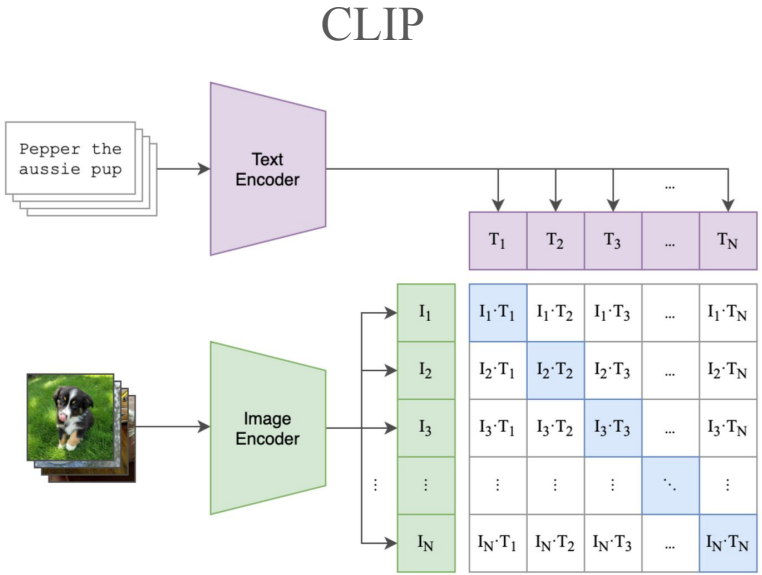
How to use text for image restoration

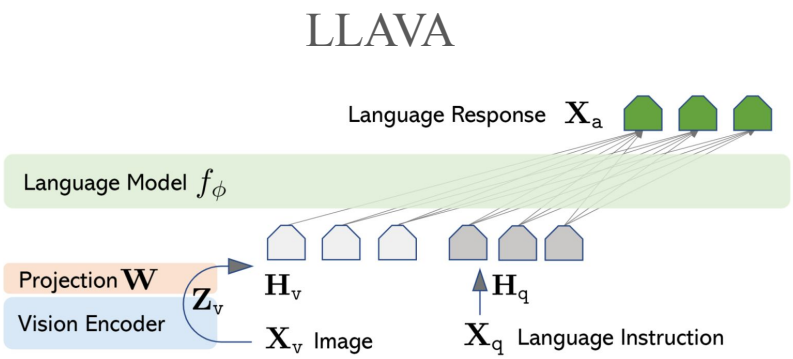
Donghun Ryou
dhryou@snu.ac.kr

2024.01.10

ComputerVisionLab
Seoul National University

Multimodal models' advancements





Usage of text in low-level vision:

1. Provide guidance in multi-task scenarios (e.g., all-in-one solutions) to decide which task to perform.

2. Offer clear guidance for ill-posed problems.

3. Serve as a simpler representation that can assist in complex image restoration tasks.

4. Leverage text's robust features

1. Provide guidance in multi-task scenarios

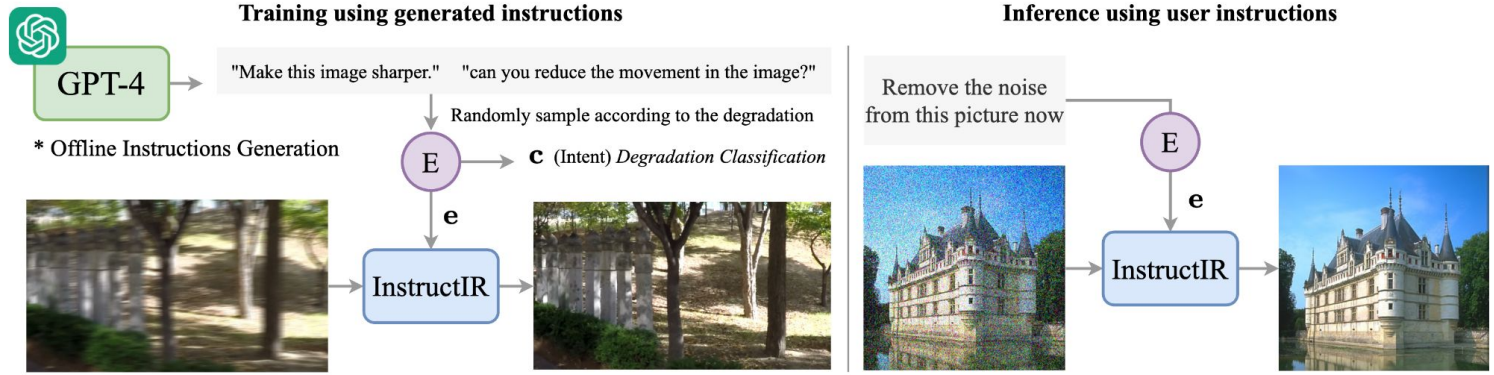
InstructIR: High-Quality Image Restoration Following Human Instructions

Marcos V. Conde[1,2], Gregor Geigle[1], and Radu Timofte[1]

[1] Computer Vision Lab, CAIDAS & IFI, University of Würzburg
[2] Visual Computing Group, FTG, Sony PlayStation

https://github.com/mv-lab/InstructIR (500 ☆)

2023 ECCV



1. Provide guidance in multi-task scenarios

- Single model can perform various low-level vision tasks in a controllable manner.

Input (1) *"Clear the rain from my picture"* ⟶ (2) *"Make it pop"*

(1) *"Retouch it as a photographer"* ⟶ "Can you remove the raindrops?" ⟶ "Increase the resolution"

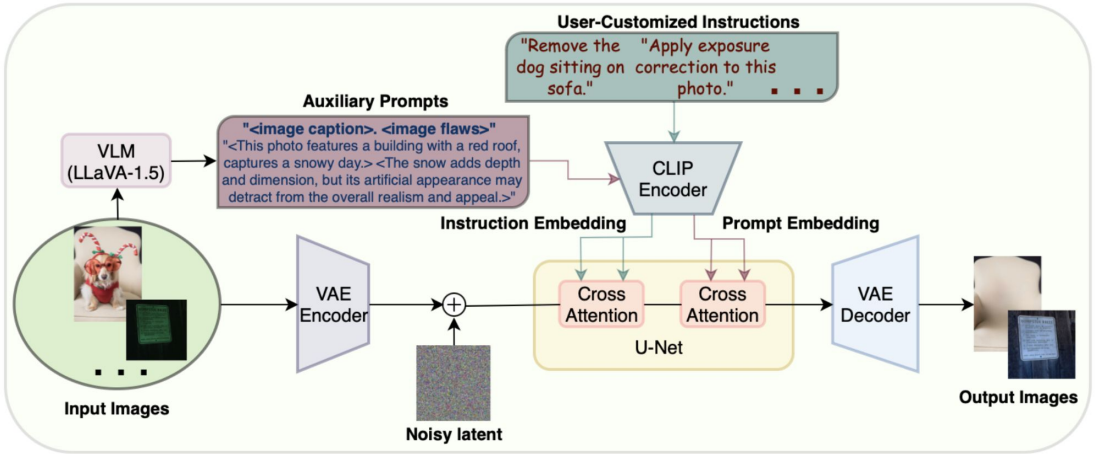
Input (1) *"My image is too dark, fix it"* ⟶ (2) *"Apply a tonemap"*

1. Provide guidance in multi-task scenarios

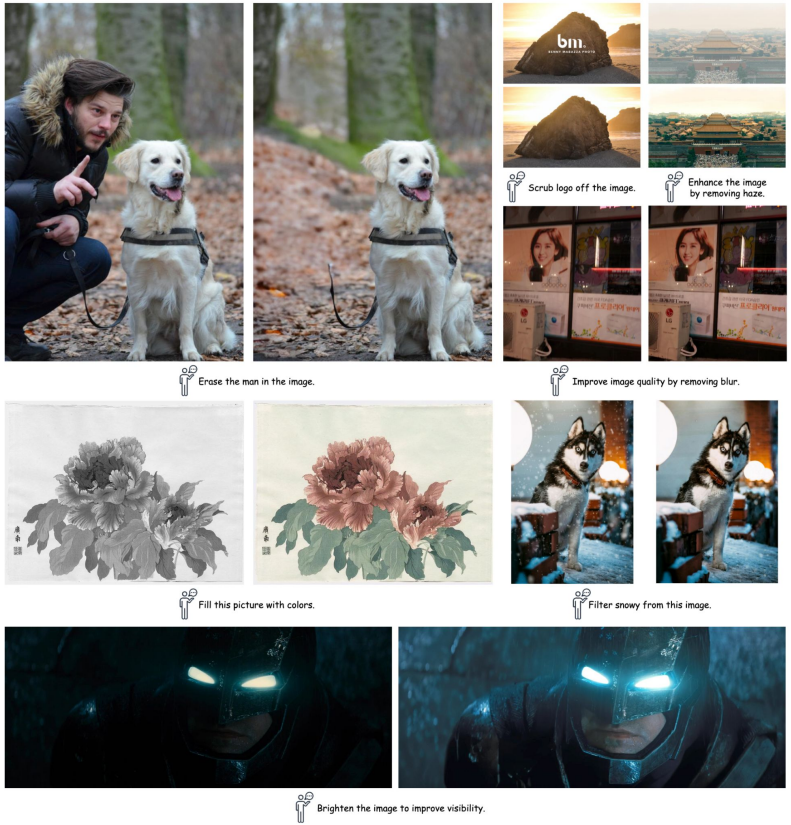
PromptFix: You Prompt and We Fix the Photo

