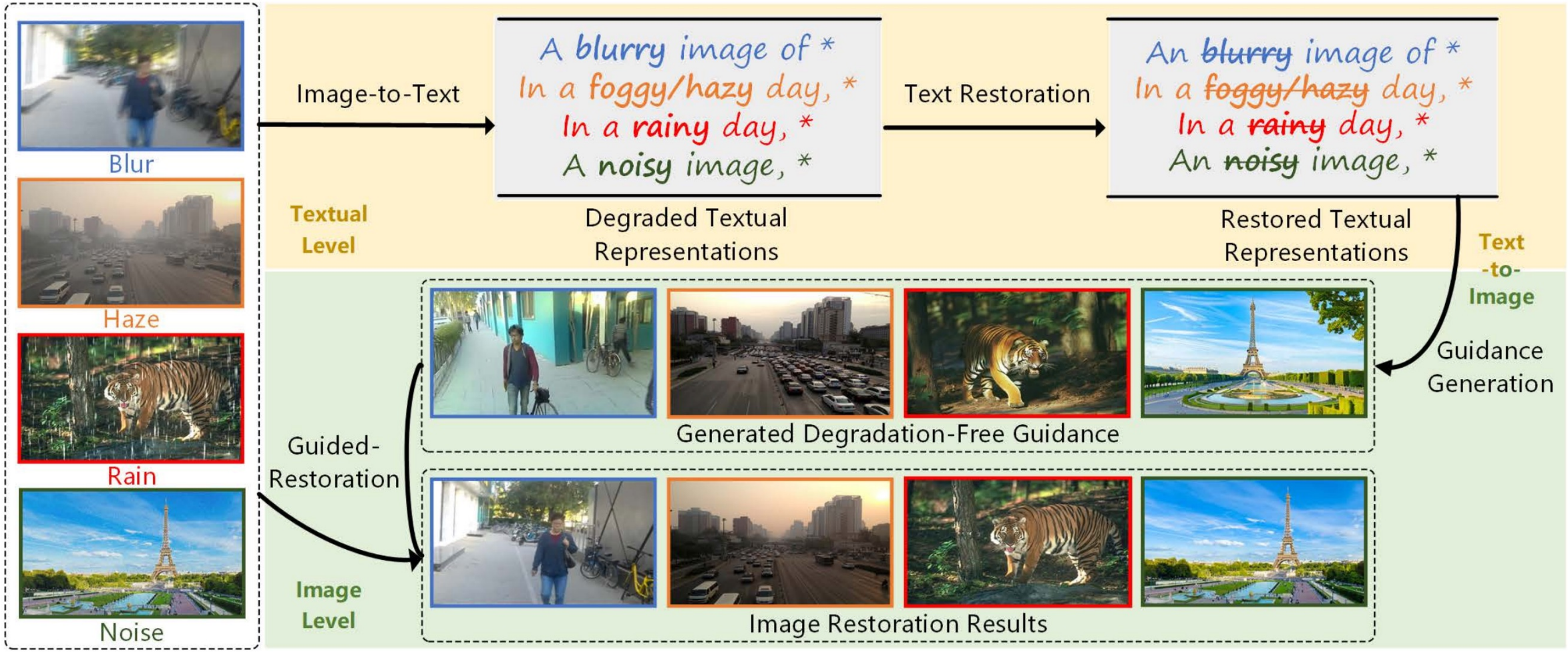
Introduction

- Introduce a new perspective for improving image restoration by **removing degradation in the textual representations** of a given degraded image.

- Propose to map the degraded images into textual representations for removing the degradations, and then **convert the restored textual representations into a guidance image** for **assisting image restoration**.

- The results showcase that our method outperforms state-of-the-art ones across all these tasks.

Introduction



Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

Related Work

- PromptIR: Prompting for All-in-One Blind Image Restoration
  - NeurIPS 2023
- TextIR: A Simple Framework for Text-based Editable Image Restoration
  - IEEE TVCG 2024

Related Work

PromptIR: Prompting for All-in-One Blind Image Restoration

Vaishnav Potlapalli[⋆], **Syed Waqas Zamir**[†], **Salman Khan**[⋆], **Fahad Shahbaz Khan**[⋆]
[⋆]Mohamed bin Zayed University of AI, [†]Inception Institute of AI
firstname.lastname@mbzuai.ac.ae

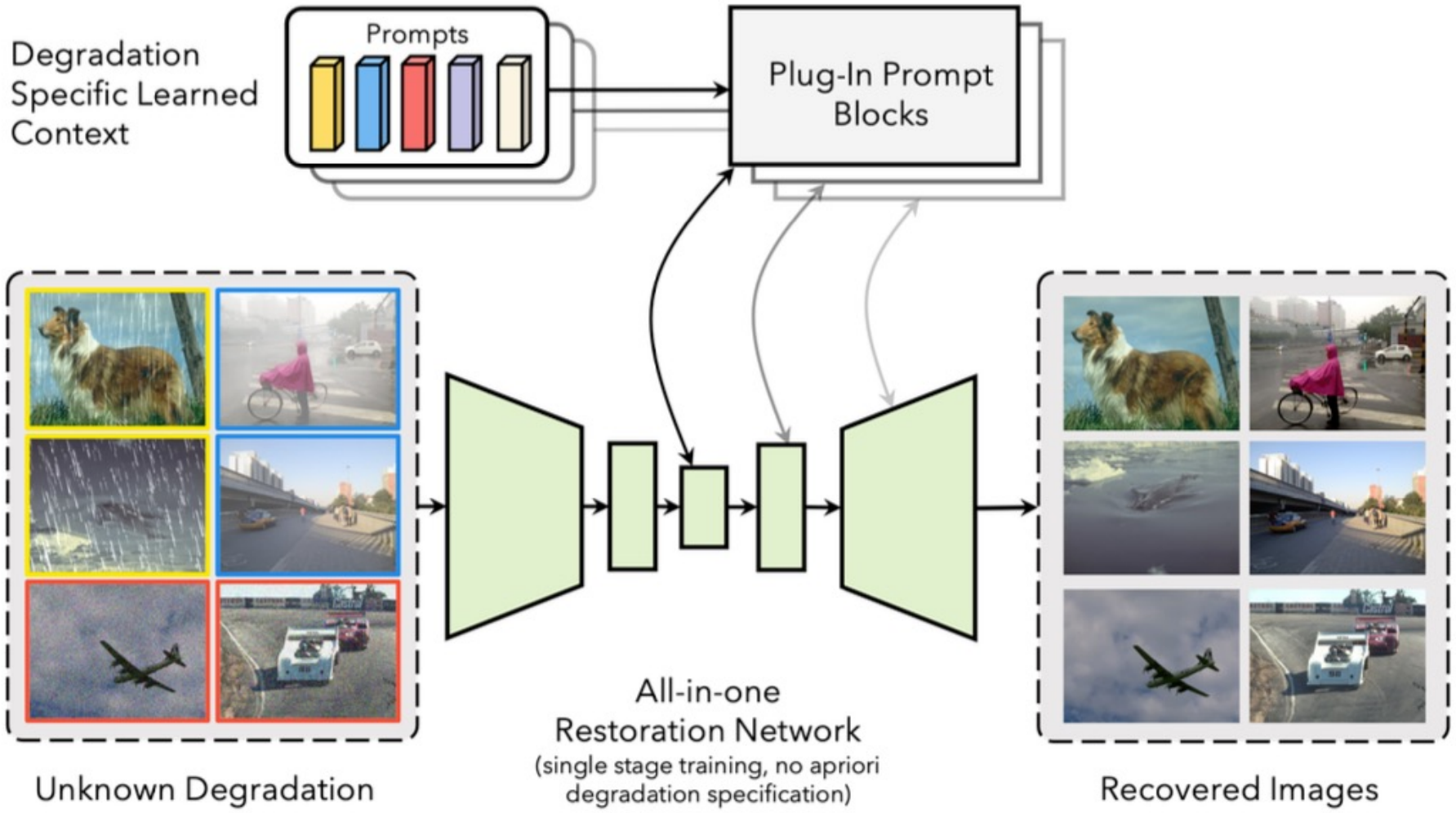
NeurIPS 2023

Introduction

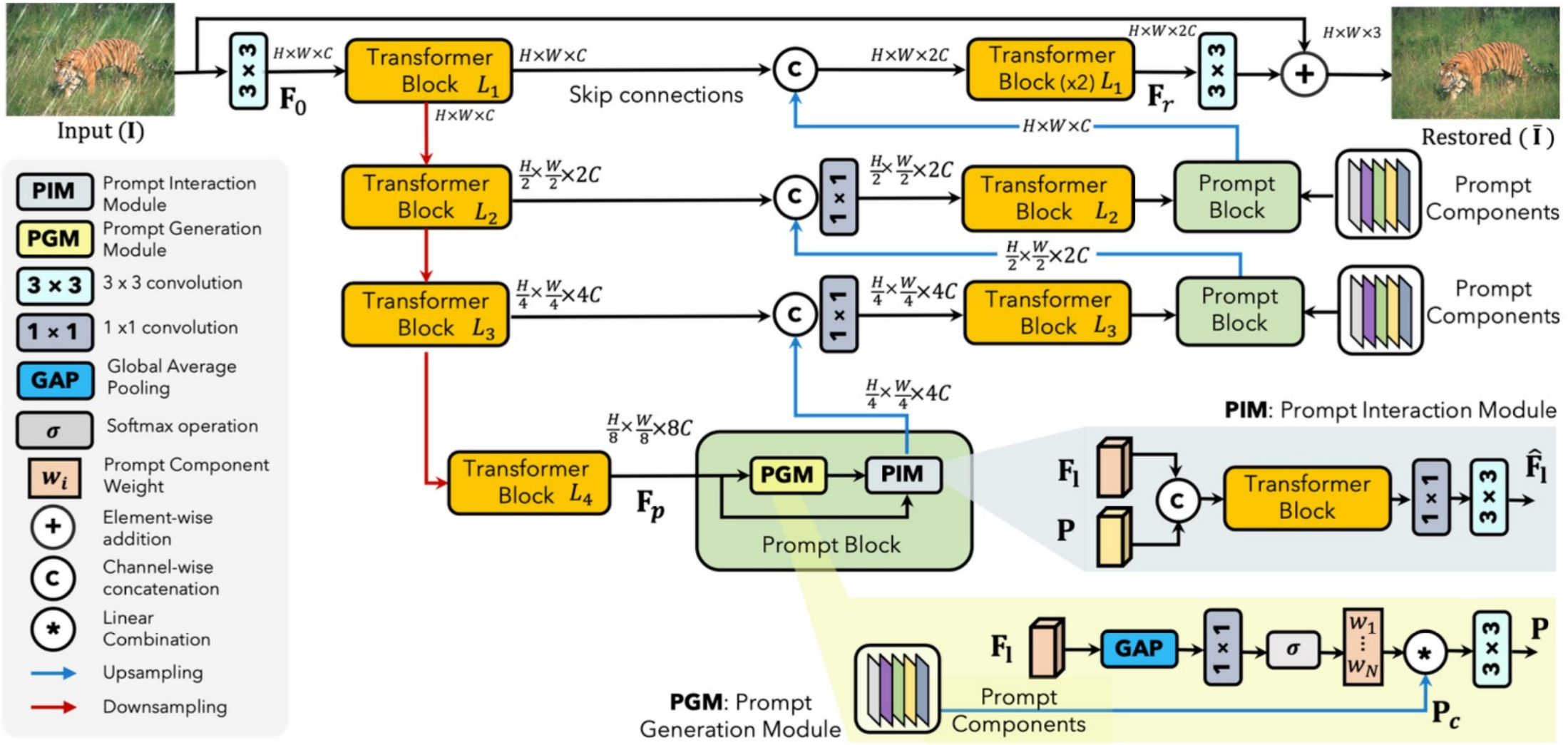
- Use **prompts to encode degradation-specific information**, which is then used to dynamically guide the restoration network.

