CONTROLLING VISION-LANGUAGE MODELS FOR MULTI-TASK IMAGE RESTORATION

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, Thomas B. Schön
Department of Information Technology, Uppsala University
{ziwei.luo,fredrik.gustafsson,zheng.zhao}@it.uu.se
{jens.sjolund,thomas.schon}@it.uu.se

ICLR 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

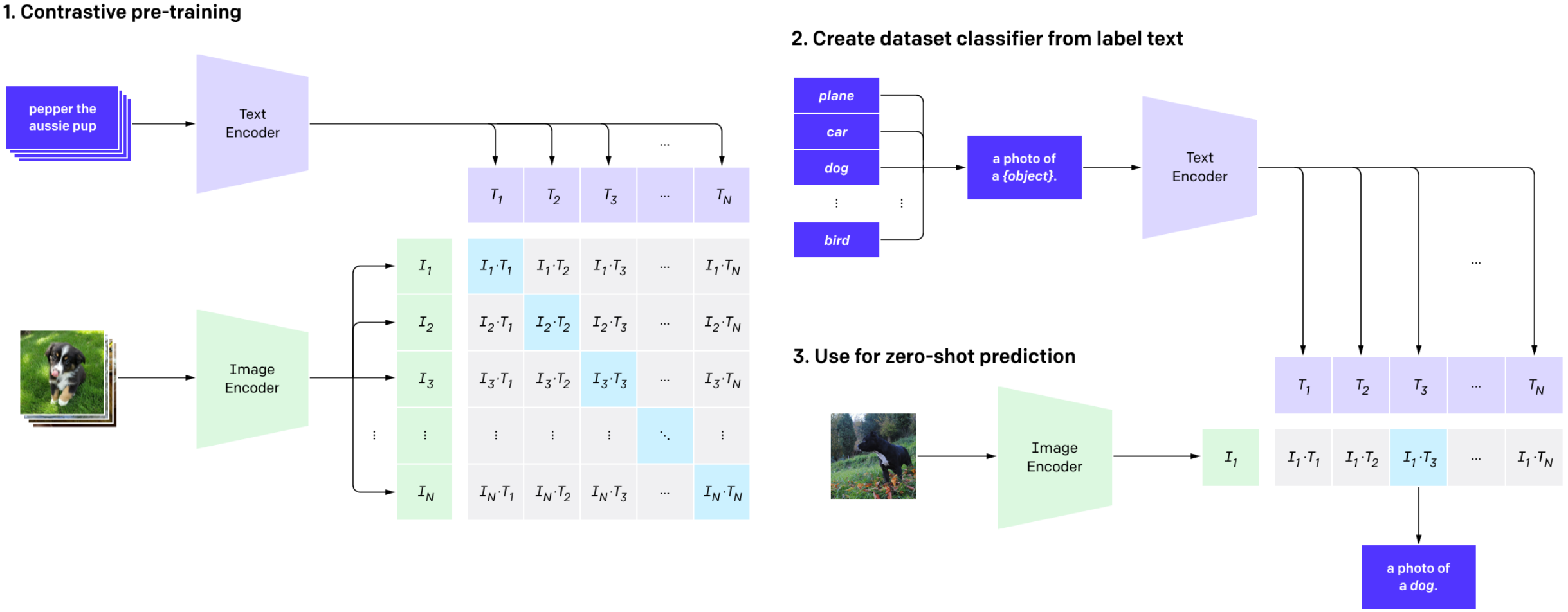
Outline

- Preliminary
- Framework
- Method
- Experiment
- Conclusion

Outline

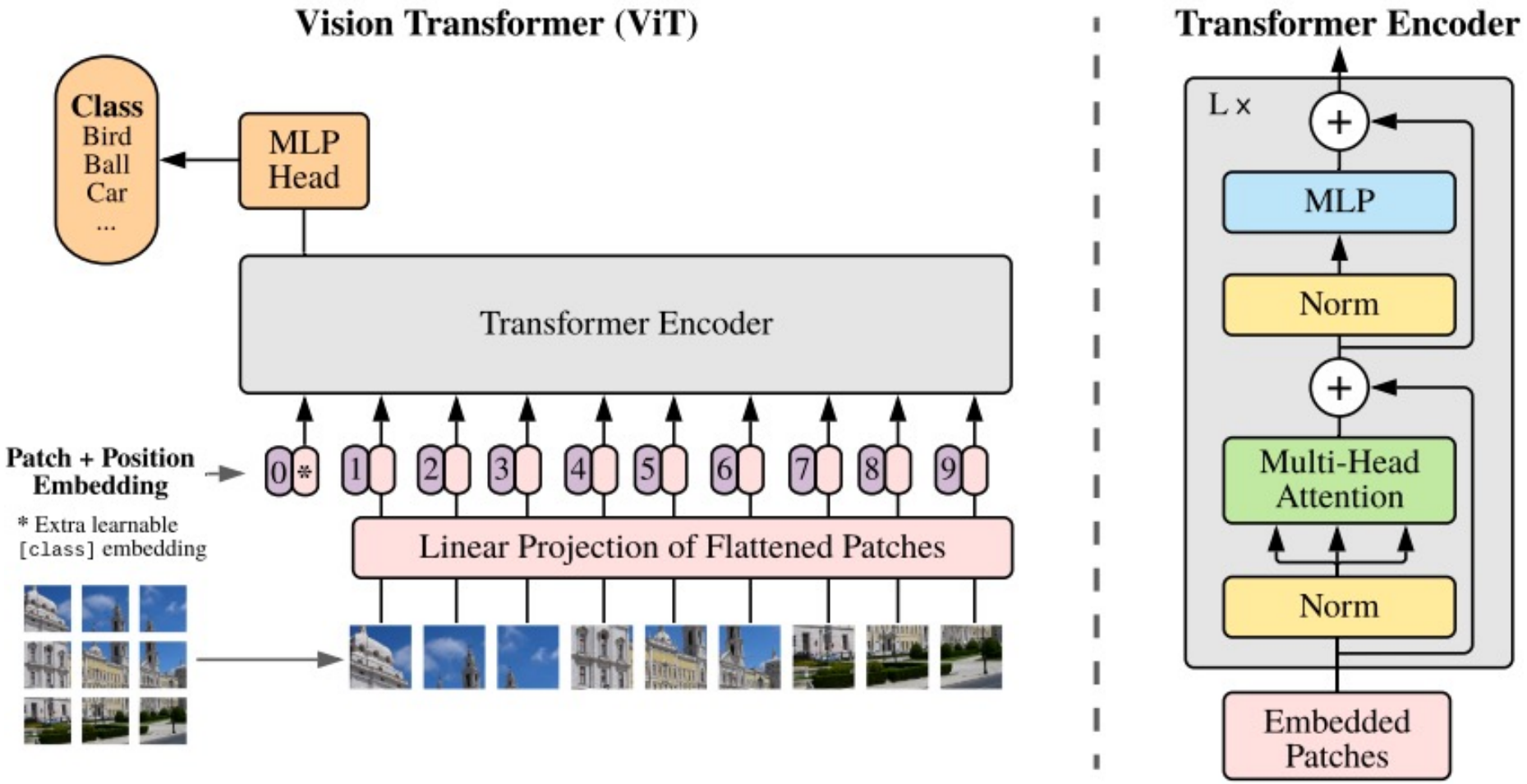
- Preliminary
- Framework
- Method
- Experiment
- Conclusion