Yongsheng Yu[1] * **Ziyun Zeng**[1] * **Hang Hua**[1] **Jianlong Fu**[2] **Jiebo Luo**[1]
[1]University of Rochester, [2]Microsoft Research
{yyu90,zzeng24}@ur.rochester.edu, {hhua2,jluo}@cs.rochester.edu, jianf@microsoft.com

2024 NeurIPS



1. Provide guidance in multi-task scenarios



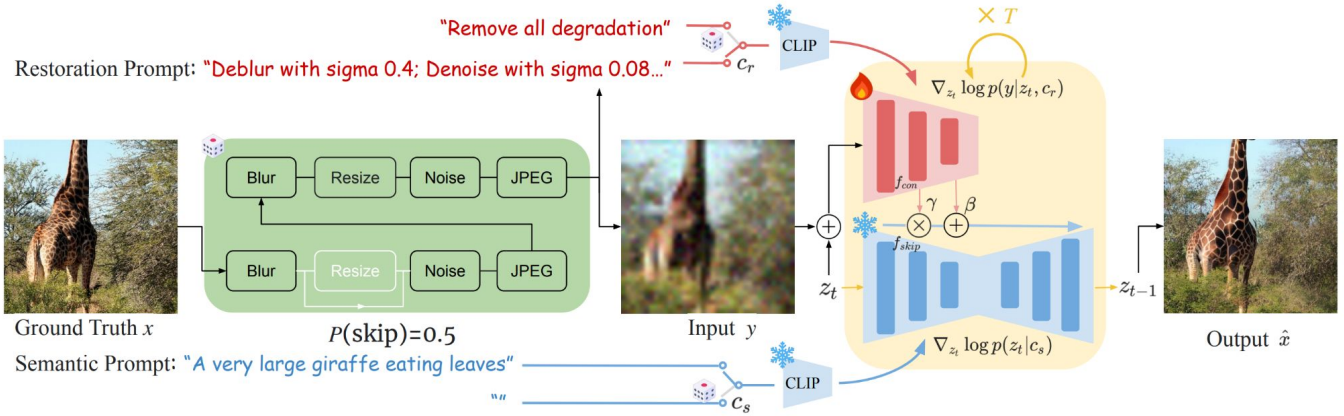
2. Offer clear guidance for ill-posed problems.

SPIRE: Semantic Prompt-Driven Image Restoration

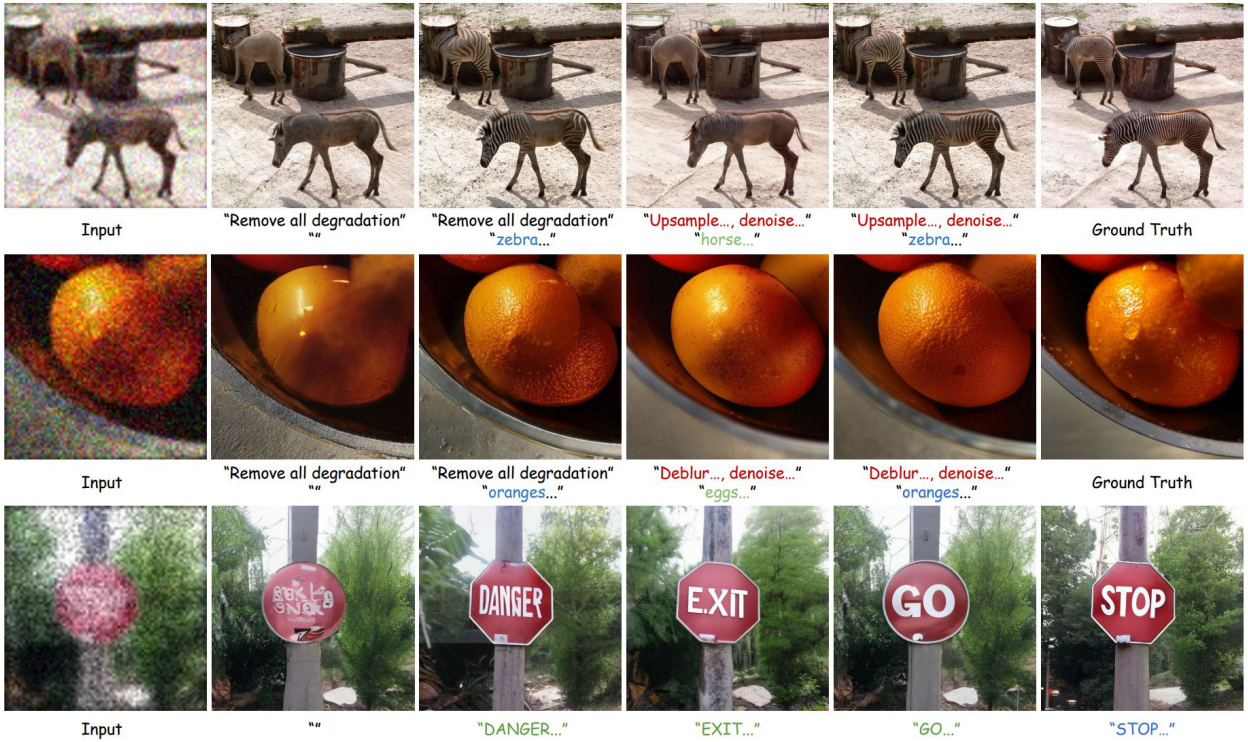
Chenyang Qi[1,2]*, Zhengzhong Tu[1], Keren Ye[1] Mauricio Delbracio[1], Peyman Milanfar[1], Qifeng Chen[2], and Hossein Talebi[1]

[1] Google Research
[2] HKUST

2024 ECCV



2. Offer clear guidance for ill-posed problems.



2. Offer clear guidance for ill-posed problems.

Scaling Up to Excellence:
Practicing Model Scaling for Photo-Realistic Image Restoration In the Wild

Fanghua Yu[1,*], Jinjin Gu[2,*], Zheyuan Li[1], Jinfan Hu[1], Xiangtao Kong[3],
Xintao Wang[4], Jingwen He[2,5], Yu Qiao[2], Chao Dong[1,2,†]
[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences [2]Shanghai AI Laboratory
[3]The Hong Kong Polytechnic University [4]ARC Lab, Tencent PCG [5]The Chinese University of Hong Kong
Project Page: https://supir.xpixel.group