- Offers a generic and efficient **plugin module with few lightweight prompts** that can be used to restore images of **various types and levels** of degradation with no prior information on the corruptions present in the image.

Introduction



Framework

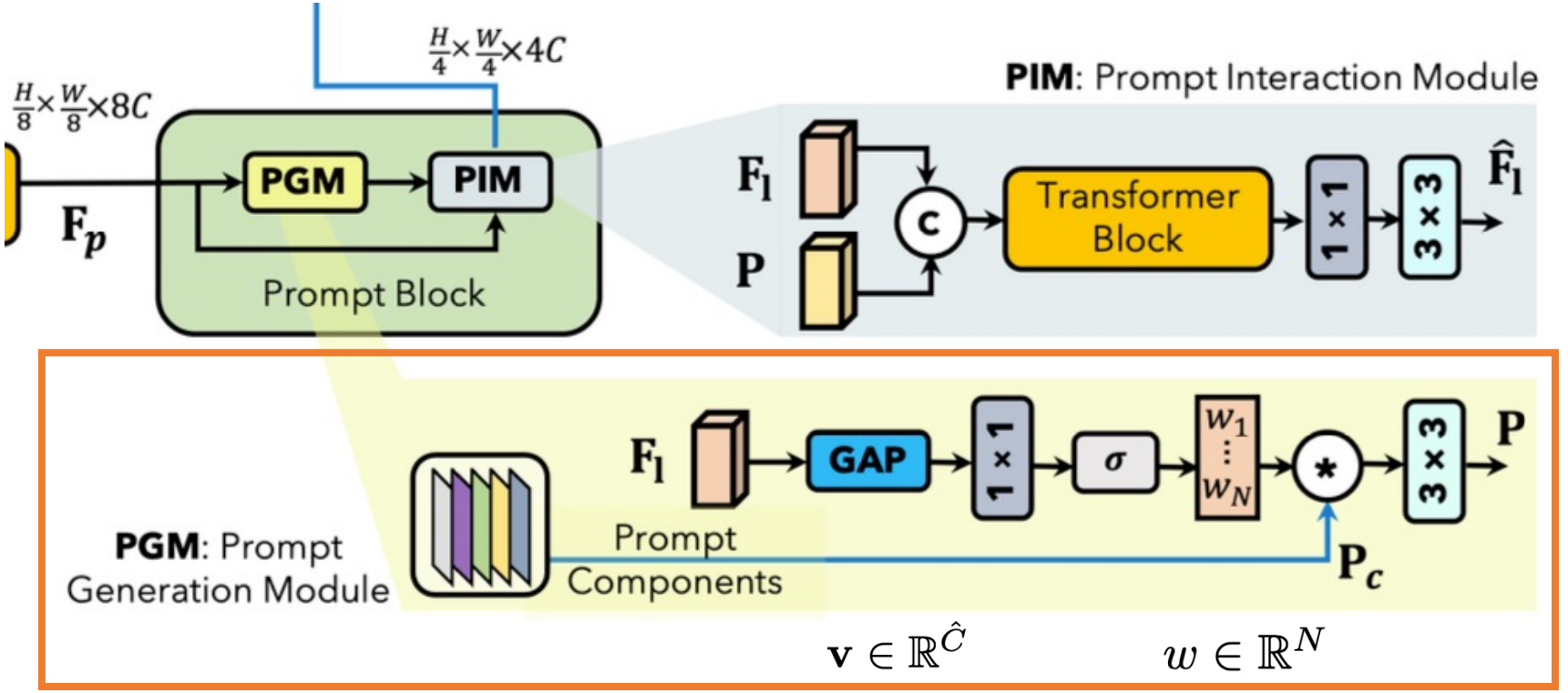


$$\hat{\mathbf{F}}_\mathbf{l} = \texttt{PIM}(\texttt{PGM}(\mathbf{P_c}, \mathbf{F_l}), \mathbf{F_l})$$

$$\mathbf{F_l} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$$

$$\mathbf{P_c} \in \mathbb{R}^{N\times\hat{H}\times\hat{W}\times\hat{C}}$$

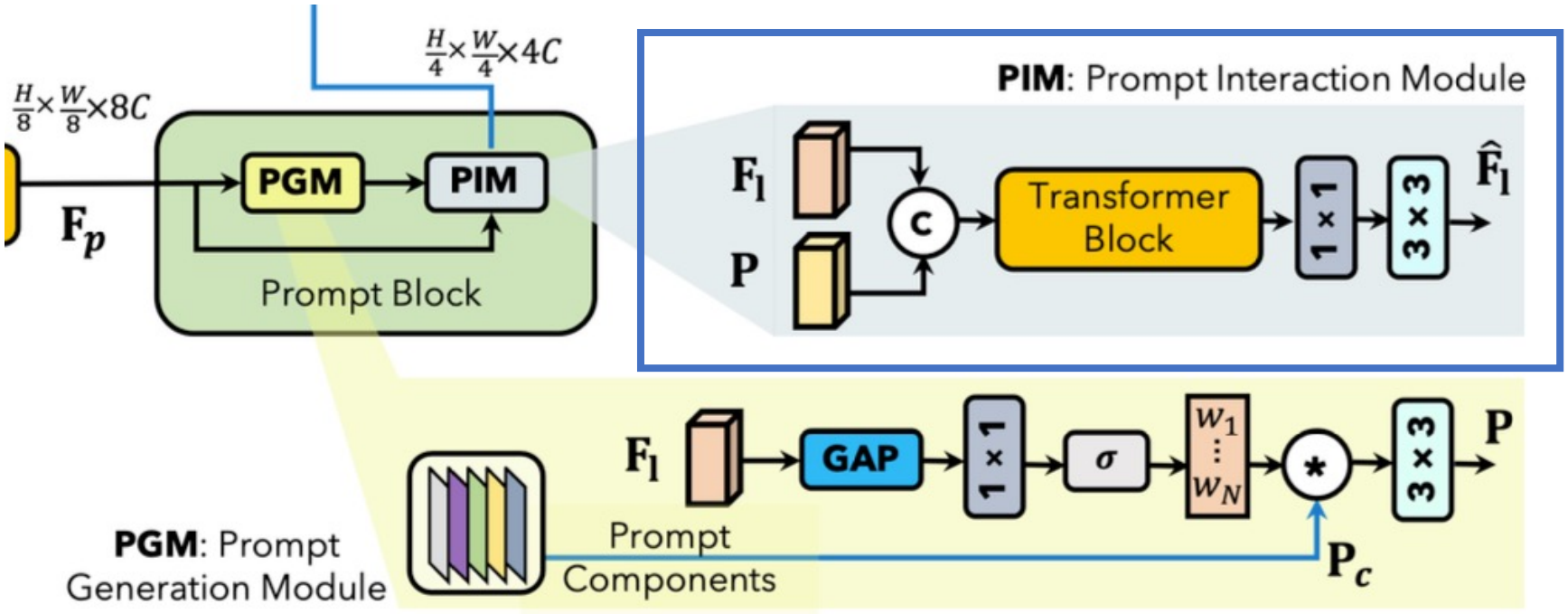
Prompt Generation Module



$$\mathbf{P} = \texttt{Conv}_{3\times3}\left(\sum_{c=1}^{N} w_i \mathbf{P}_c\right), \qquad w_i = \texttt{Softmax}(\texttt{Conv}_{1\times1}(\texttt{GAP}(\mathbf{F_l}))) \tag{2}$$

- Prompt components Pc interact with the incoming features to embed degradation information, dynamically yield input-conditioned prompts P.