CLIP



Learning Transferable Visual Models From Natural Language Supervision, ICML 2021

Vision Transformer



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021

Stochastic Differential Equations

$$dX_t = \mu(X_t, t)\, dt + \sigma(X_t, t)\, dW_t$$

$$dx_t = \left[\mu(t, x_t) - \sigma(t)^2 \nabla_{x_t} \log p_t(x_t)\right] dt + \sigma(t)\, d\bar{W}_t$$

IR-SDE

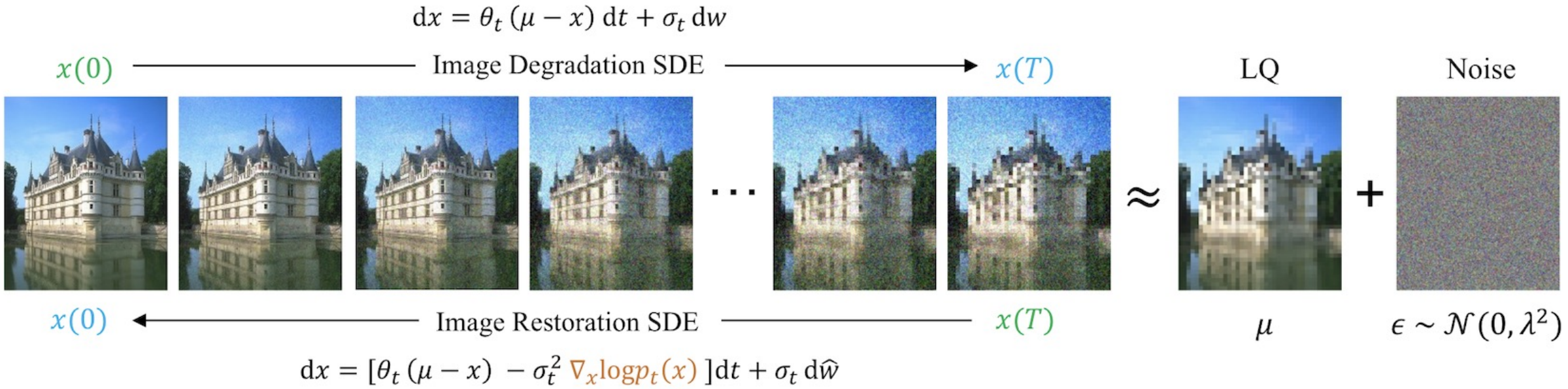


Image Restoration with Mean-Reverting Stochastic Differential Equations, ICML 2023

Outline

- Preliminary
- Framework
- Method
- Experiment
- Conclusion

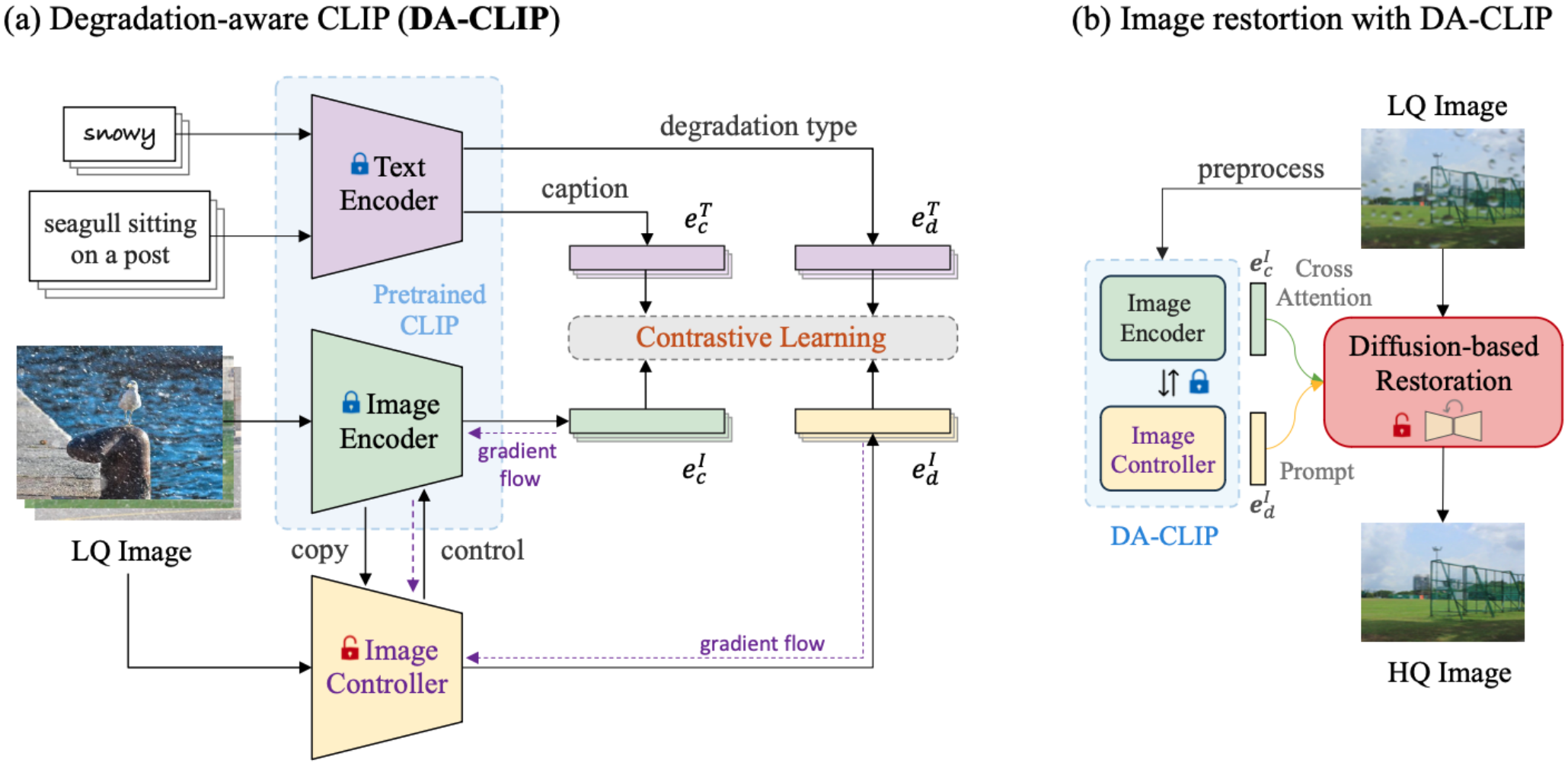
Introduction

- Present a degradation-aware vision-language model (**DA-CLIP**) to better transfer pretrained **vision-language models** to **low-level vision tasks** as a multi-task framework for image restoration.