2024 CVPR

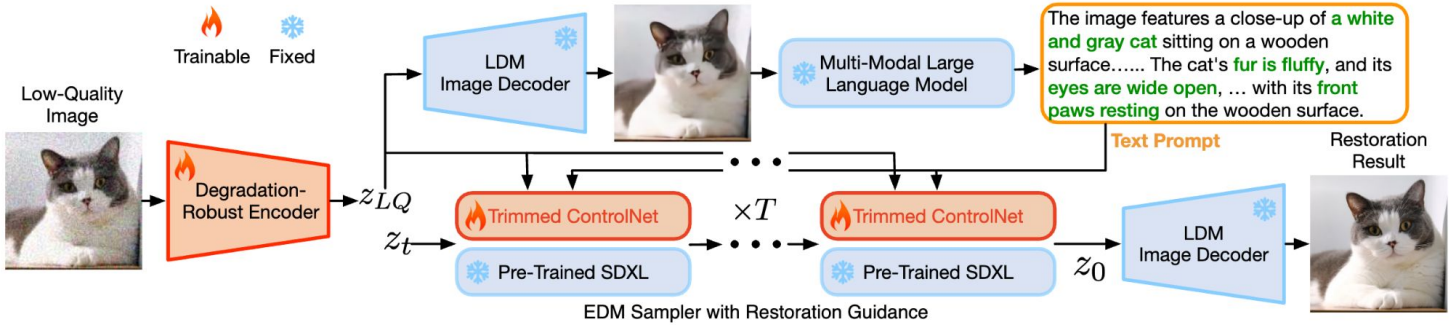
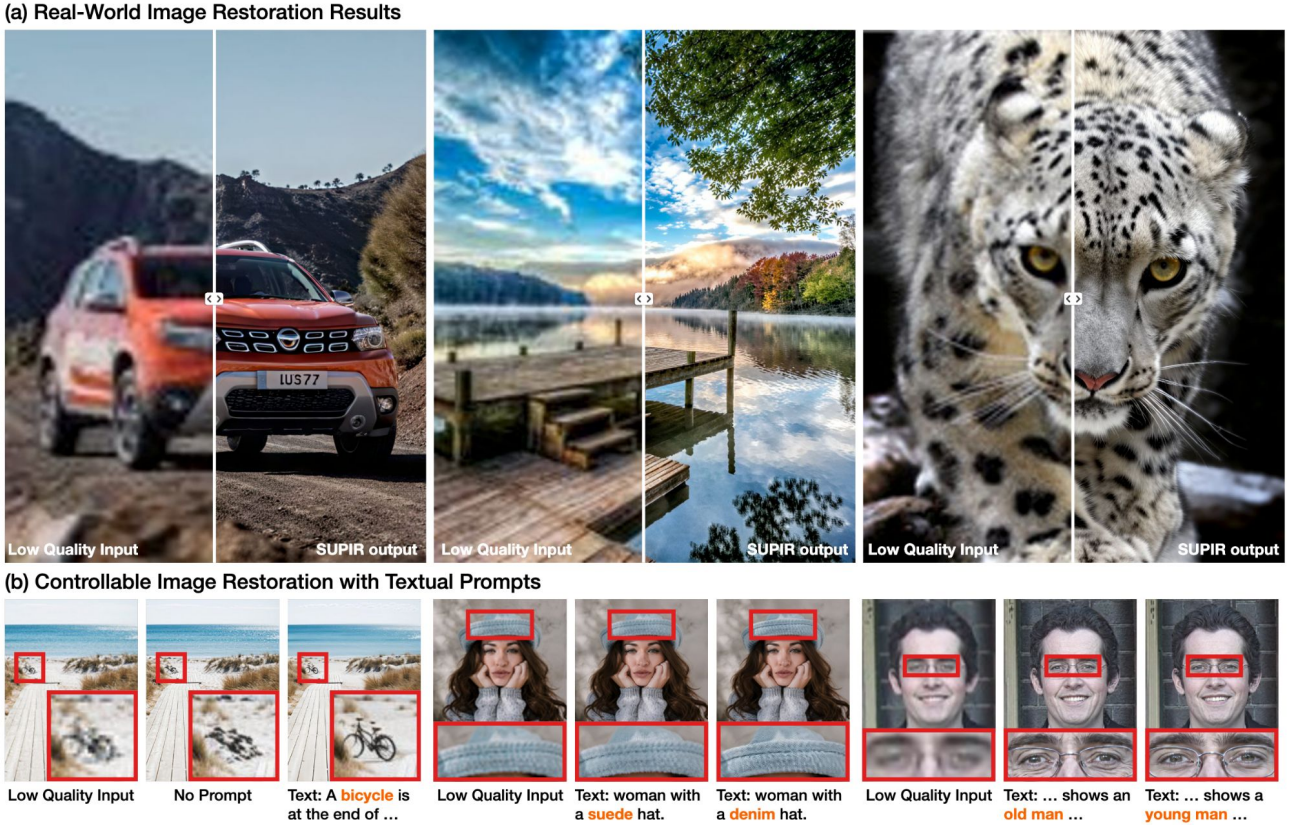


Figure 2. This figure briefly shows the workflow of the proposed SUPIR model.

2. Offer clear guidance for ill-posed problems.



Usage of text in low-level vision:

3. Serve as a simpler representation that can assist in complex image restoration tasks.

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

2024 CVPR

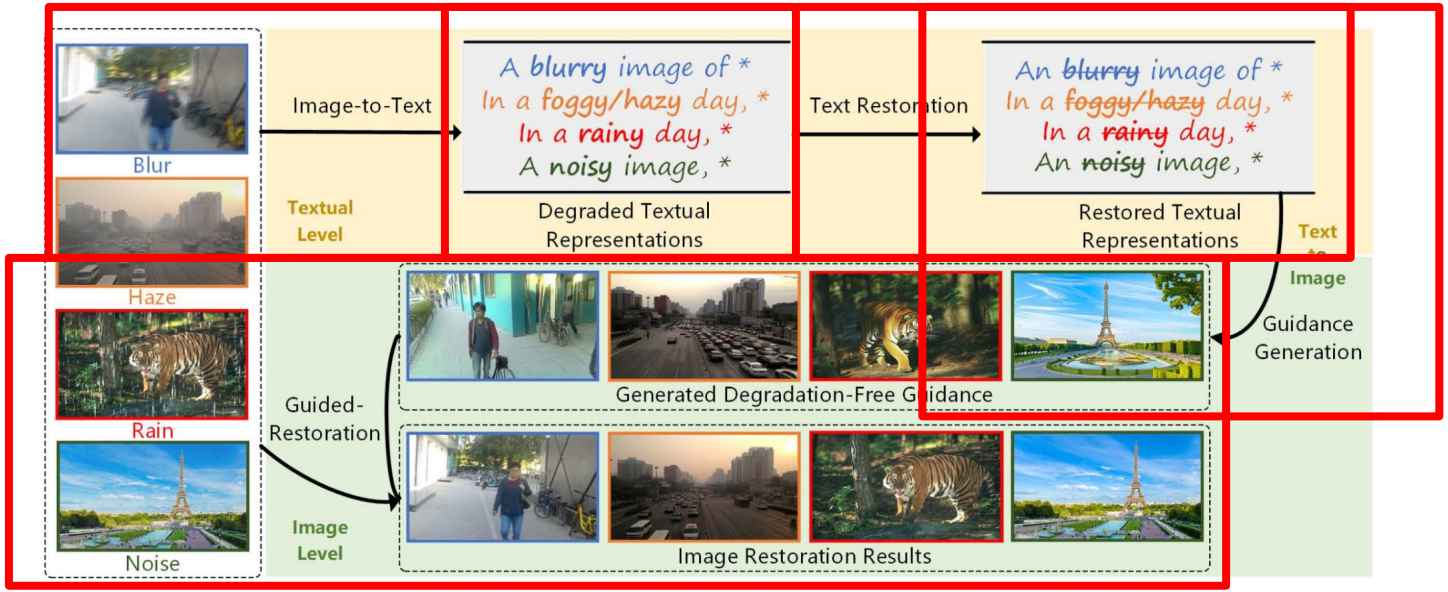
4. Leverage text's robust features

Beyond Pixels: Text Enhances Generalization in Real-World Image Restoration

Haoze Sun[1] Wenbo Li[2*] Jiayue Liu[1] Kaiwen Zhou[2] Yongqiang Chen[3]
Yong Guo[2] Yanwei Li[3] Renjing Pei[2] Long Peng[4] Yujiu Yang[1*]
[1]Tsinghua University [2]Huawei [3]CUHK [4]USTC
shz22@mails.tsinghua.edu.cn fenglinglwb@gmail.com

12.01.2024 Arxiv

Improving Image Restoration through Removing Degradations in Textual Representations



Improving Image Restoration through Removing Degradations in Textual Representations

Motivation in this paper

- text is loosly coupled with content, easy to remove degradation
  - ex. “a scene of ∗” ←→ “a rainy scene of ∗”

My opinion…

- generated “content-related clean prior” is the key

Method

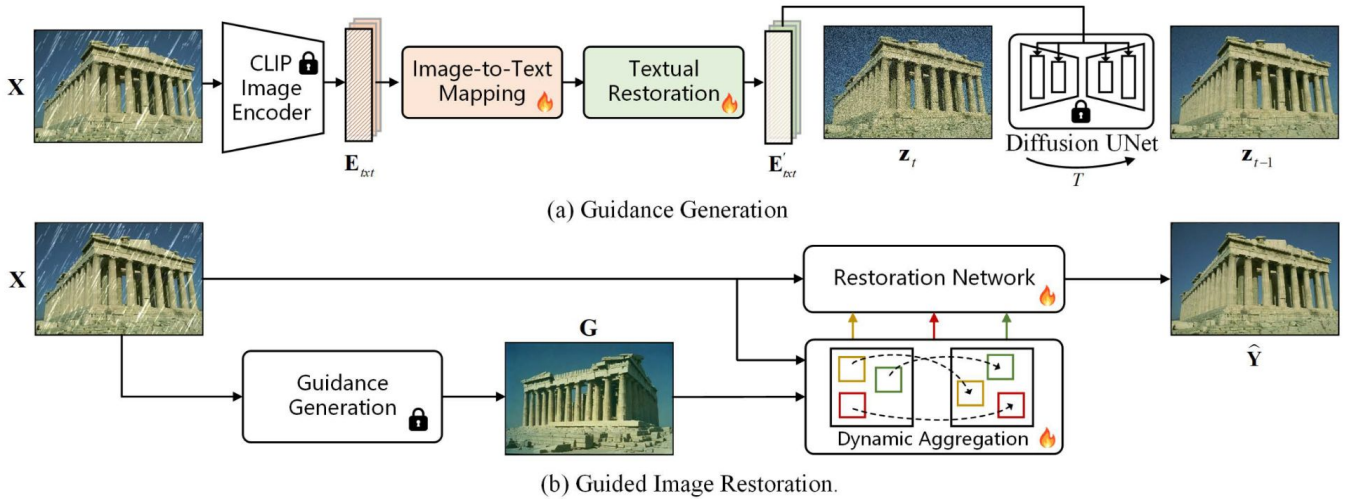


Figure 2. **Illustration of the proposed pipeline.** **(a)** We sequentially train image-to-text mapper $\mathcal{M}_{i2t}$ and textual restoration module $\mathcal{M}_{clean}$ to convert image concepts into textual representations and remove textual degradation information, respectively. **(b)** The guidance image is used to assist the image restoration process.