Prompt Interaction Module

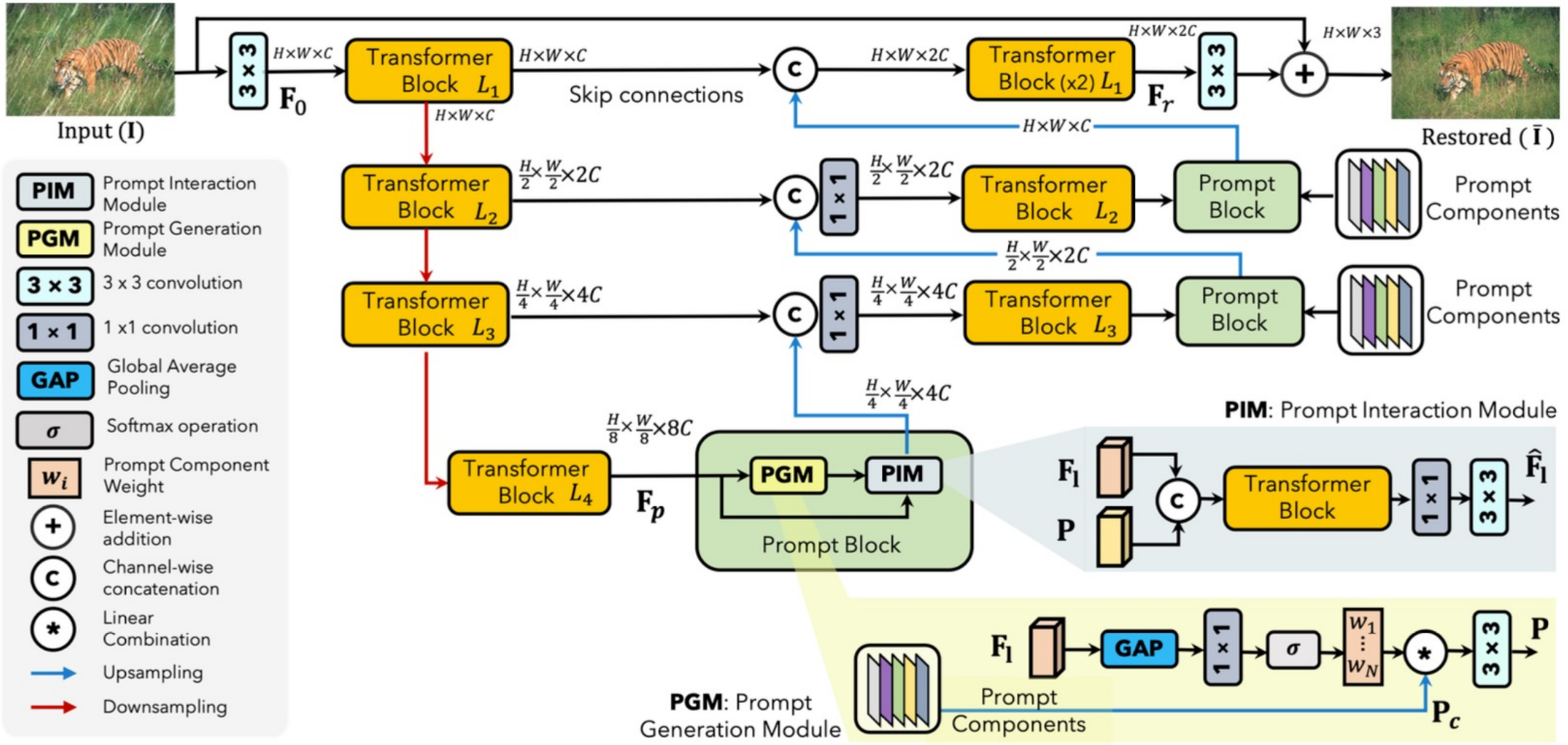


$$\hat{\mathbf{F}}_l = \texttt{Conv}_{3\times 3}(\texttt{GDFN}(\texttt{MDTA}[\mathbf{F}_l; \mathbf{P}])$$

MDTA is formulated as $\mathbf{Y} = W_p\mathbf{V} \cdot \texttt{Softmax}\left(\mathbf{K} \cdot \mathbf{Q}/\alpha\right) + \mathbf{X}.$

GDFN is defined as $\mathbf{Z} = W_p^0\left(\phi(W_d^1 W_p^1(\texttt{LN}(\mathbf{Y}))) \odot W_d^2 W_p^2(\texttt{LN}(\mathbf{Y}))\right) + \mathbf{Y}.$

Framework

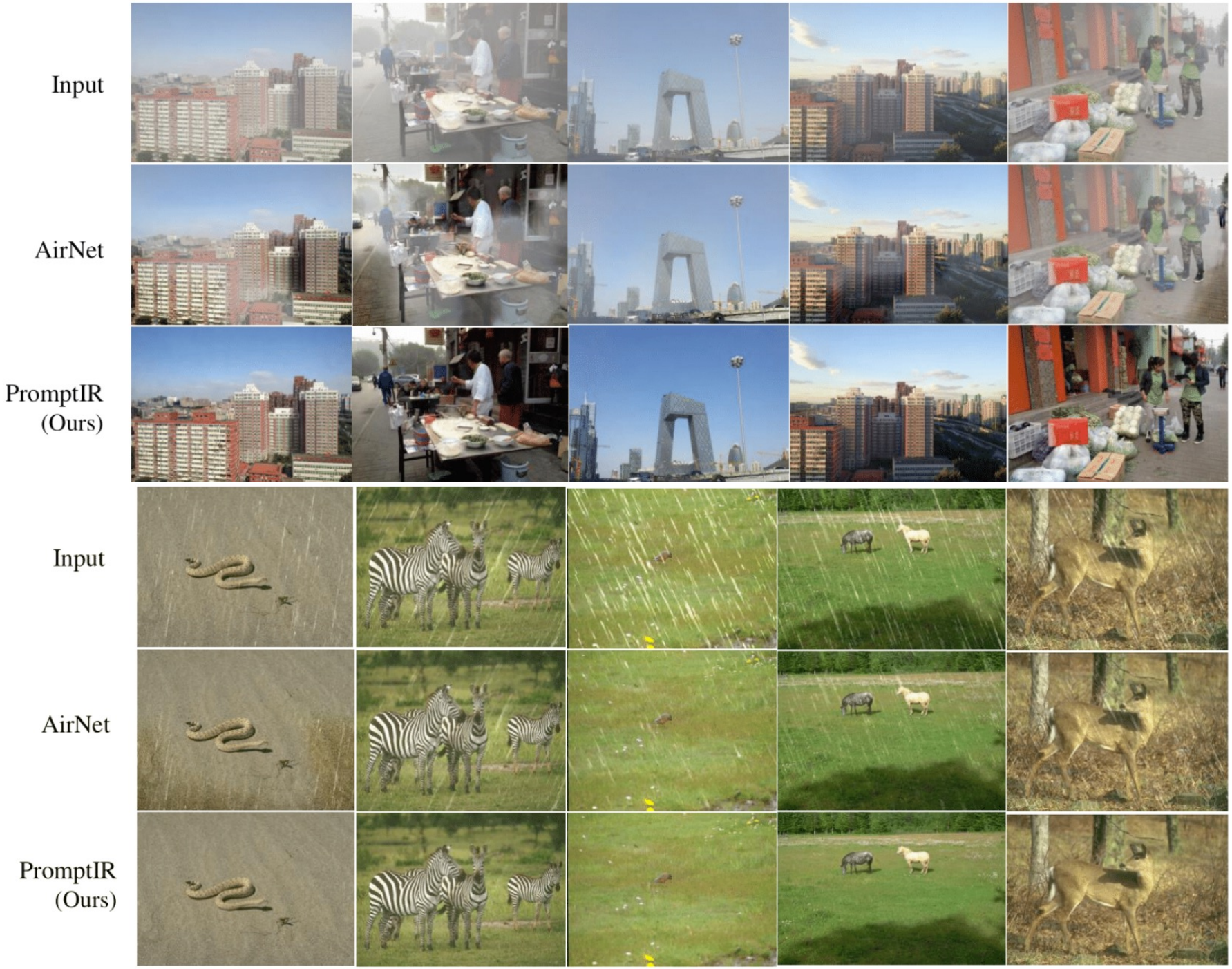


$$\hat{\mathbf{F}}_\mathbf{l} = \texttt{PIM}(\texttt{PGM}(\mathbf{P_c}, \mathbf{F_l}), \mathbf{F_l})$$

$$\mathbf{F_l} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$$

$$\mathbf{P_c} \in \mathbb{R}^{N\times\hat{H}\times\hat{W}\times\hat{C}}$$

Experiment



Experiment

Method	Dehazing on SOTS [31]	Deraining on Rain100L [16]	Denoising on BSD68 dataset [41]) $\sigma = 15$	$\sigma = 25$	$\sigma = 50$	Average
BRDNet [51]	23.23/0.895	27.42/0.895	32.26/0.898	29.76/0.836	26.34/0.836	27.80/0.843
LPNet [17]	20.84/0.828	24.88/0.784	26.47/0.7782	24.77/0.748	21.26/0.552	23.64/0.738
FDGAN [14]	24.71/0.924	29.89/0.933	30.25/0.910	28.81/0.868	26.43/0.776	28.02/0.883
MPRNet [70]	25.28/0.954	33.57/0.954	33.54/0.927	30.89/0.880	27.56/0.779	30.17/0.899
DL[16]	26.92/0.391	32.62/0.931	33.05/0.914	30.41/0.861	26.90/0.740	29.98/0.875
AirNet [29]	27.94/0.962	34.90/0.967	33.92/0.933	31.26/0.888	28.00/0.797	31.20/0.910
PromptIR (Ours)	**30.58/0.974**	**36.37/0.972**	**33.98/0.933**	**31.31/0.888**	**28.06/0.799**	**32.06/0.913**

Related Work

TextIR: A Simple Framework for Text-based Editable Image Restoration

Yunpeng Bai[1], Cairong Wang[1], Shuzhao Xie[1], Chao Dong[2,3], Chun Yuan[1,4], Zhi Wang[1,4]

[1] Tsinghua University, [2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3] Shanghai AI Laboratory, China, [4]Peng Cheng Laboratory, Shenzhen, China