- Trains an additional controller predict **HQ content embeddings** and **degradation embedding**. By integrating the embedding into an image restoration network.

- Advances **SOTA** performance on both degradation-specific and unified image restoration tasks, showing a promising direction of **prompting image restoration with large-scale pretrained vision-language models**.

Framework



e_c^T: Caption Embedding from Text-encoder

e_d^T: Degradation Embedding from Text-encoder

e_c^I: Content Embedding from Image-encoder

e_d^I: Degradation Embedding from Image-controller

Outline

- Preliminary
- Framework
- Method
- Experiment
- Conclusion

Image controller

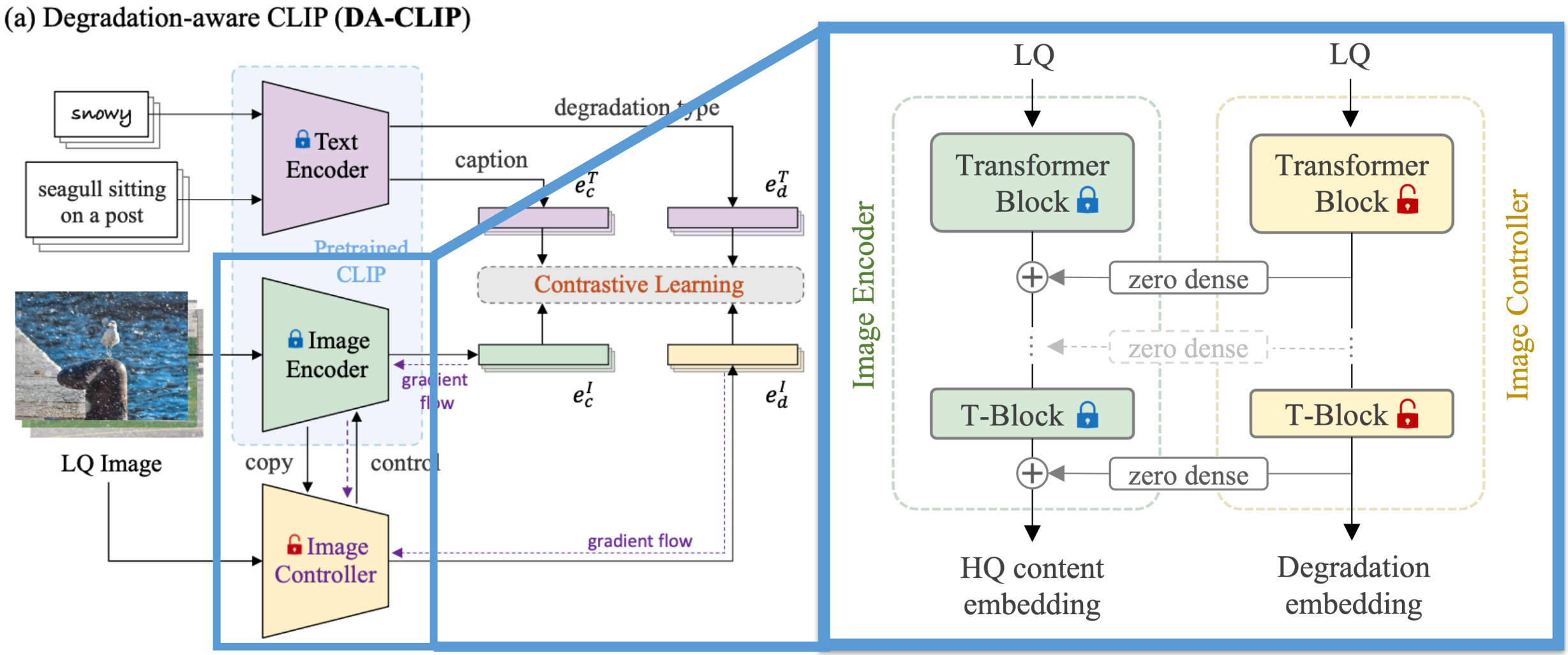


Image controller

$$\mathcal{L}_{\text{con}}(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(\boldsymbol{x}_i^{\intercal}\boldsymbol{y}_i/\tau\right)}{\sum_{j=1}^{N}\exp(\boldsymbol{x}_i^{\intercal}\boldsymbol{y}_j/\tau)}\right),$$

$$\mathcal{L}_c(\omega) = \mathcal{L}_{\text{con}}(\boldsymbol{e}_c^I, \boldsymbol{e}_c^T; \omega) + \mathcal{L}_{\text{con}}(\boldsymbol{e}_d^I, \boldsymbol{e}_d^T; \omega),$$

- Learning to align these embeddings enables DA-CLIP to predict real degradation types and HQ content features for corrupted image inputs.