1. How to generate content-related, degradation-free images
2. How to guide restoration

Degradation-Free Guidance Generation

Using Stable Diffusion as the text-to-image generation model.

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

ϵ : noise, t : timestep, z : latent, τ : CLIP text encoder, p : text

Training 2 models (MLP)

$$\mathbf{E}_{txt} = \mathcal{M}_{i2t}(\tau_\theta^i(\mathbf{X})), \quad X \text{ : image, } \tau \text{ : CLIP image encoder}$$

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

Degradation-Free Guidance Generation

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

Training sequentially

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

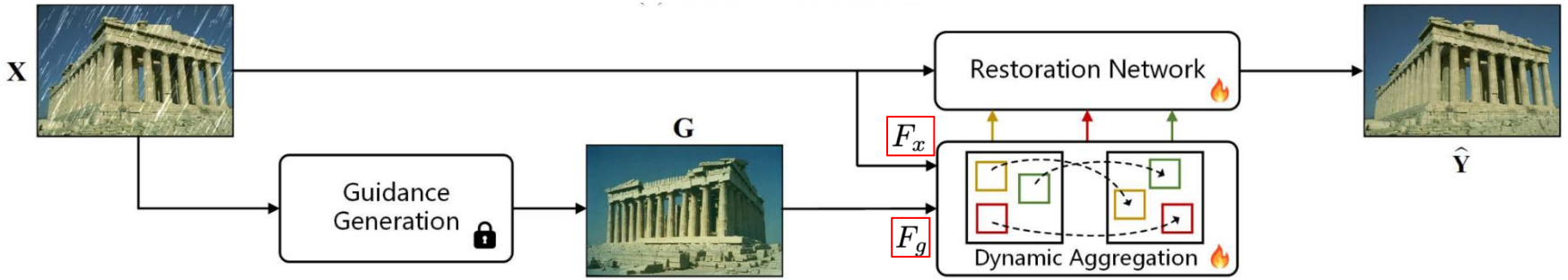
   X : degraded or clean image, z : corresponding degraded or clean image's latent vector

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

   X : degraded image, z : paired clean image's latent vector

- Restore implicite textual representations for faithful reconstruction

Guided Restoration



(b) Guided Image Restoration.

F_x : degraded image's multi-scale features, F_g : generated image's multi-scale features

search useful feature based similarity score

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

$\mathcal{B}$: one CNN-based block or transformer-based block, α : hyper-parameter

Effect of Textual Restoration

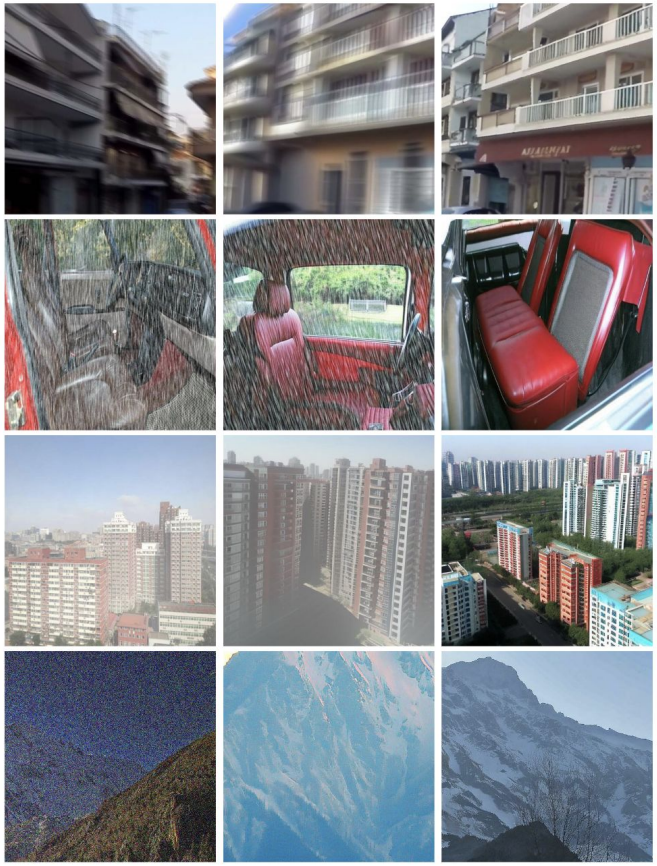
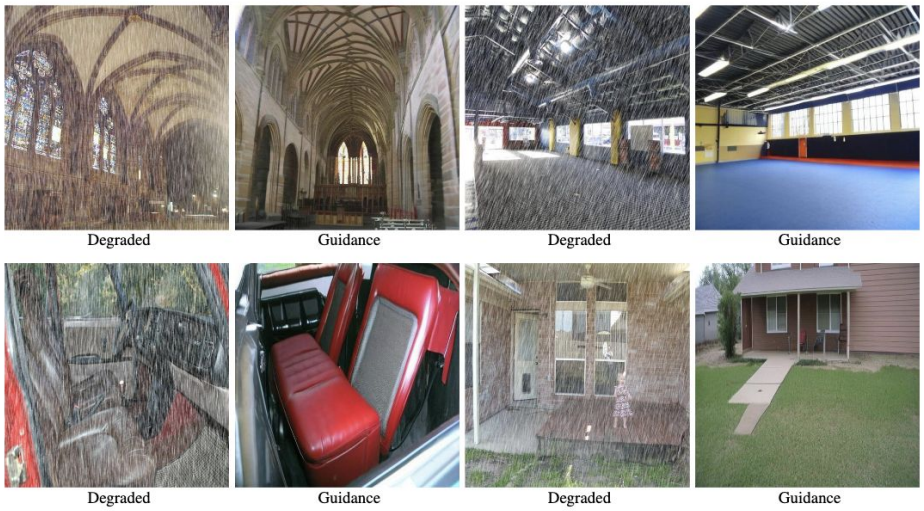


Figure A. Visual comparison of w/o. textual restoration and w/. textual restoration.

Explicit Textual Restoration v.s. Implicit Textual Restoration

Figure B. Visual comparison of synthetic guidance by explicit and implicit textual representation on image deblurring task.

Examples of generated images for guidance



Experimental results

Table 1. **All-in-one image restoration results**. Following PromptIR [73], we train and evaluate the proposed method in all-in-one image restoration task, our method outperforms PromptIR across all the benchmark datasets.

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**

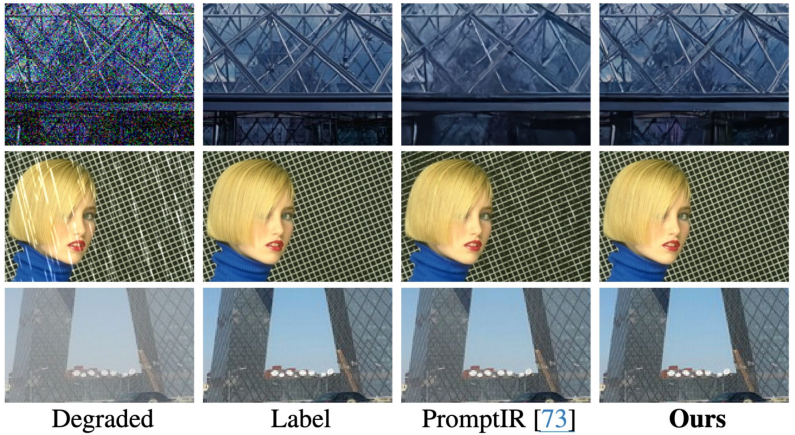


Figure 3. **All-in-one image restoration results**. Top: image denoising, mid: image deraining, bottom: image dehazing.

Experimental results

Table 2. **Motion image deblurring results**. We train models with GoPro training data. We evaluate our method on GoPro, HIDE, RealBlur benchmark datasets. PSNR and SSIM scores are calculated on RGB-channels.

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

Table 3. **Defocus image deblurring results**. We train and evaluate methods on DPDD dataset [2]. S denotes single-image defocus deblurring model. D denotes dual-pixel defocus deblurring. PSNR and SSIM scores are calculated on RGB channels.