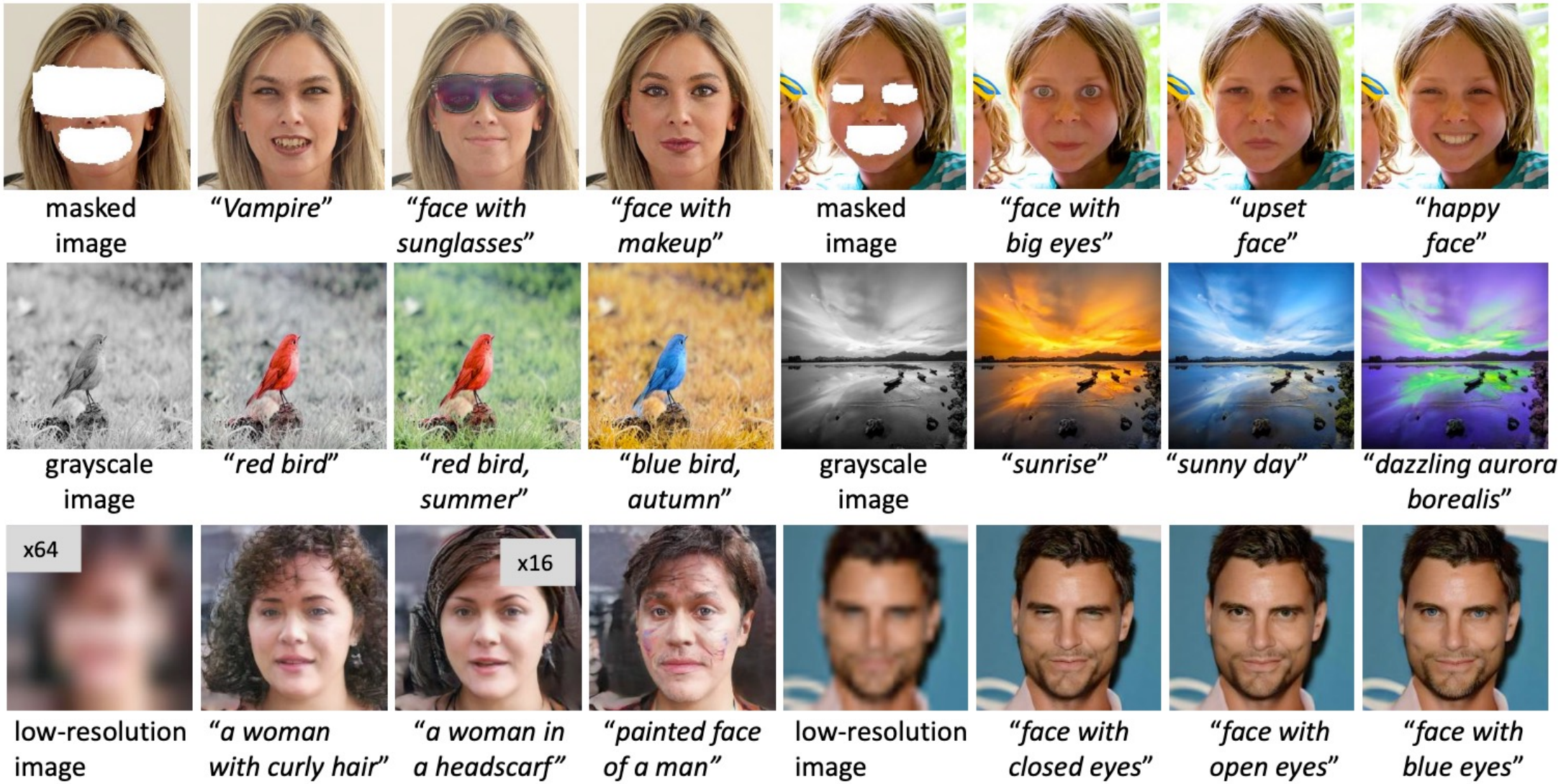
IEEE TVCG 2024

Introduction

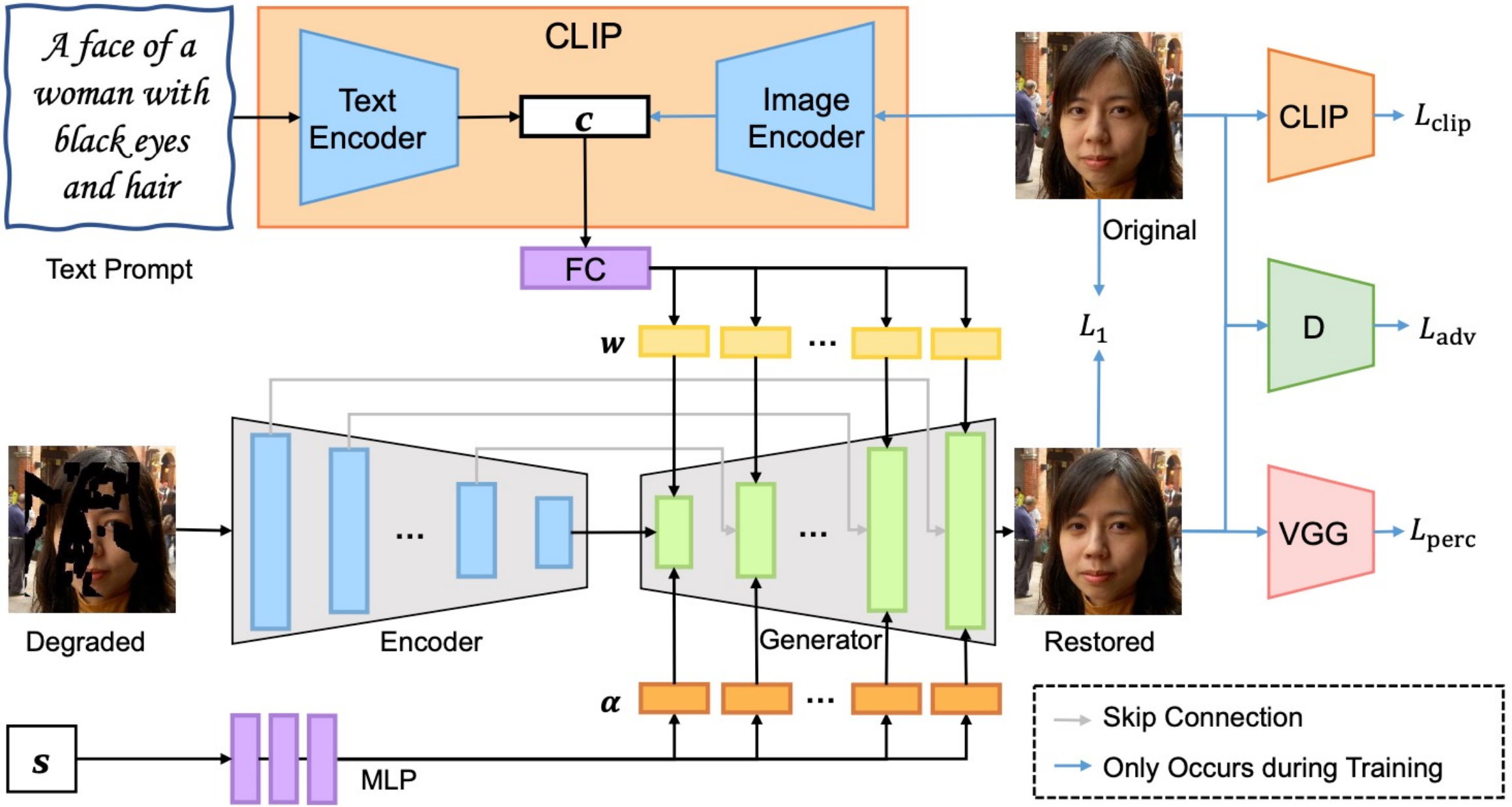
- Design a simple and effective framework that allows the user to **use text input to get desired image restoration results**, including image inpainting, image super-resolution, and image colorization.

- Take advantage of CLIP's text-image feature compatibility to enable a better fusion of image and text features and propose the **first text-based image restoration framework**.

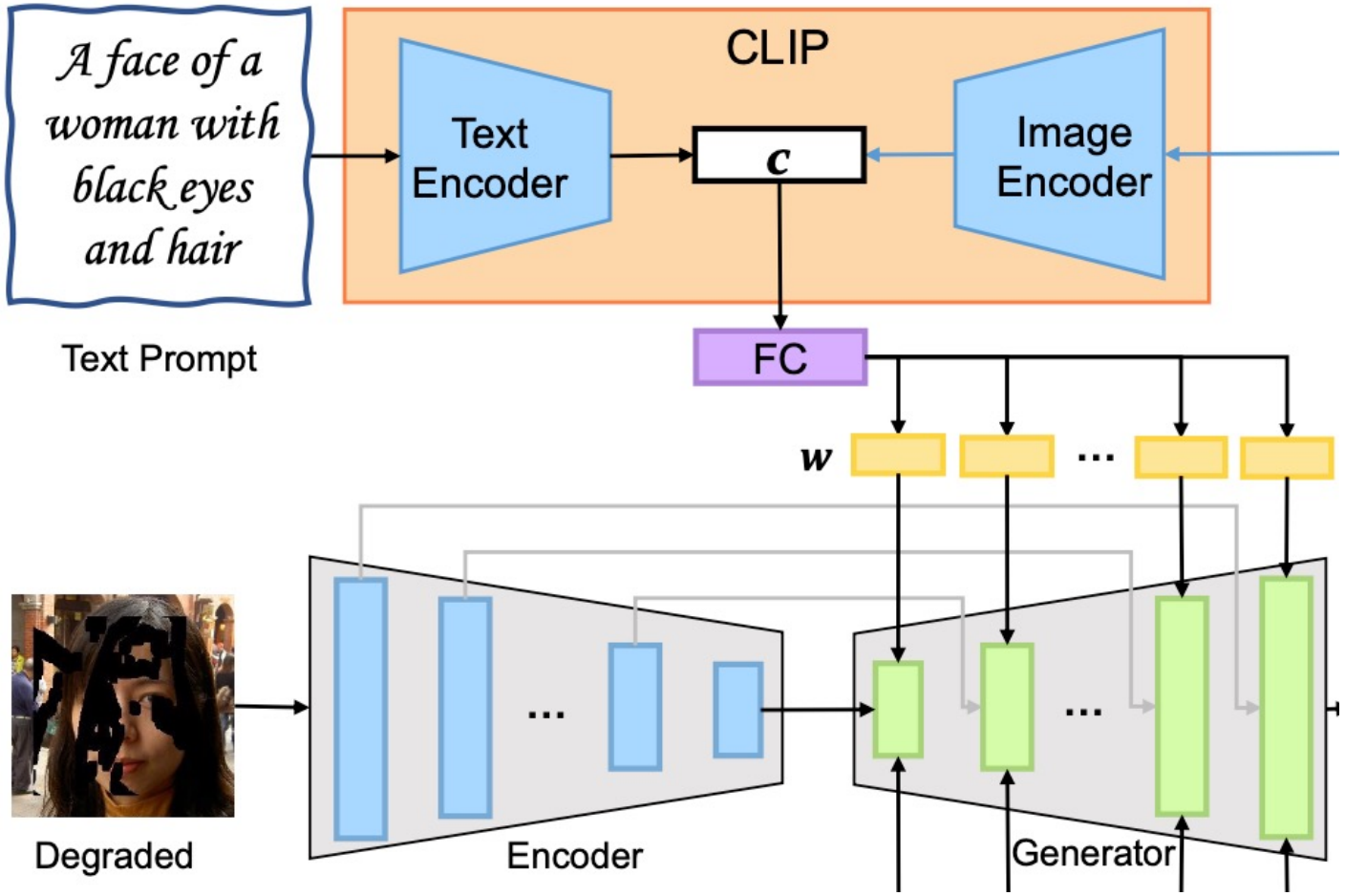
Introduction



Framework


A face of a woman with black eyes and hair
Text Prompt
CLIP
Text Encoder
c
Image Encoder
FC
w
Original
CLIP
L_{clip}
L_1
D
L_{adv}
Degraded
Encoder
Generator
Restored
VGG
L_{perc}
α
s
MLP
Skip Connection
Only Occurs during Training