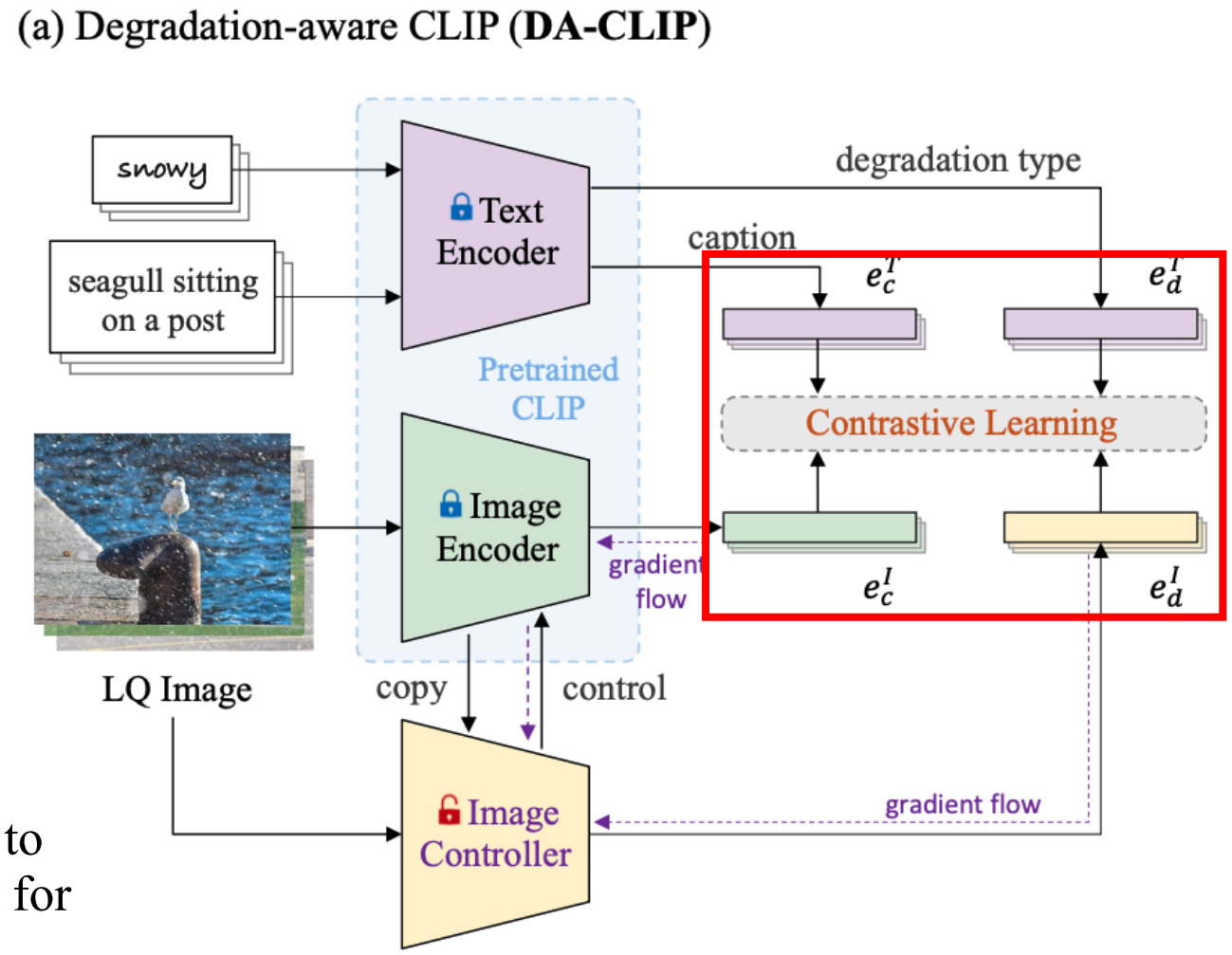
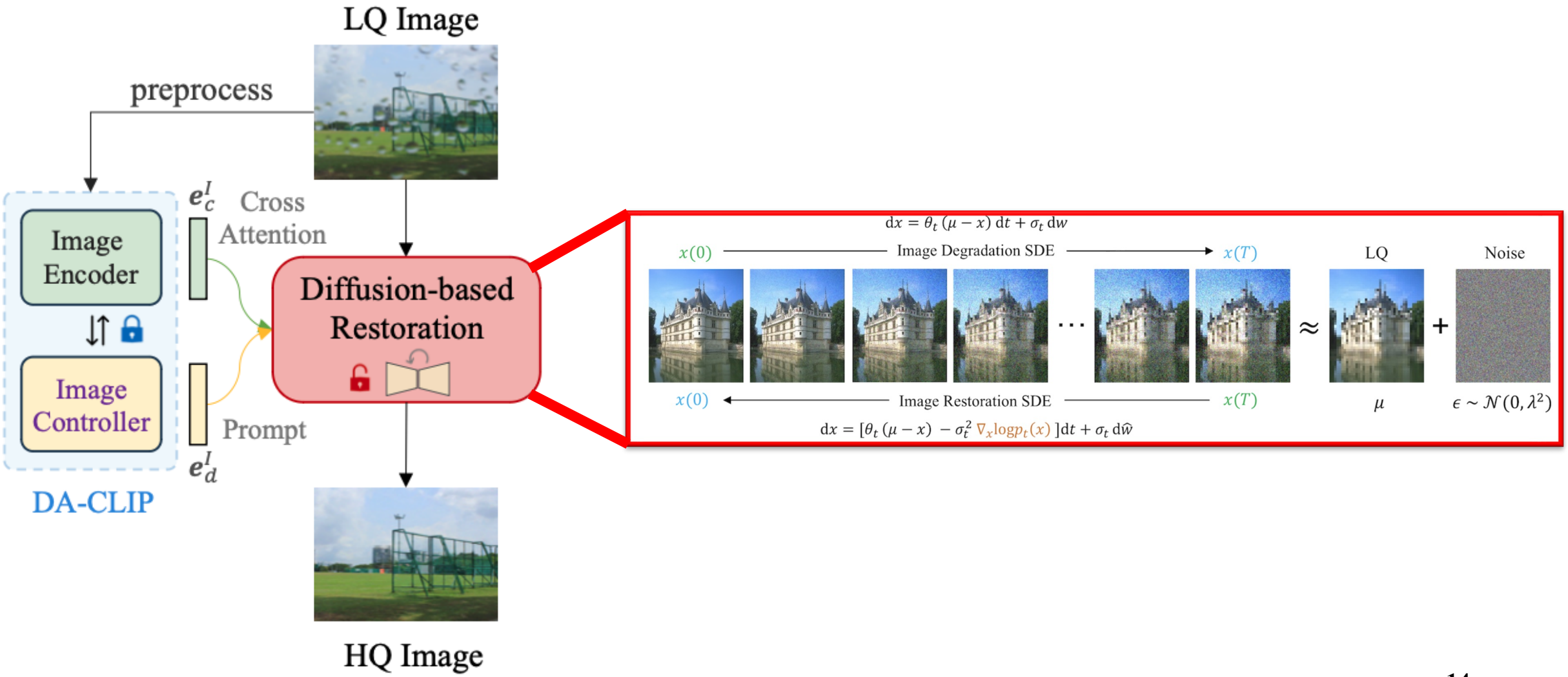