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]	28.87	0.882	23.24	0.743	25.98	0.811
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]	29.48	0.895	23.97	0.773	26.66	0.833
Ours$_D$	**29.62**	**0.899**	**24.16**	**0.775**	**26.82**	**0.835**

Experimental results

Table 4. **Image dehazing results**. We separately train and evaluate our method indoor scene and outdoor scene. PSNR and SSIM scores are calculated on RGB-channels.

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]	41.24	0.996	40.05	0.996
Ours	**41.48**	**0.996**	**40.29**	**0.996**

Table 5. **Image deraining results**. We separately train and evaluate our method on Rain200H, Rain200L, DID-Data, and DDN-Data. PSNR and SSIM scores are calculated on Y channel in YCbCr color space.

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	0.964	34.20	0.957
IDT [100]	40.74	0.988	32.10	**0.934**	34.89	0.962	33.84	0.955
DRSformer [17]	41.21	0.989	**32.16**	0.933	35.24	0.962	34.23	0.955
Ours	**41.59**	**0.990**	31.97	0.931	**35.46**	**0.964**	**34.57**	**0.958**

Table 6. **Grayscale image denoising on Gaussian noise**. Upper-bracket: models are trained on a range of noise levels. Lower-bracket: models are trained on the fixed noise level.

Method	Set12 [118]			BSD68 [65]			Urban100 [40]		
	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
DnCNN [118]	32.67	30.35	27.18	31.62	29.16	26.23	32.28	29.80	26.35
FFDNet [120]	32.75	30.43	27.32	31.63	29.19	26.29	32.40	29.90	26.50
IRCNN [119]	32.76	30.37	27.12	31.63	29.15	26.19	32.46	29.80	26.22
DRUNet [122]	33.25	30.94	27.90	31.91	29.48	26.59	33.44	31.11	27.96
Restormer[114]	33.35	31.04	28.01	31.95	29.51	26.62	**33.67**	31.39	28.33
Ours	**33.35**	**31.30**	**28.13**	**31.98**	**29.58**	**26.77**	33.62	**31.47**	**28.46**
FOCNet [41]	33.07	30.73	27.68	31.83	29.38	26.50	33.15	30.64	27.40
MWCNN [60]	33.15	30.79	27.74	31.86	29.41	26.53	33.17	30.66	27.42
NLRN [59]	33.16	30.80	27.64	31.88	29.41	26.47	33.45	30.94	27.49
RNAN [125]	-	-	27.70	-	-	26.48	-	-	27.65
DeamNet [79]	33.19	30.81	27.74	31.91	29.44	26.54	33.37	30.85	27.53
DAGL [67]	33.28	30.93	27.81	31.93	29.46	26.51	33.79	31.39	27.97
SwinIR [57]	33.36	31.01	27.91	**31.97**	29.50	26.58	33.70	31.30	27.98
Restormer [114]	33.42	31.08	28.00	31.96	29.52	26.62	**33.79**	31.46	28.29
Ours	**33.47**	**31.15**	**28.12**	31.92	**29.67**	**26.78**	33.78	**31.58**	**28.38**

Table 7. **Color image denoising on Gaussian noise**. Upper-bracket: models are trained on a range of noise levels. Lower-bracket: models are trained on the fixed noise level. PSNR is calculated on RGB channels.

Method	CBSD68 [66]			Kodak24			McMaster [123]			Urban100 [40]		
	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
IRCNN [119]	33.86	31.16	27.86	34.69	32.18	28.93	34.58	32.18	28.91	33.78	31.20	27.70
FFDNet [120]	33.87	31.21	27.96	34.63	32.13	28.98	34.66	32.35	29.18	33.83	31.40	28.05
DnCNN [118]	33.90	31.24	27.95	34.60	32.14	28.95	33.45	31.52	28.62	32.98	30.81	27.59
DSNet [72]	33.91	31.28	28.05	34.63	32.16	29.05	34.67	32.40	29.28	-	-	-
DRUNet [122]	34.30	31.69	28.51	35.31	32.89	29.86	35.40	33.14	30.08	34.81	32.60	29.61
Restormer [114]	**34.39**	31.78	28.59	35.44	33.02	30.00	35.55	33.31	30.29	**35.06**	32.91	30.02
Ours	34.37	**31.87**	**28.68**	**35.52**	**33.13**	**30.15**	**35.62**	**33.38**	**30.40**	35.03	**32.97**	**30.19**
RPCNN [99]	-	31.24	28.06	-	32.34	29.25	-	32.33	29.33	-	31.81	28.62
BRDNet [91]	34.10	31.43	28.16	34.88	32.41	29.22	35.08	32.75	29.52	34.42	31.99	28.56
RNAN [125]	-	-	28.27	-	-	29.58	-	-	29.72	-	-	29.08
RDN [126]	-	-	28.31	-	-	29.66	-	-	-	-	-	29.38
IPT [13]	-	-	28.39	-	-	29.64	-	-	29.98	-	-	29.71
SwinIR [57]	34.42	31.78	28.56	35.34	32.89	29.79	35.61	33.20	30.22	35.13	32.90	29.82
Restormer [114]	34.40	31.79	28.60	35.47	33.04	30.01	35.61	33.34	30.30	**35.13**	32.96	30.02
Ours	**34.48**	**31.97**	**28.83**	**35.58**	**33.21**	**30.23**	**35.75**	**33.56**	**30.46**	35.11	**33.13**	**30.27**

Ablation Studies (for deblurring task)

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

N : text descriptions' length

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

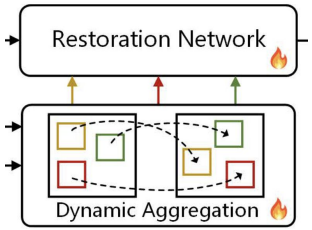
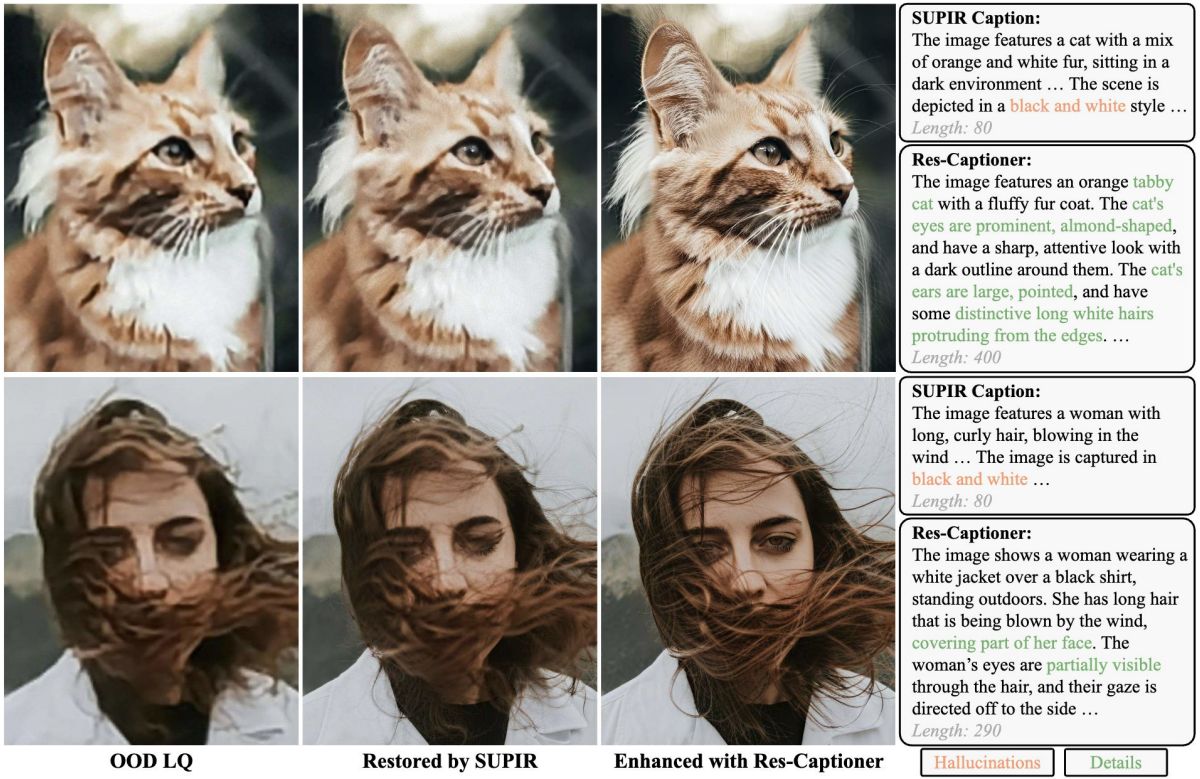


Table 11. Effect of generated guidance.