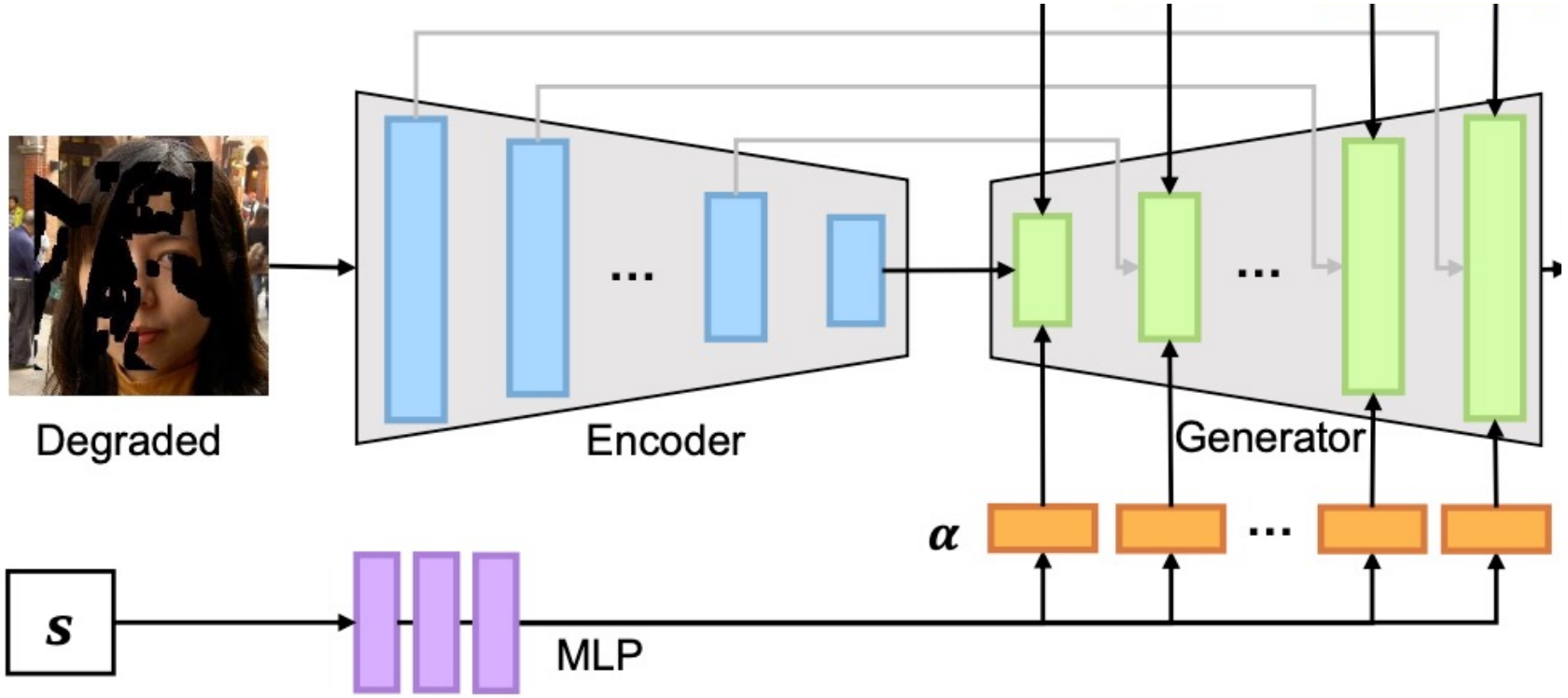
Style modulation



$$w = \text{Reshape}(\text{FC}(c)),\ c = E_I(I_{gt}) \text{ or } c = E_T(\text{text}). \quad (2)$$

$$g^i = \begin{cases} \text{StyleConv}(f^{l-i}, w^i) & i = 0, \\ \text{StyleConv}(\text{StyleConv}(\uparrow_2 (x^{i-1}), w_1^i), w_2^i) & i > 0, \end{cases} \quad (3)$$

Feature fusion



$$\boldsymbol{\alpha} = \{\alpha_{1,2}^{0}, ..., \alpha_{1,2}^{l-1}, \alpha_{1,2}^{l}\} = \text{Reshape}(\text{MLP}(s)). \quad (4)$$

$$\alpha_{enc/gen}^{i} = \frac{|\alpha_{1/2}^{i}|}{\sqrt{{\alpha_{1}^{i}}^{2} + {\alpha_{2}^{i}}^{2} + \epsilon}}, \quad \alpha_{1/2}^{i} \in \boldsymbol{\alpha}, \quad (5)$$

$$x^{i} = \alpha_{enc}^{i} \cdot \text{Conv}(f^{l-i}) + \alpha_{gen}^{i} \cdot g^{i},$$

Objective Functions

- non-saturating adversarial loss

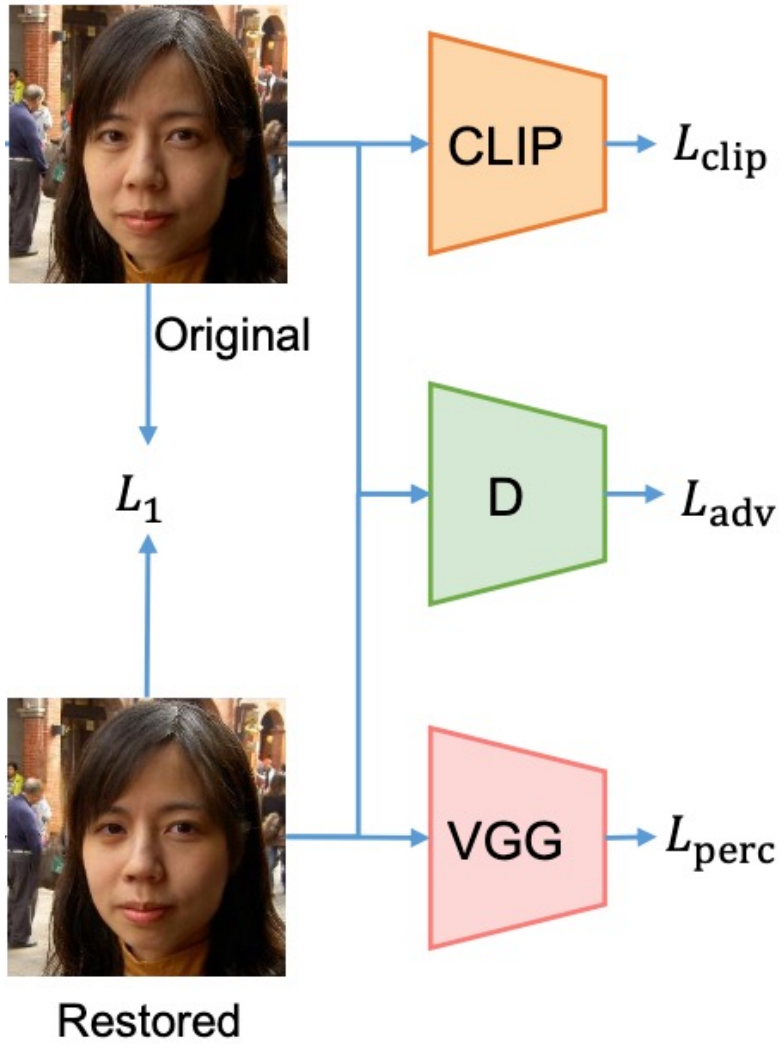
$$\begin{aligned}\mathcal{L}_{adv,D} &= \mathbb{E}[\log(1+\exp(-D(I_{gt})))\\ &\quad + \log(1+\exp(D(G(I_d,c))))],\\ \mathcal{L}_{adv,G} &= \mathbb{E}[\log(1+\exp(-D(G(I_d,c))))],\end{aligned} \tag{6}$$

- CLIP loss

$$\mathcal{L}_{clip} = 1 - \frac{E_I(I_r)\cdot E_I(I_{gt})}{|E_I(I_r)||E_I(I_{gt})|}. \tag{7}$$

- Total loss

$$\begin{aligned}\mathcal{L}_D &= \lambda_{adv}\mathcal{L}_{adv,D},\\ \mathcal{L}_G &= \lambda_{adv}\mathcal{L}_{adv,G} + \lambda_{clip}\mathcal{L}_{clip} + \lambda_{l_1}\mathcal{L}_1 + \lambda_{perc}\mathcal{L}_{perc}.\end{aligned} \tag{8}$$



Experiment

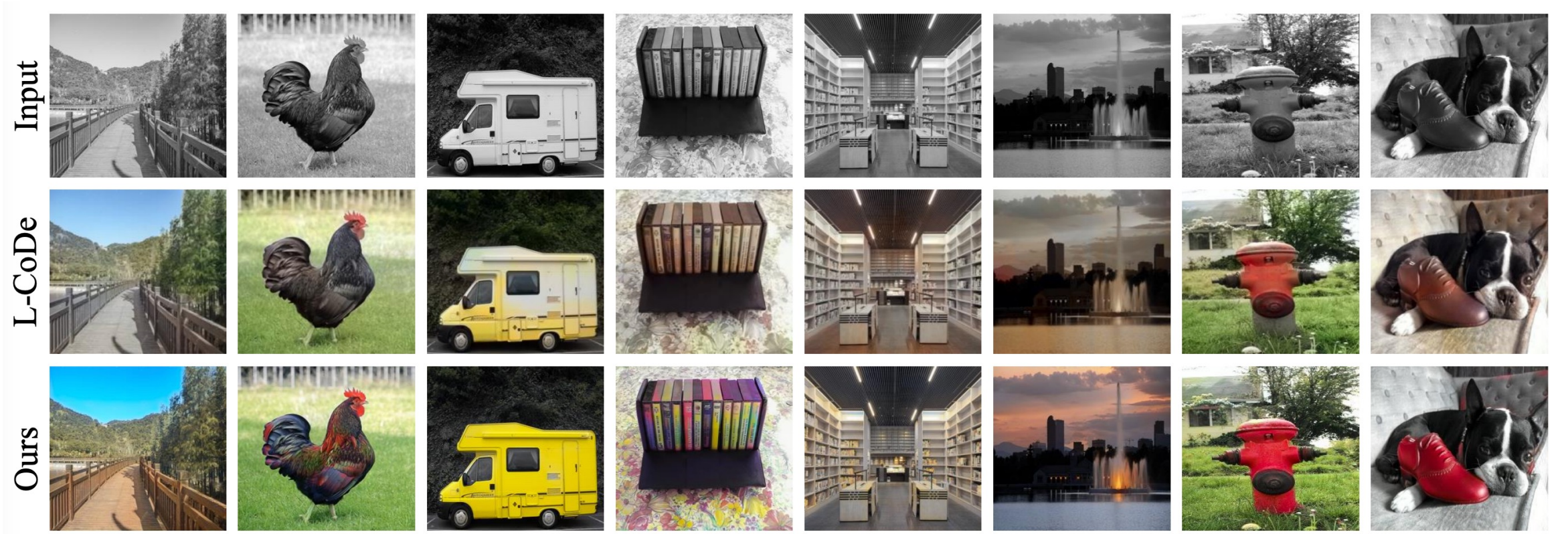
Table 1. Quantitative evaluation of inpainting experiment.