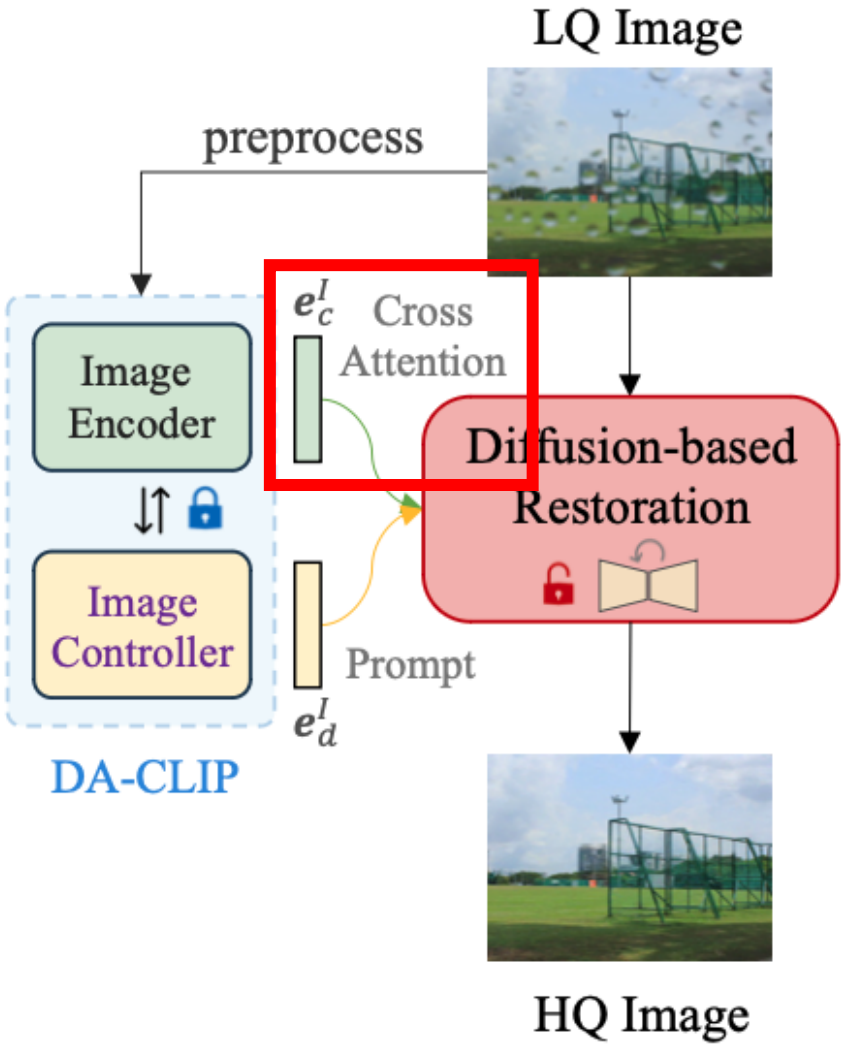
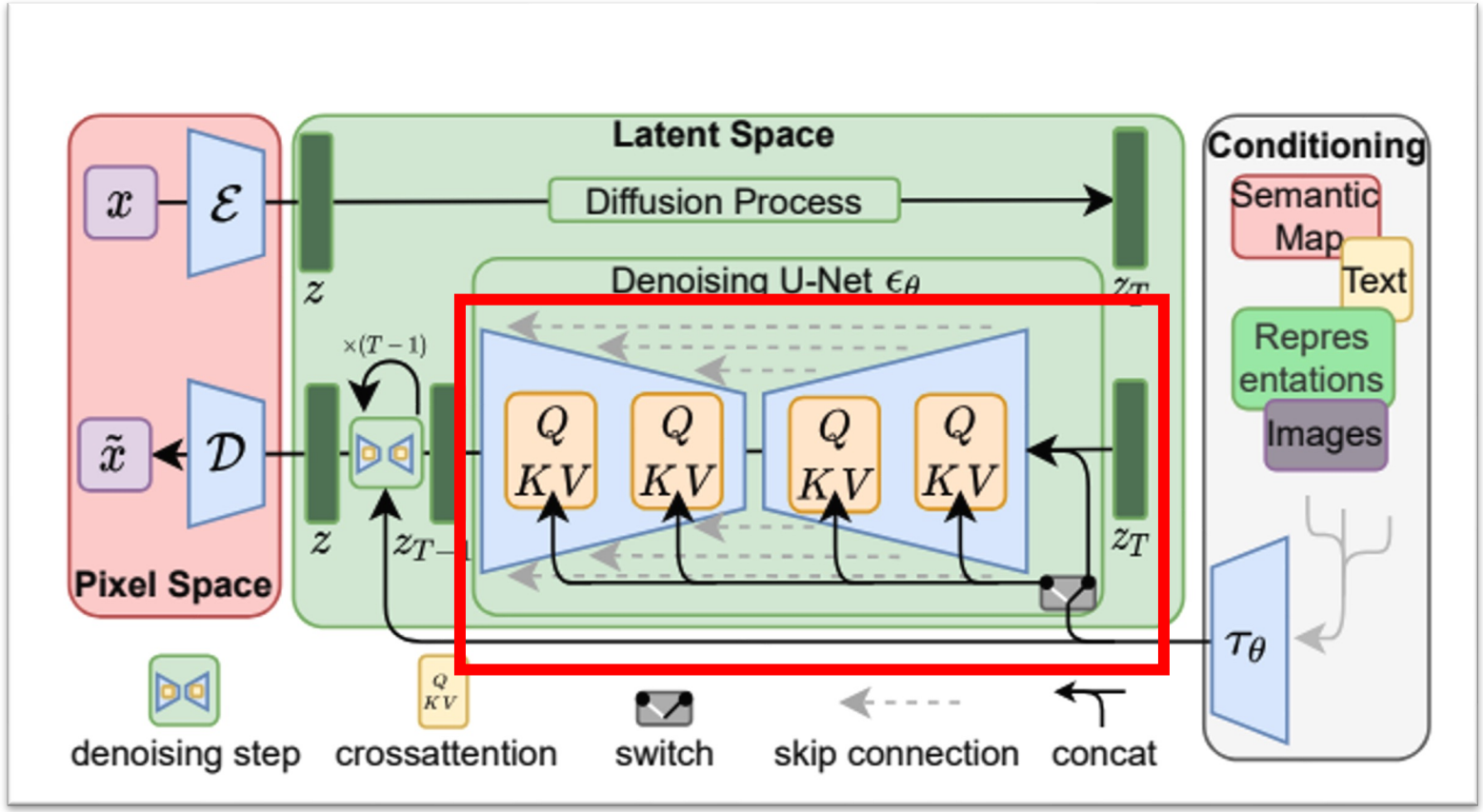