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

Degra : guided restoration by degraded input

Beyond Pixels: Text Enhances Generalization in Real-World Image Restoration



Motivation - domain invariant feature

$$\boldsymbol{x} = \mathcal{R}(\boldsymbol{x}_{lq}) \qquad x : \text{high-quality image}, x_{lq} : \text{low-quality image}$$

- Need cross-domain invariant feature **z**

$$\boldsymbol{z} = \mathcal{G}(\boldsymbol{x}_{lq}) \quad \boldsymbol{x} = \mathcal{H}(\boldsymbol{z})$$

- Propose content-related image caption as domain invariant feature

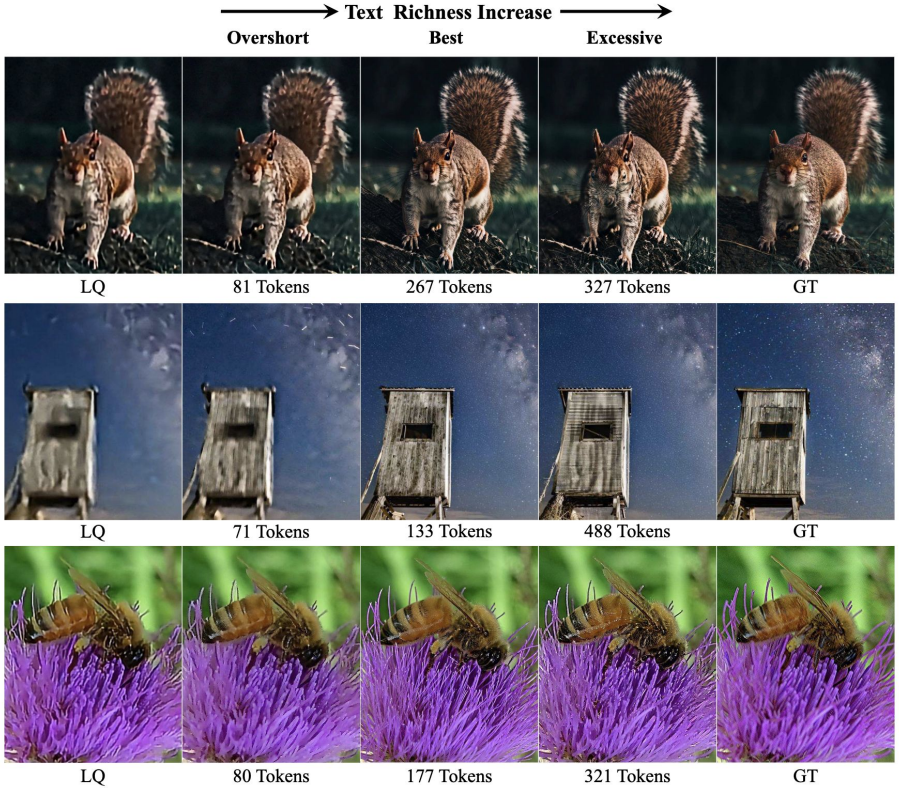
$$\boldsymbol{y} = \mathcal{C}(\boldsymbol{x}_{lq}) \quad \mathcal{C} : \text{image captioner}$$

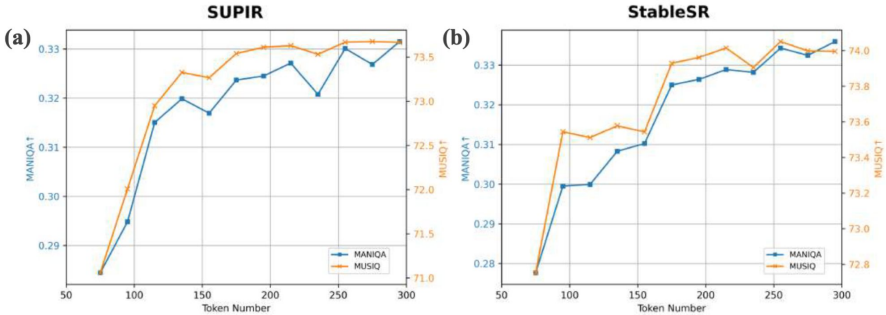
$$\boldsymbol{y}_{cont} = \{w \mid w \in \boldsymbol{y}, w \notin \boldsymbol{y}_{deg}\}$$

y_{deg} : degardation related caption, y_{cont} : content related caption

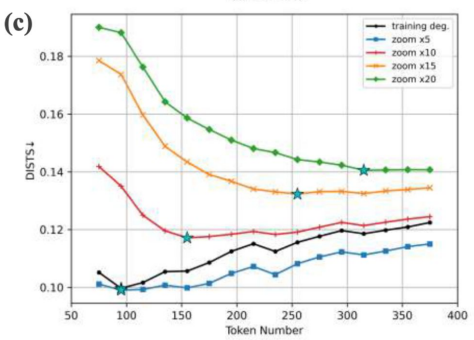
$$\boldsymbol{x} = \mathcal{R}(\boldsymbol{x}_{lq}, \boldsymbol{y}_{cont})$$

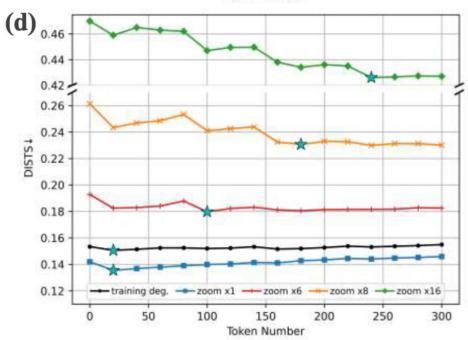
Observation 1. The richness of restored textures and details increases proportionally with the text richness.





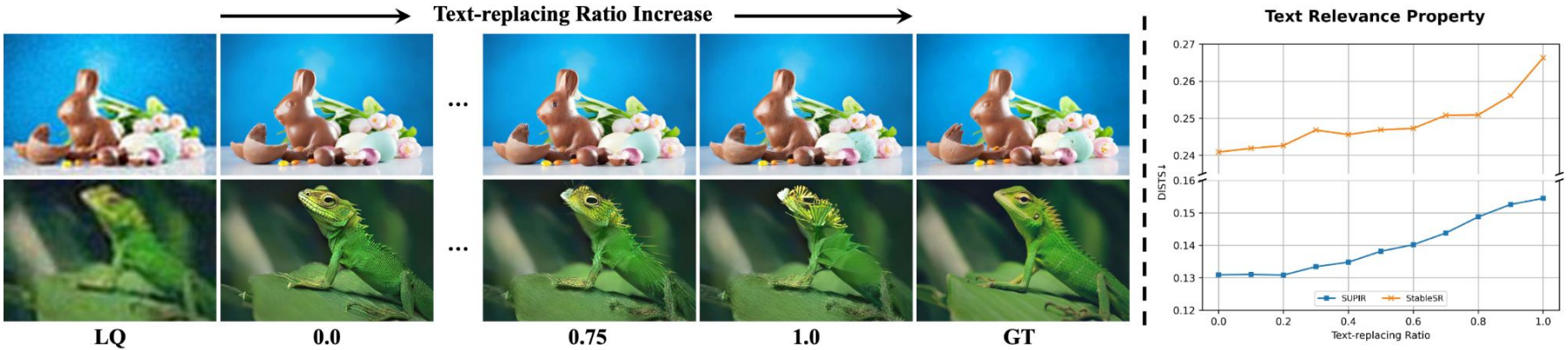
Observation 2. The optimal level of text richness is influenced by degradation severity and image content.





Other observations

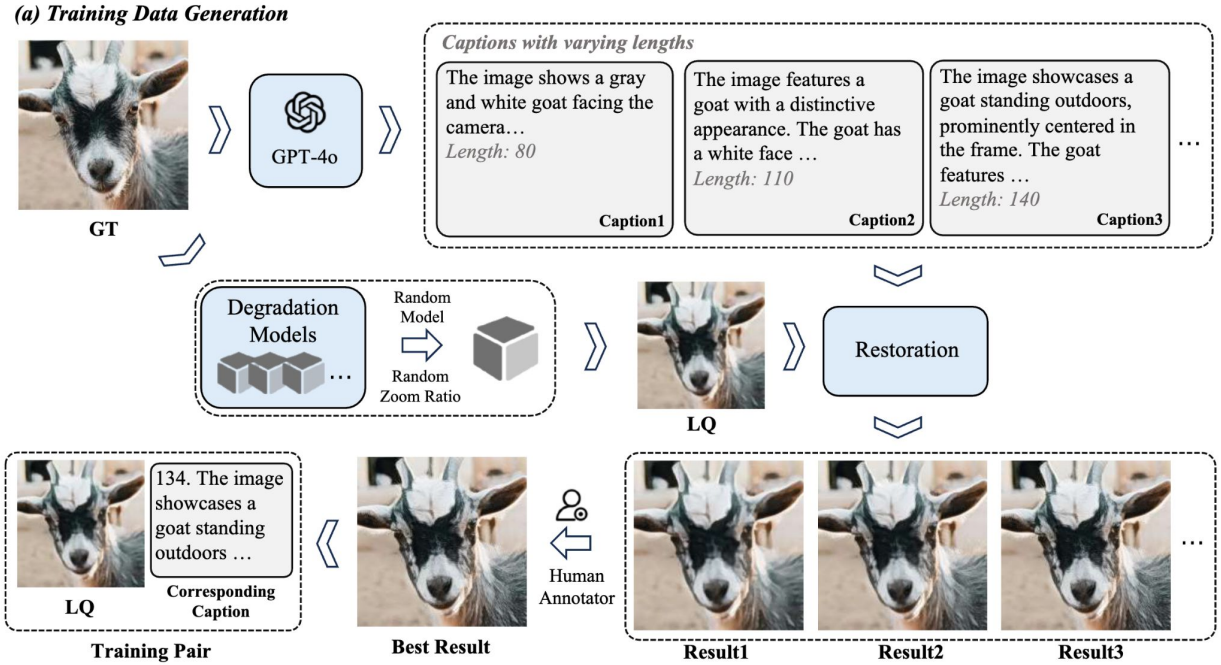
- Observation 3. The fidelity of restored textures improves incorrelation with the relevance of the text description.