Methods	Metrics		
	PSNR ↑	SSIM↑	LPIPS↓
Blended-diffusion [3]	23.16	0.901	0.226
Ours	**29.83**	**0.932**	**0.068**

Experiment

Table 2. Quantitative evaluation of colorization experiment.

Methods	Metrics		
	PSNR ↑	SSIM↑	LPIPS↓
L-CoDe [50]	24.965	0.916	0.169
Ours	**26.70**	**0.923**	**0.124**



Experiment

Table 3. Quantitative evaluation of super-resolution experiment.

Methods	Metrics		
	PSNR ↑	SSIM↑	LPIPS↓
GPEN [53]	26.82	0.704	0.273
Ours	**27.41**	**0.784**	**0.227**

Conclusion of related works

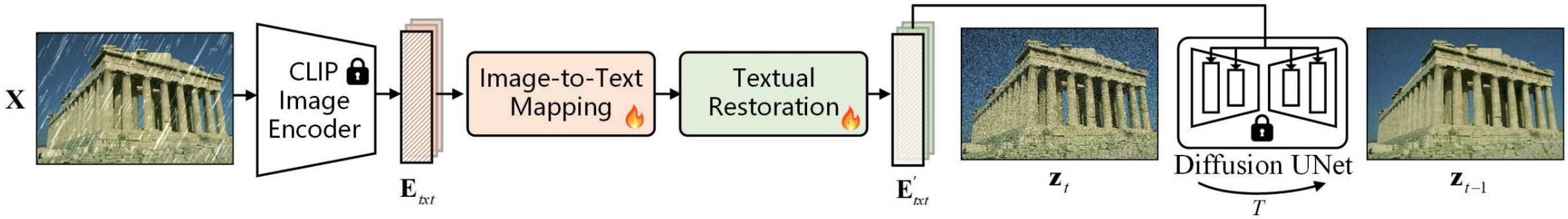
- PromptIR: Prompting for All-in-One Blind Image Restoration
  - ✓ Proposed a **drop-in prompt block** that can interact with the input features to **dynamically adjust the representations** such that the restoration process is **adapted for the relevant degradation**.

- TextIR: A Simple Framework for Text-based Editable Image Restoration
  - ✓ **Frist model use text information** to assist in **image restoration** because text input is more readily available and provides information with higher flexibility.

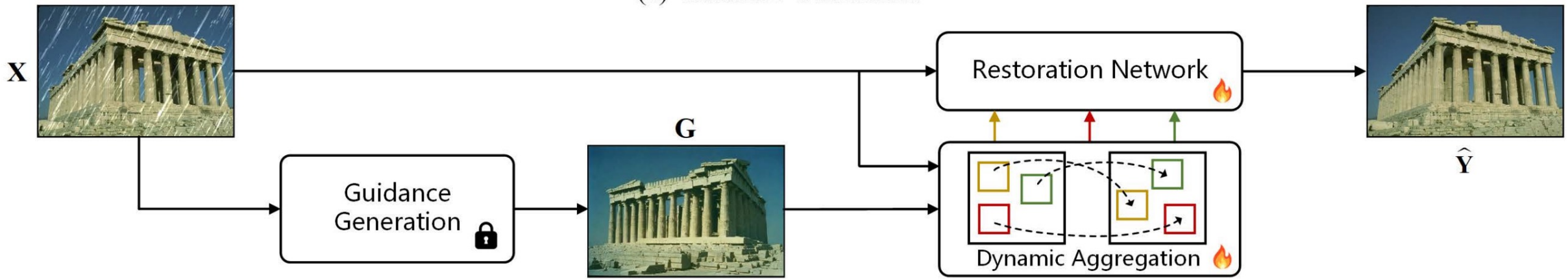
Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

Framework



(a) Guidance Generation



(b) Guided Image Restoration.

Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

Image-to-Text Mapping

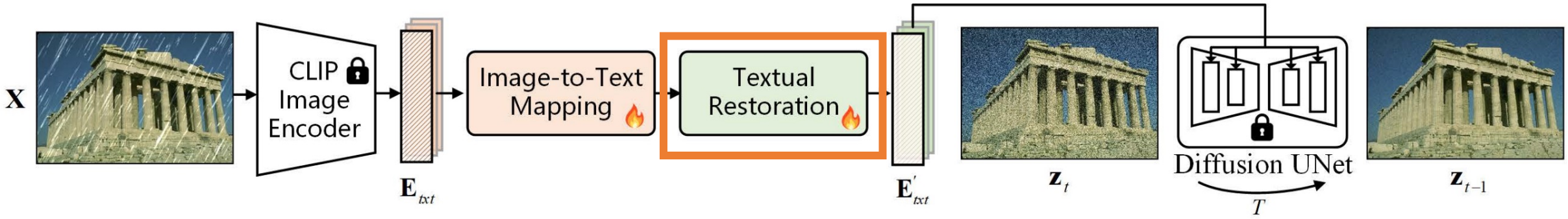
(a) Guidance Generation

$$\mathbf{E}_{txt} = \mathcal{M}_{i2t}(\tau_\theta^i(\mathbf{X})), \tag{2}$$

$$\mathbf{E}_{txt} \in \mathbb{R}^{N \times D} \qquad N=20$$

$$L_{LDM} = \mathbb{E}_{\mathbf{z}\sim\mathcal{E}(\mathbf{X}),\mathbf{p},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta^t(\mathbf{p}))\|_2^2\right], \tag{1}$$

Textual Restoration

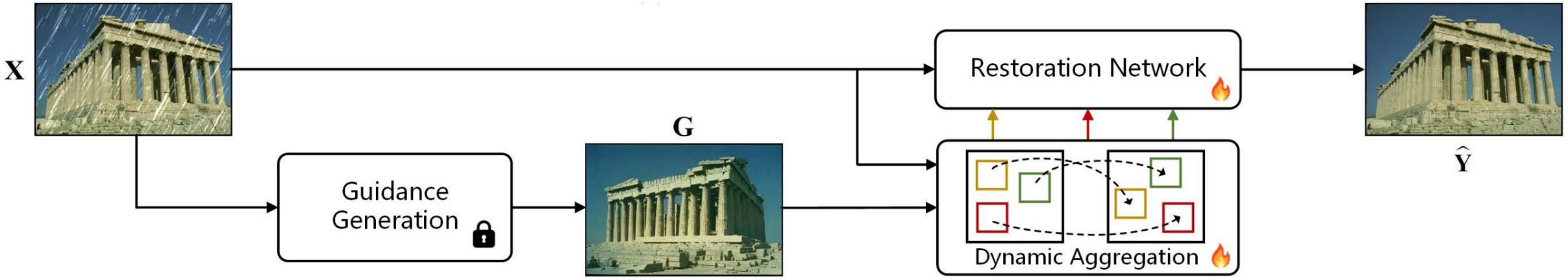


(a) Guidance Generation

$$\mathbf{E}'_{txt} = \mathcal{M}_{clean}(\mathbf{E}_{txt}), \tag{3}$$

$$L_{LDM} = \mathbb{E}_{\mathbf{z}\sim\mathcal{E}(\mathbf{X}),\mathbf{p},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta^t(\mathbf{p}))\|_2^2\right], \tag{1}$$

Guided Restoration

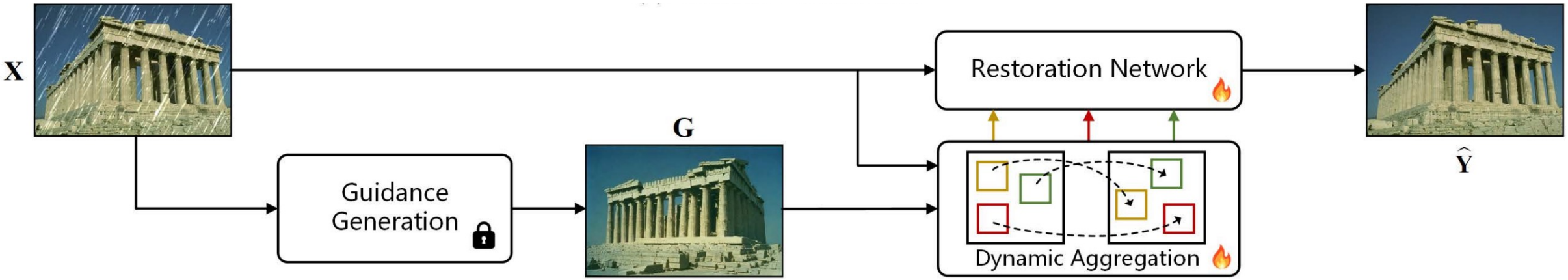


(b) Guided Image Restoration.

$$\mathbf{F}_x = \mathbf{F}_x + \alpha \cdot \mathcal{B}([\mathbf{F}_x, \hat{\mathbf{F}}_g]),$$

- Feature matching
  - shared encoder to extract multi-scale feature from degraded input and clean guidance
  - coarse-to-fine manner to match useful information
- Feature aggregation
  - concatenation & residual/self-attention blocks