Image Restoration With DA-CLIP



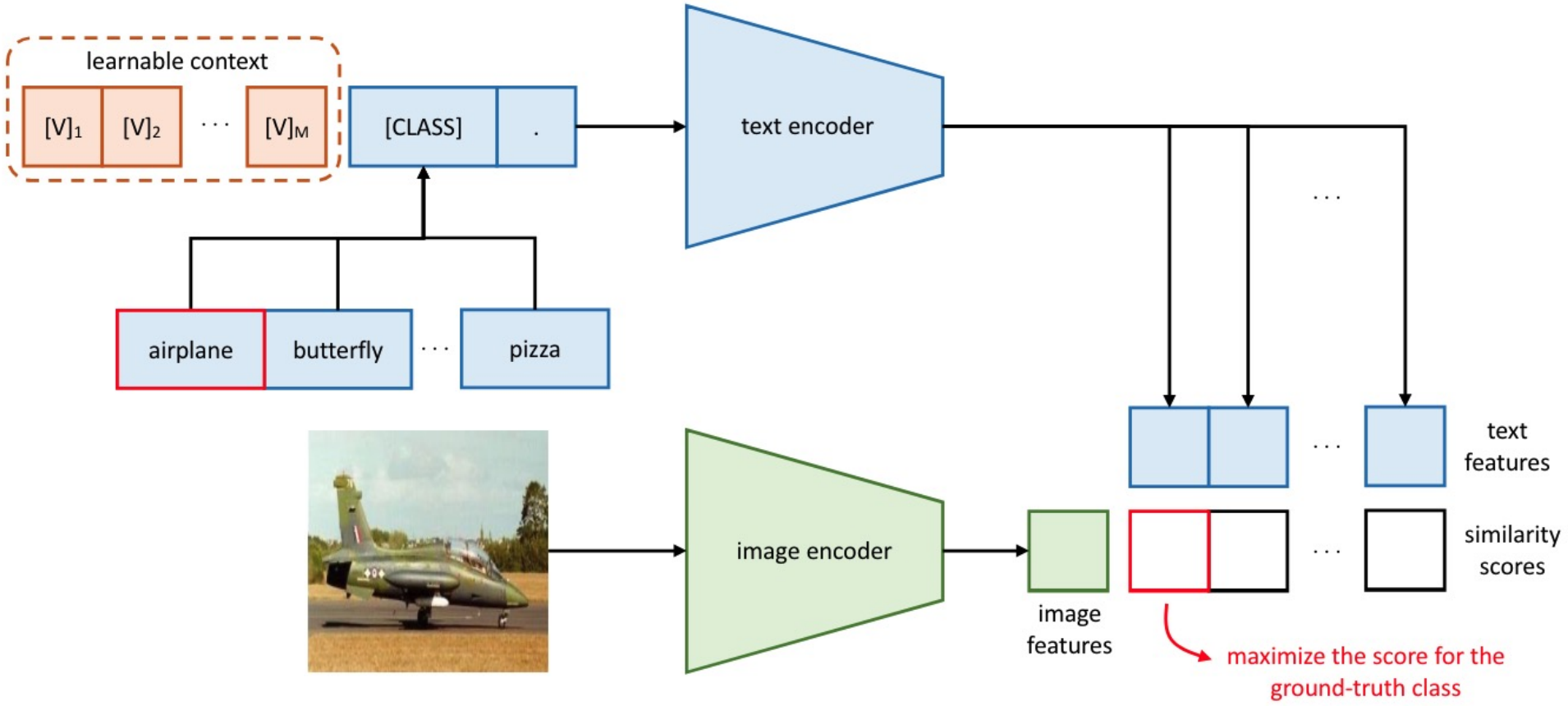
Cross attention





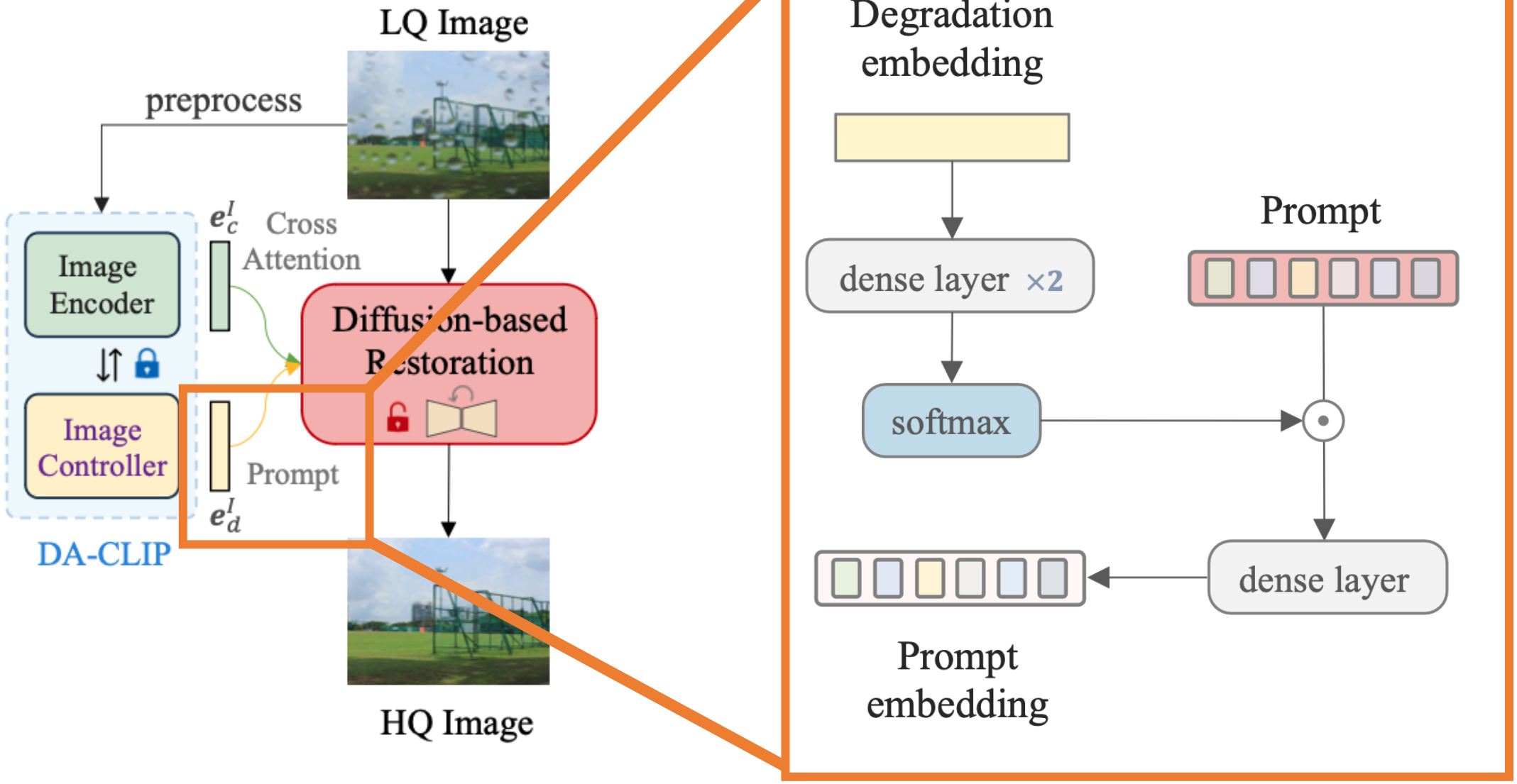
High-Resolution Image Synthesis with Latent Diffusion Models, CVPR2022