- Observation 4. Descriptions related to degradation or photography can lead to blurring in the restored images.

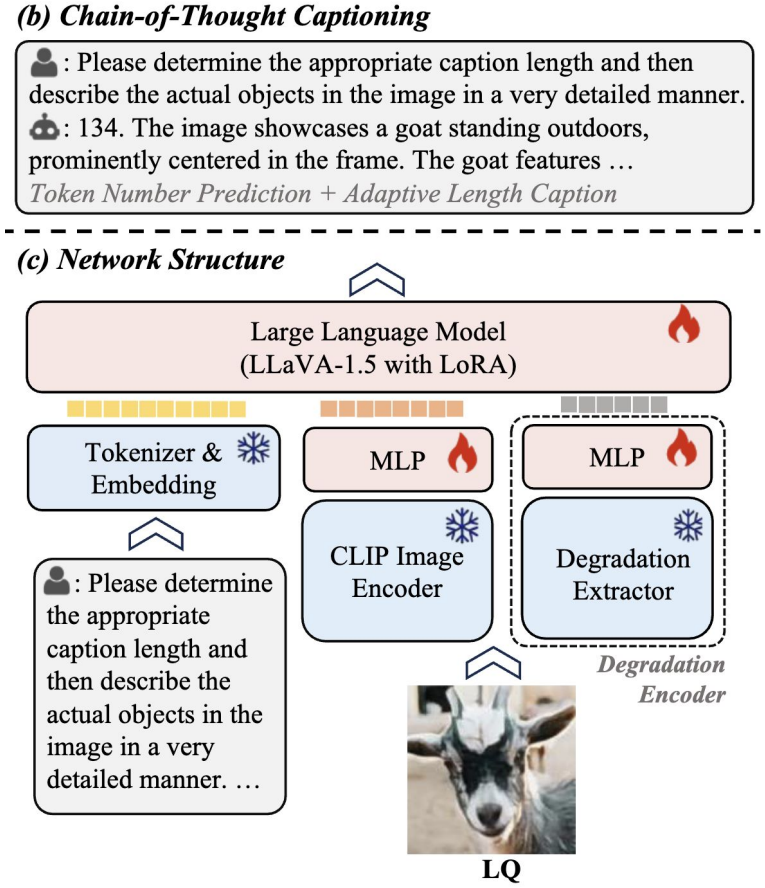
Method

- Goal : Training “Res-captioner” that can generate text descriptions of an appropriate length and accuracy, serving as a caption that supports robust restoration.



5,500 LQ image-caption pairs for training Res-Captioner.

Method


(b) Chain-of-Thought Captioning
: Please determine the appropriate caption length and then describe the actual objects in the image in a very detailed manner.
: 134. The image showcases a goat standing outdoors, prominently centered in the frame. The goat features …
Token Number Prediction + Adaptive Length Caption
(c) Network Structure
Large Language Model (LLaVA-1.5 with LoRA)
Tokenizer & Embedding
MLP
MLP
CLIP Image Encoder
Degradation Extractor
Degradation Encoder
: Please determine the appropriate caption length and then describe the actual objects in the image in a very detailed manner. …
LQ

Experimental results

Methods	RealIR (Cameras)					RealIR (Internet)				
	MUSIQ↑	MANIQA↑	LIQE↑	NIQE↓	CLIP-IQA↑	MUSIQ↑	MANIQA↑	LIQE↑	NIQE↓	CLIP-IQA↑
Real-ESRGAN+ [58]	58.54	0.1784	2.425	5.049	0.4900	58.34	0.2048	2.157	5.646	0.4458
DASR [31]	53.82	0.1487	2.208	6.038	0.4045	50.84	0.1397	1.594	6.748	0.3290
CoSeR [50]	56.91	0.1163	2.597	4.766	0.4789	66.67	0.1842	3.822	4.042	0.5831
SeeSR [62]	70.19	0.2138	3.768	3.705	0.6401	72.65	0.2694	4.243	3.749	0.6706
StableSR [55]	66.15	0.1924	3.466	4.208	0.6345	67.66	0.2012	3.913	4.033	0.6400
StableSR w/ Ours	69.28	0.2389	3.693	3.891	**0.6956**	71.64	0.2690	4.279	3.784	**0.7031**
SUPIR [68]	60.43	0.1651	2.983	4.213	0.4793	71.94	0.2727	4.425	3.492	0.6362
SUPIR w/ Ours	**71.38**	**0.2543**	**4.056**	**3.454**	0.6235	**73.26**	**0.3055**	**4.578**	**3.389**	0.6749

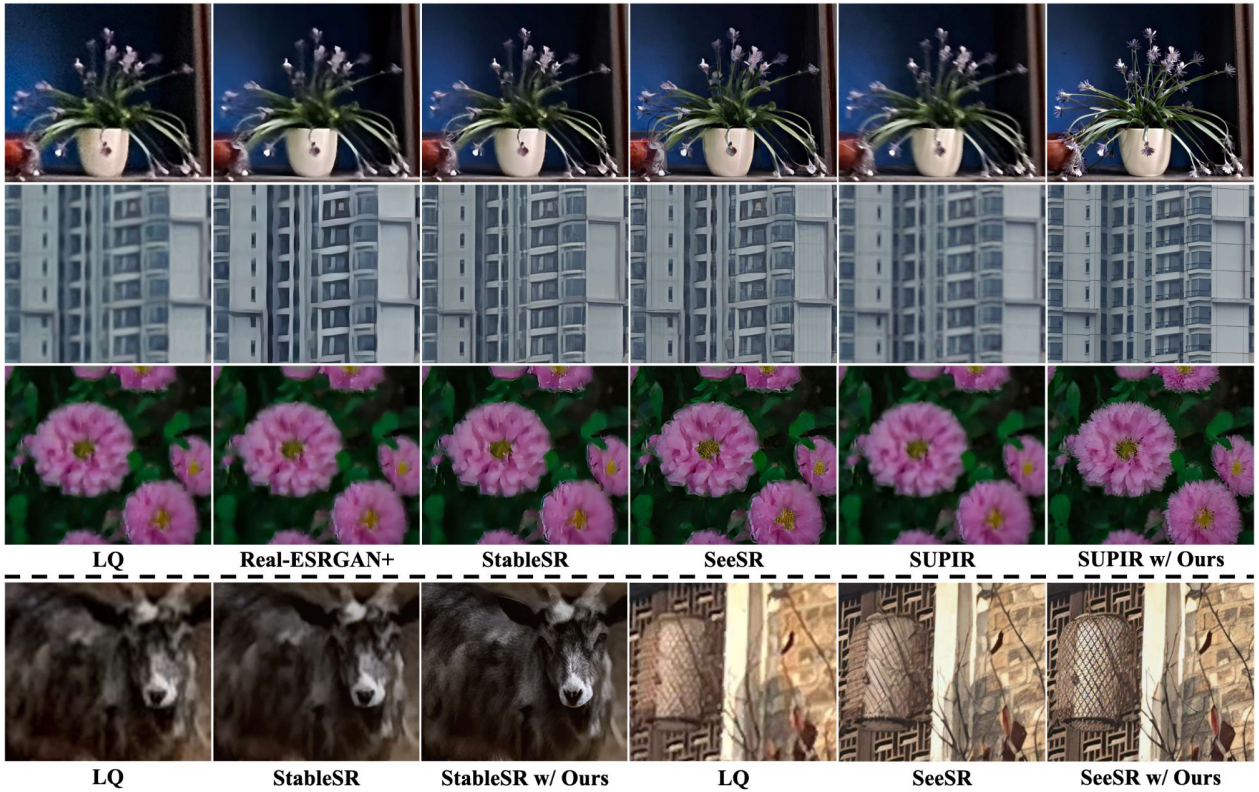
Table 1. Quantitative comparisons on our RealIR benchmark. We highlight **best** values and results of Res-Captioner-enhanced models.