Reconstruction loss



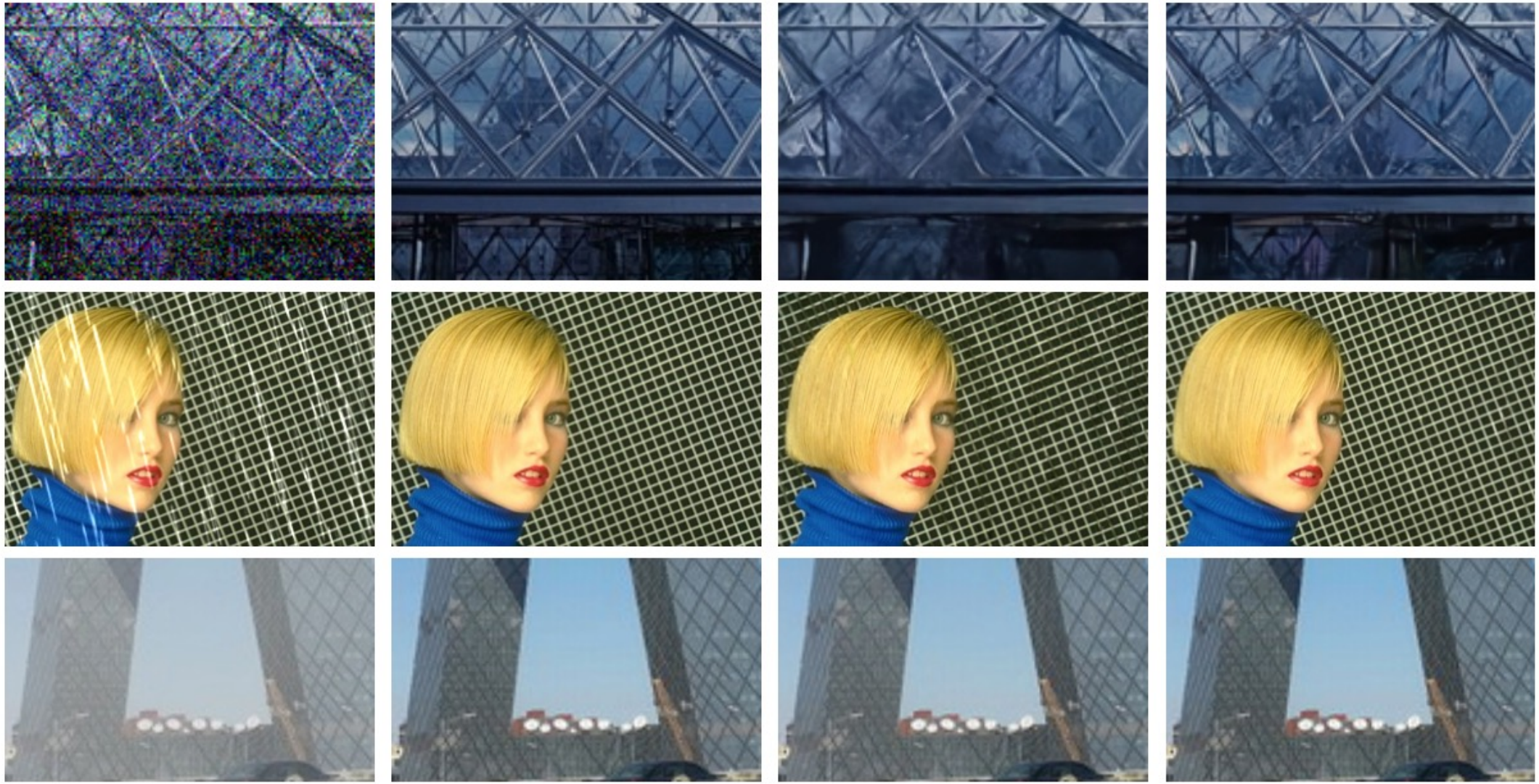
(b) Guided Image Restoration.

$$\ell_1 = ||\hat{\mathbf{Y}} - \mathbf{Y}||_1$$

Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

All-in-one Results



All-in-one Results

Method	Dehazing on SOTS	Derain on Rain100L	Denoise on BSD68 $\sigma = 15$	Denoise on BSD68 $\sigma = 25$	Denoise on BSD68 $\sigma = 50$	Average
BRDNet [91]	23.23/0.895	27.42/0.895	32.26/0.898	29.74/0.836	26.34/0.836	27.80/0.843
LPNet [34]	20.84/0.828	24.88/0.784	26.47/0.778	24.77/0.748	21.26/0.552	23.64/0.738
FDGAN [33]	24.71/0.924	29.89/0.933	30.25/0.910	28.81/0.868	26.43/0.776	28.02/0.883
MPRNet [113]	25.28/0.954	33.57/0.954	33.54/0.927	30.89/0.880	27.56/0.779	30.17/0.899
DL [28]	26.92/0.391	32.62/0.931	33.05/0.914	30.41/0.861	26.90/0.740	29.98/0.875
AirNet [51]	27.94/0.962	34.90/0.967	33.92/0.933	31.26/0.888	28.00/0.797	31.20/0.910
PromptIR [73]	30.58/0.974	36.37/0.972	33.98/0.933	31.31/0.888	28.06/0.799	32.06/0.913
Ours	**31.63/0.980**	**37.58/0.979**	**34.01/0.933**	**31.39/0.890**	**28.18/0.802**	**32.56/0.916**

Image Deblurring Results

- Motion image deblurring

Method	GoPro [68]		HIDE [86]		RealBlur-R [81]		RealBlur-J [81]	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DBGAN [121]	31.10	0.942	28.94	0.915	33.78	0.909	24.93	0.745
MT-RNN [70]	31.15	0.945	29.15	0.918	35.79	0.951	28.44	0.862
DMPHN [116]	31.20	0.940	29.09	0.924	35.70	0.948	28.42	0.860
SPAIR [74]	32.06	0.953	30.29	0.931	-	-	28.81	0.875
MIMO-Unet+ [19]	32.45	0.957	29.99	0.930	35.54	0.947	27.63	0.837
IPT [13]	32.52	-	-	-	-	-	-	-
MPRNet [113]	32.66	0.959	30.96	0.939	35.99	0.952	28.70	0.873
HINet [14]	32.71	0.959	30.32	0.932	-	-	-	-
Uformer [95]	32.97	0.967	-	-	-	-	-	-
Restormer [114]	32.92	0.961	31.22	0.942	36.19	0.957	28.96	**0.879**
Ours-Restormer	33.11	0.962	31.26	0.943	**36.47**	**0.959**	**29.17**	0.875
NAFNet [15]	33.69	0.966	31.32	0.943	33.62	0.944	26.33	0.856
Ours-NAFNet	**33.97**	**0.968**	**31.57**	**0.946**	33.87	0.950	26.76	0.861

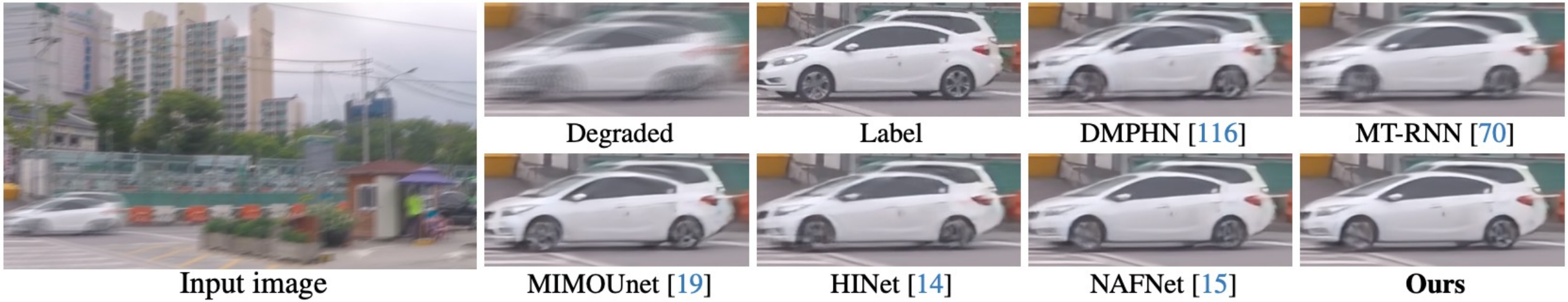
Image Deblurring Results

- Defocus image deblurring results

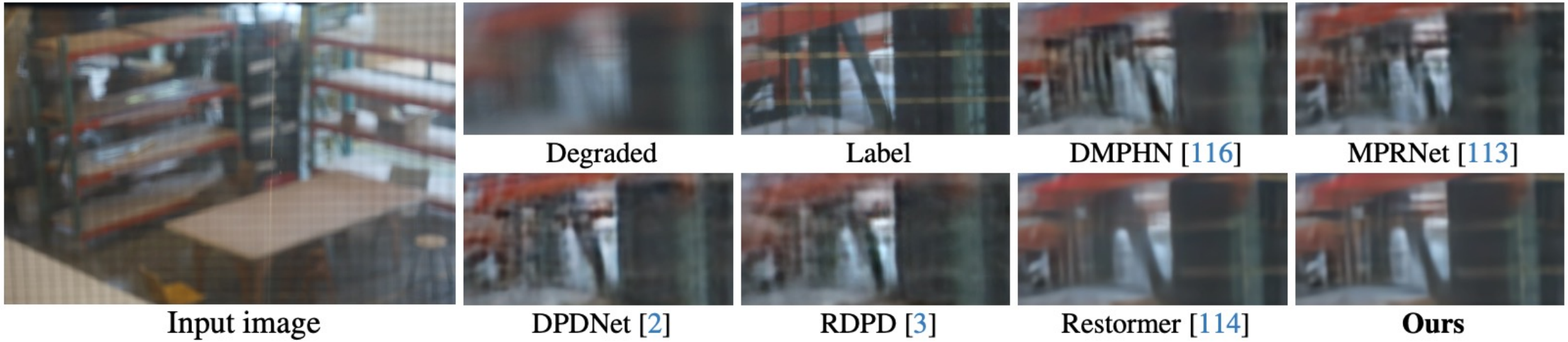
Method	Indoor Scenes		Outdoor Scenes		Combined	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
EBDB$_S$ [44]	25.77	0.772	21.25	0.599	23.45	0.683
DMENet$_S$ [46]	25.50	0.788	21.43	0.644	23.41	0.714
JNB$_S$ [87]	26.73	0.828	21.10	0.608	23.84	0.715
DPDNet$_S$ [2]	26.54	0.816	22.25	0.682	24.34	0.747
KPAC$_S$ [88]	27.97	0.852	22.62	0.701	25.22	0.774
IFAN$_S$ [47]	28.11	0.861	22.76	0.720	25.37	0.789
Restormer$_S$ [114]	<u>28.87</u>	<u>0.882</u>	<u>23.24</u>	<u>0.743</u>	<u>25.98</u>	<u>0.811</u>
Ours$_S$	**29.11**	**0.889**	**23.35**	**0.748**	**26.15**	**0.817**
DPDNet$_D$ [2]	27.48	0.849	22.90	0.726	25.13	0.786
RDPD$_D$ [3]	28.10	0.843	22.82	0.704	25.39	0.772
Uformer$_D$ [95]	28.23	0.860	23.10	0.728	25.65	0.795
IFAN$_D$ [47]	28.66	0.868	23.46	0.743	25.99	0.804
Restormer$_D$ [114]	<u>29.48</u>	<u>0.895</u>	<u>23.97</u>	<u>0.773</u>	<u>26.66</u>	<u>0.833</u>
Ours$_D$	**29.62**	**0.899**	**24.16**	**0.775**	**26.82**	**0.835**