Prompt learning



Learning to Prompt for Vision-Language Models, IJCV 2022

Prompt module



Dataset Construction

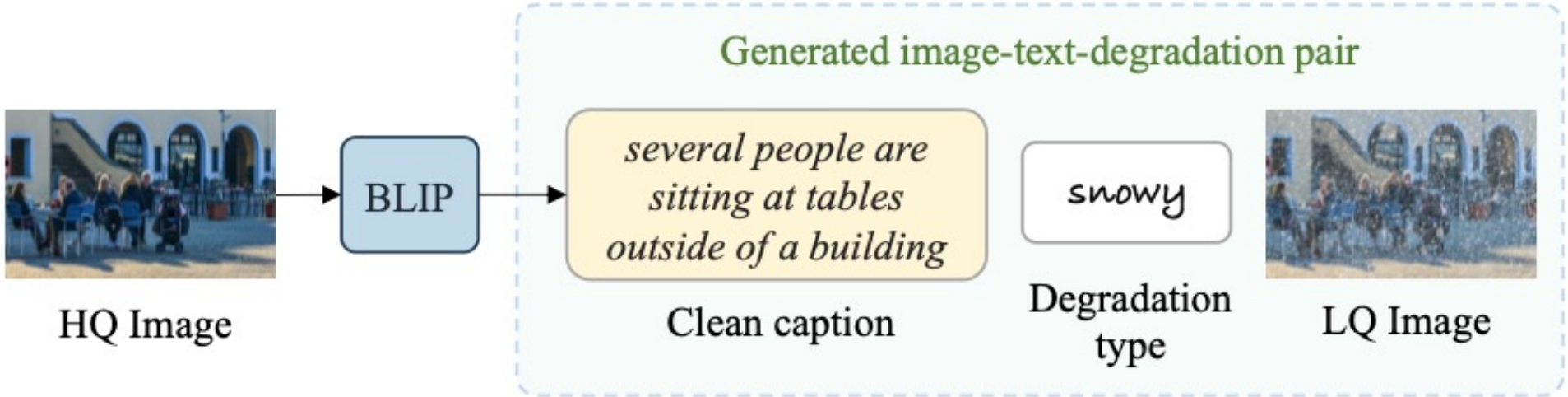


Table 1: Details of the collected training and testing datasets with different image degradation types.

Dataset	Blurry	Hazy	JPEG	Low-light	Noisy	Raindrop	Rainy	Shadowed	Snowy	Inpainting
#Train	2 103	6 000	3 550	485	3 550	861	1 800	2 680	1 872	29 900
#Test	1 111	1 000	29	15	68	58	100	408	601	100

Outline

- Preliminary
- Framework
- Method
- Experiment
- Conclusion

Degradation-Specific Image Restoration

Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
JORDER	26.25	0.835	0.197	94.58
PReNet	29.46	0.899	0.128	52.67
MPRNet	30.41	0.891	0.158	61.59
MAXIM	30.81	0.903	0.133	58.72
IR-SDE	31.65	0.904	0.047	18.64
Ours	**33.91**	**0.926**	**0.031**	**11.79**

(a) Deraining results on the Rain100H dataset.

Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
EnlightenGAN	17.61	0.653	0.372	94.71
MIRNet	**24.14**	0.830	0.250	69.18
URetinex-Net	19.84	0.824	0.237	52.38
MAXIM	23.43	**0.863**	0.098	48.59
IR-SDE	20.45	0.787	0.129	47.28
Ours	23.77	0.830	**0.083**	**34.03**

(b) Low-light enhancement on the LOL dataset.

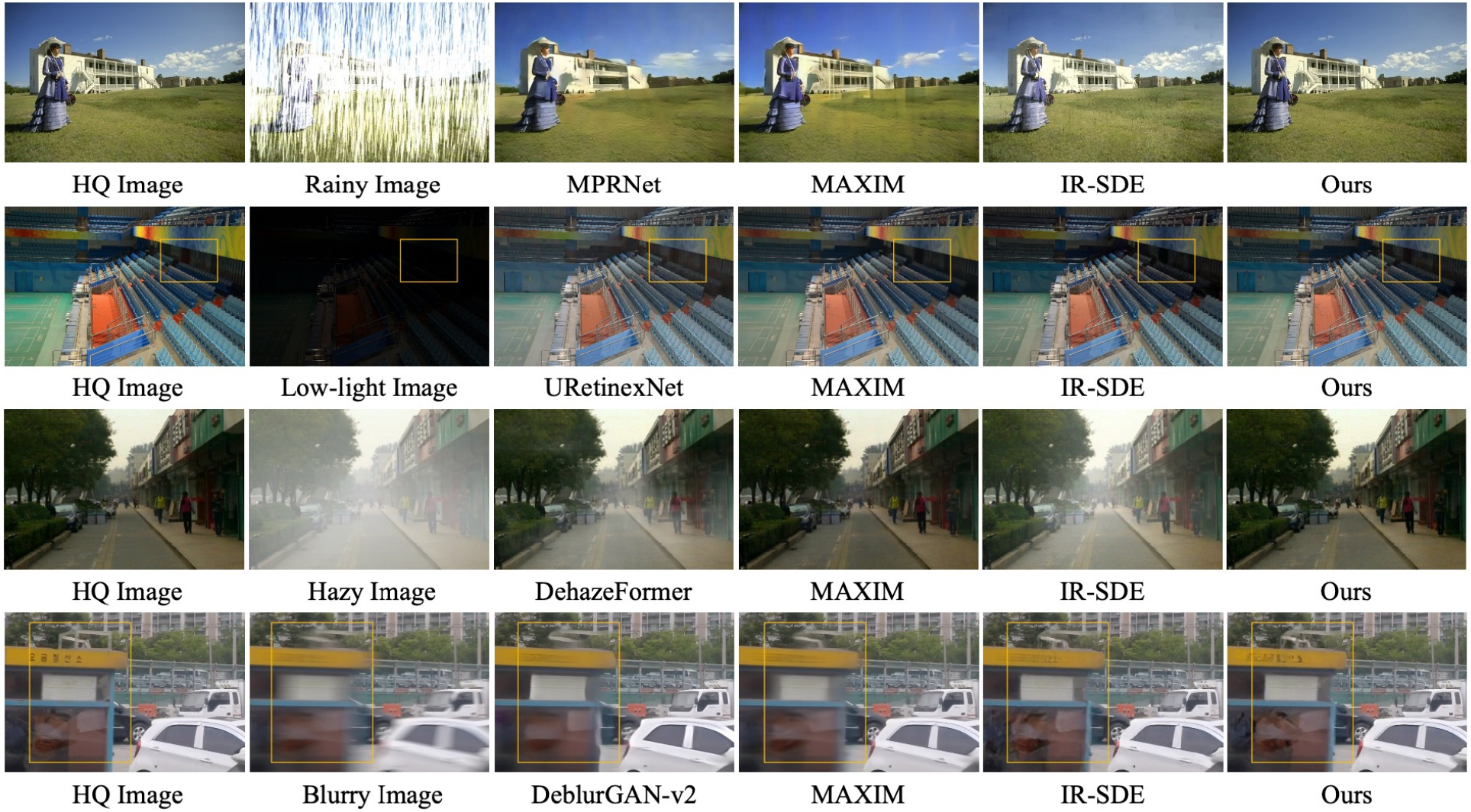
Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
DeepDeblur	29.08	0.913	0.135	15.14
DeblurGAN	28.70	0.858	0.178	27.02
DeblurGANv2	29.55	0.934	0.117	13.40
MAXIM	**32.86**	**0.940**	0.089	11.57
IR-SDE	30.70	0.901	0.064	6.32
Ours	30.88	0.903	**0.058**	**6.15**

(c) Deblurring results on the GoPro dataset.