Experimental results

Methods	Light Degradation				Moderate Degradation				Heavy Degradation			
	DISTS↓	LPIPS↓	MANIQA↑	LIQE↑	DISTS↓	LPIPS↓	MANIQA↑	LIQE↑	DISTS↓	LPIPS↓	MANIQA↑	LIQE↑
StableSR	0.1791	0.3311	0.2256	3.699	0.1864	0.3209	0.2297	3.603	0.2181	0.4008	0.1676	3.047
StableSR w/ Ours	**0.1748**	**0.3271**	**0.2712**	**3.733**	**0.1774**	**0.3121**	**0.2614**	**3.872**	**0.1993**	**0.3883**	**0.2298**	**3.502**
	2.4%	1.2%	20.2%	0.9%	4.8%	2.7%	13.8%	7.5%	8.6%	3.1%	37.1%	14.9%
SUPIR	0.1821	0.3444	0.2042	3.148	0.1883	0.3473	0.2182	3.349	0.2159	0.4106	0.1749	2.840
SUPIR w/ Ours	**0.1680**	**0.3178**	**0.3065**	**4.011**	**0.1621**	**0.3052**	**0.3294**	**4.226**	**0.1873**	**0.3754**	**0.3033**	**3.991**
	7.7%	7.7%	50.0%	27.4%	13.9%	12.1%	51.0%	26.2%	13.3%	8.6%	73.4%	40.5%

Table 2. Quantitative comparisons between the official model and the Res-Captioner-enhanced model under different degradation levels. We show the improvement percentage on each metric.

Experimental results



Experimental results

Method	Light Degradation		Moderate Degradation		Heavy Degradation	
	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓
Ours	**0.1680**	**0.3178**	**0.1621**	**0.3052**	0.1873	**0.3754**
w/ Min Len.	0.1718	0.3274	0.1753	0.3252	0.2033	0.4009
w/ Max Len.	0.1864	0.3525	0.1770	0.3184	0.1964	0.4039
w/ Low Rel.	0.1738	0.3389	0.1655	0.3061	0.1907	0.3914
w/ Harmful Des.	0.1686	0.3191	0.1678	0.3178	**0.1868**	0.3883

Table 4. Ablation studies on text richness, relevance, and harmful descriptions. We highlight **best** values for each metric.

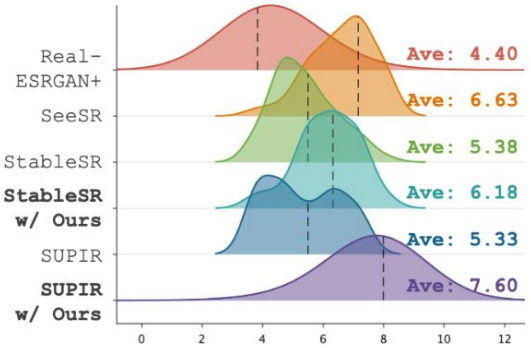


Figure 7. User study.

Experimental results

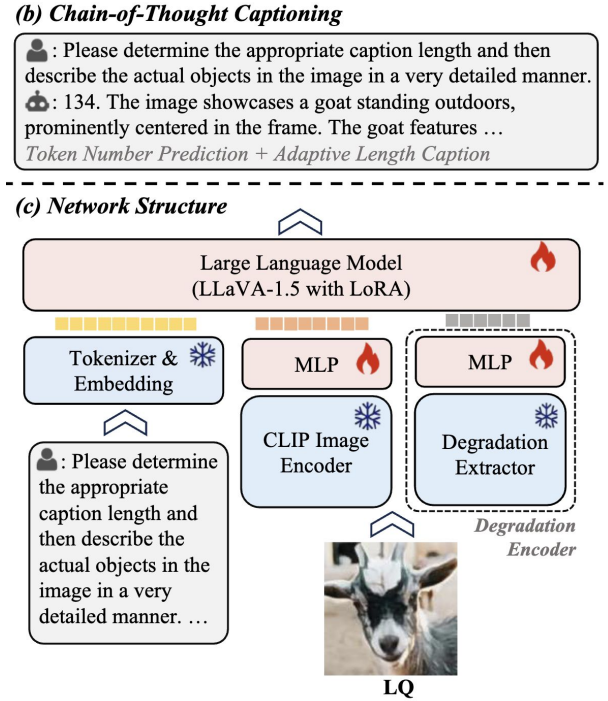
CoT captioning and degradation-aware visual encoder.

E : Offset level,
L0 : optimal length annotated by human,
L : output length of Res-captioner (using RealIR dataset)

$$E = \max\left(|L_o - L| - 15, 0\right)/30$$

E = 1.27 for RealIR dataset (Out-of-distirbution samples)

- 66.7% increase without CoT captioning
- 31.5% increase without degradation aware visual encoder



Effect of Harmful Description

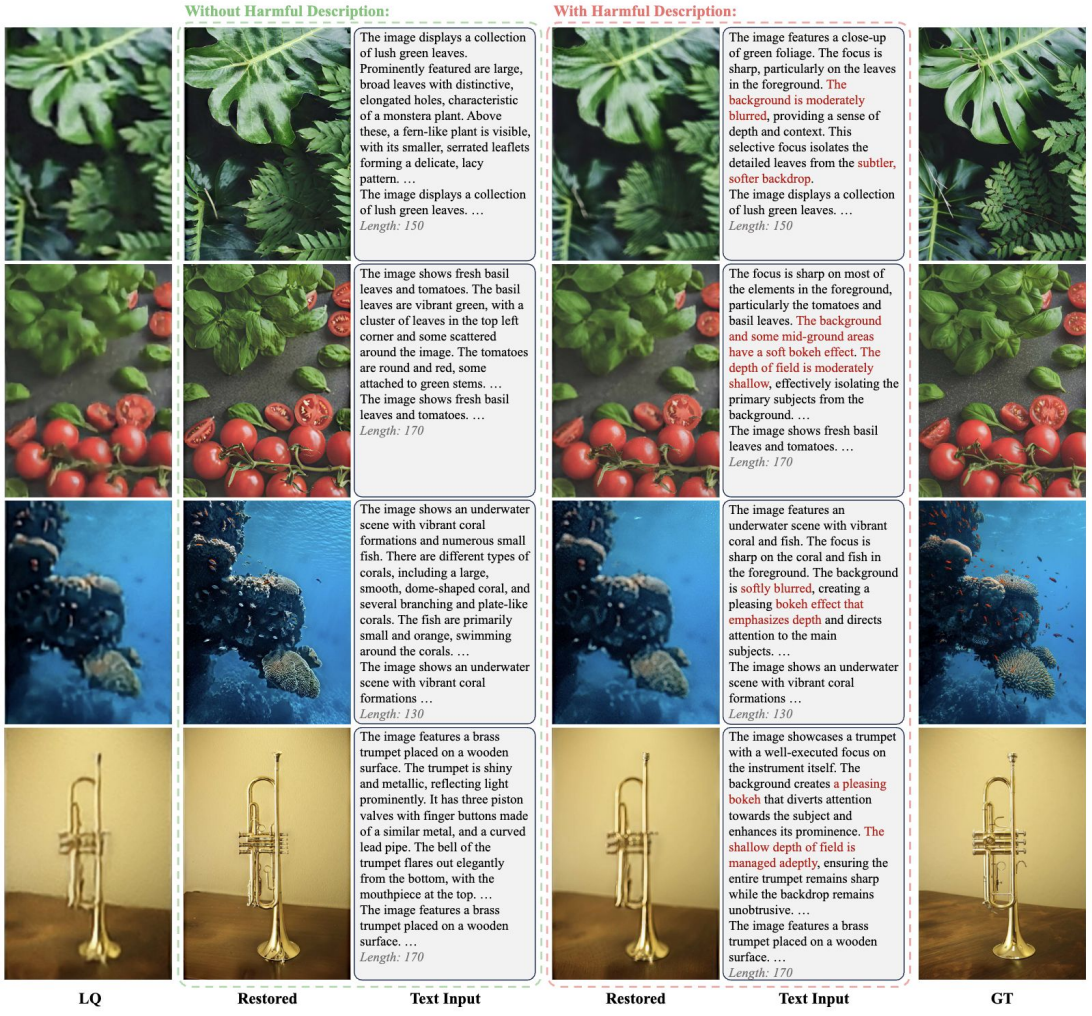


Figure A.8. Harmful descriptions to the image restoration.

Qualitative results

(a) Additional qualitative comparisons of Res-Captioner applied to StableSR on in-the-wild images.

(b) Qualitative comparisons of Res-Captioner on de-hazing and de-snowing.

Conclusion

- Low-level vision can also benefit from LLM or VLM improvement

- Importance of methods for achieving scalable performance increases with LLM or VLM

Thank you!