Image Deblurring Results

- Motion image deblurring



- Defocus image deblurring results

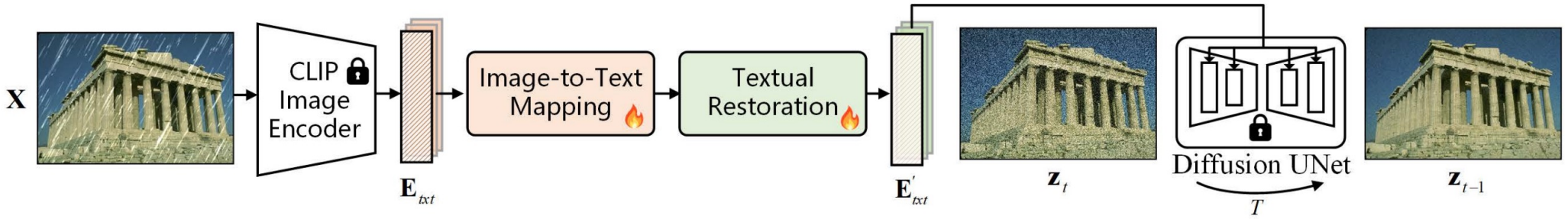


Dehaze & Derain Results

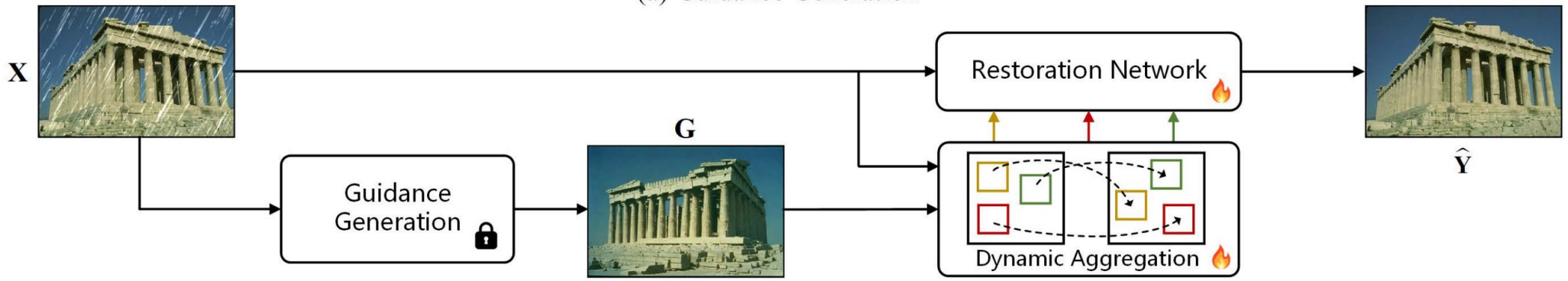
Method	SOTS-Indoor [50]		SOTS-Outdoor [50]	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DehazeNet [9]	19.82	0.821	24.75	0.927
AOD-Net [48]	20.51	0.861	24.14	0.920
GridDehazeNet [61]	32.16	0.984	30.86	0.982
MSBDN [26]	33.67	0.985	33.48	0.982
FFA-Net [75]	36.39	0.989	33.57	0.984
ACER-Net [97]	37.17	0.990	-	-
DeHamer [37]	36.63	0.988	35.18	0.986
MAXIM-2S [92]	38.11	0.991	34.19	0.985
PMNet [105]	38.41	0.990	34.74	0.985
DehazeFormer-L [90]	40.05	0.996	-	-
SFNet [20]	<u>41.24</u>	<u>0.996</u>	<u>40.05</u>	<u>0.996</u>
Ours	**41.48**	**0.996**	**40.29**	**0.996**

Method	Rain200L [104]		Rain200H [104]		DID-Data [115]		DDN-Data [30]	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DDN [29]	34.68	0.967	26.05	0.805	30.97	0.911	30.00	0.904
RESCAN [54]	36.09	0.967	26.75	0.835	33.38	0.941	31.94	0.935
PReNet [80]	37.80	0.981	29.04	0.899	33.17	0.948	32.60	0.946
MSPFN [42]	38.53	0.983	29.36	0.903	33.72	0.955	32.99	0.933
RCDNet [93]	39.17	0.989	30.24	0.904	34.08	0.953	33.04	0.947
MPRNet [113]	39.47	0.982	30.67	0.911	33.99	0.959	33.10	0.935
DualGCN [31]	40.73	0.989	31.15	0.912	34.37	0.962	33.01	0.949
SPDNet [106]	40.50	0.988	31.28	0.920	34.57	0.956	33.15	0.946
Uformer [95]	40.20	0.986	30.80	0.910	35.02	0.962	33.95	0.955
Restormer [114]	40.99	0.989	32.00	0.932	35.29	<u>0.964</u>	34.20	<u>0.957</u>
IDT [100]	40.74	0.988	<u>32.10</u>	**0.934**	34.89	0.962	33.84	0.955
DRSformer [17]	<u>41.21</u>	<u>0.989</u>	**32.16**	<u>0.933</u>	<u>35.24</u>	0.962	<u>34.23</u>	0.955
Ours	**41.59**	**0.990**	31.97	0.931	**35.46**	**0.964**	**34.57**	**0.958**

Ablation study



(a) Guidance Generation



(b) Guided Image Restoration.

Table 9. Effect of condition information.

Method	baseline	N=5	N=10	N=20	N=30	N=40
PSNR↑	30.16	31.13	31.36	31.57	31.51	31.60
SSIM↑	0.932	0.941	0.945	0.947	0.947	0.948

Table 10. Effect of integration strategy.

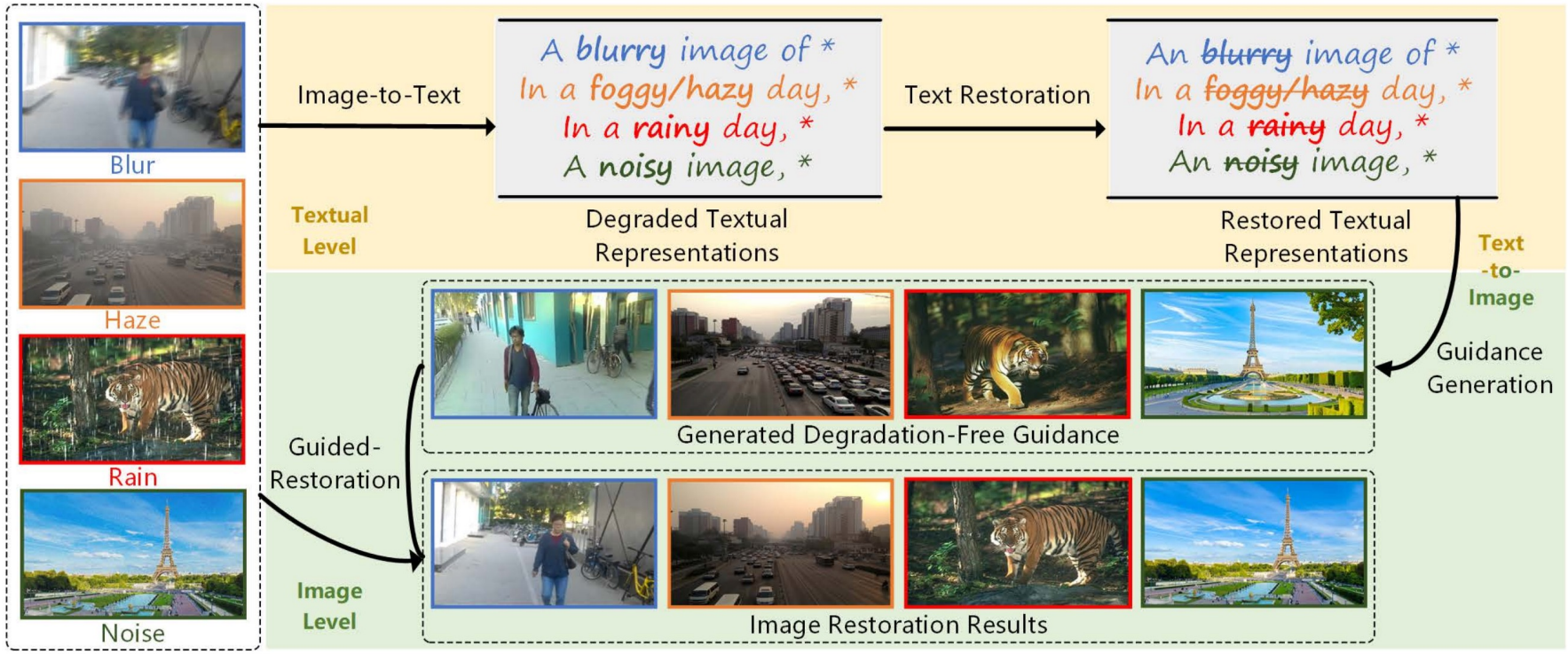
Method	baseline	Enc.	Dec.	Enc. & Dec.
PSNR↑	30.16	31.37	30.31	31.57
SSIM↑	0.932	0.946	0.934	0.947

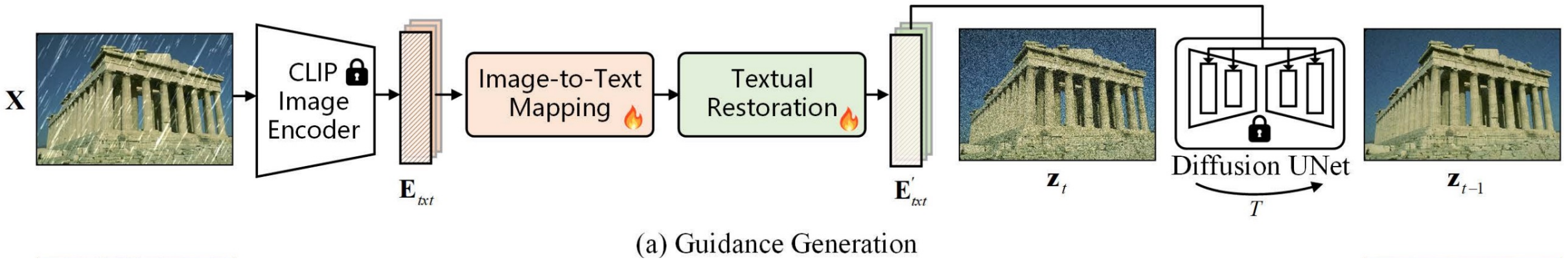
Table 11. Effect of generated guidance.

Method	baseline	Degra.	Ours
PSNR↑	30.16	30.13	31.57
SSIM↑	0.932	0.931	0.947

Ablation study

- Implicit vs Explicit textual representation





(a) Guidance Generation

Ablation study

- Implicit vs Explicit textual representation

Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- Consider that content and degradation are tightly coupled in image representation, while **decoupled in textual representation**.

- Propose to embed an **image-to-text mapper** and **textual restoration module** into CLIP-equipped **text-to-image models** to generate **clear guidance** from degraded images.

- Extensive experiments on multiple tasks demonstrate that method improves the performance of SOTA image restoration networks.