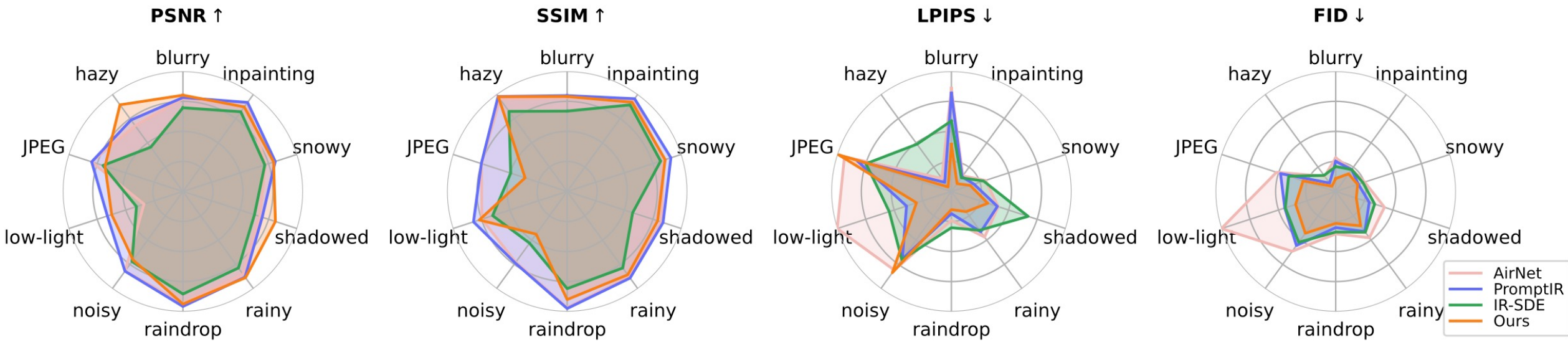
Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
GCANet	26.59	0.935	0.052	11.52
GridDehazeNet	25.86	0.944	0.048	10.62
DehazeFormer	**30.29**	**0.964**	0.045	7.58
MAXIM	29.12	0.932	0.043	8.12
IR-SDE	25.25	0.906	0.060	8.33
Ours	30.16	0.936	**0.030**	**5.52**

(d) Dehazing results on the RESIDE-6k dataset.

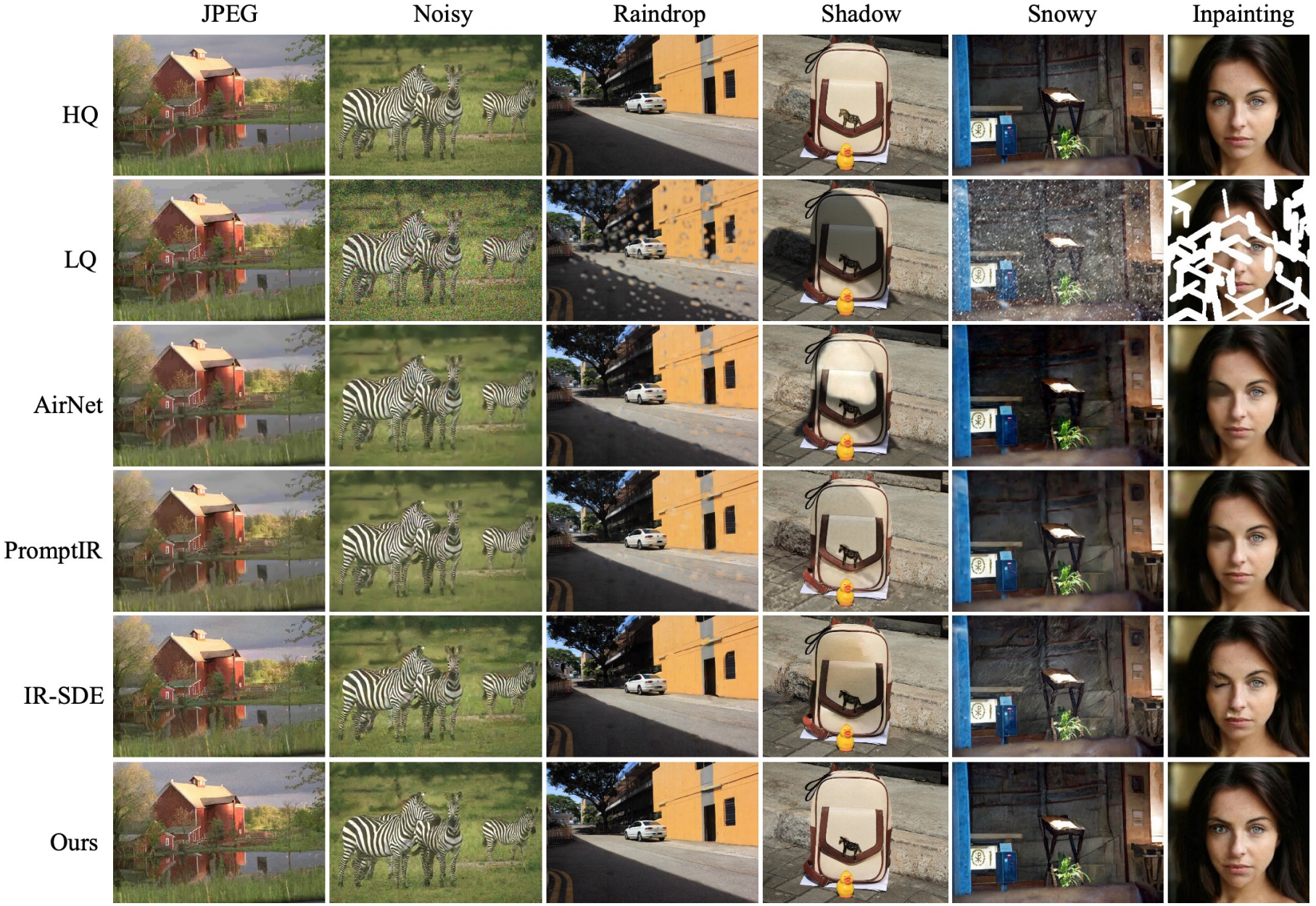
Degradation-Specific Image Restoration



Unified Image Restoration



Unified Image Restoration



Unified Image Restoration

PromptIR

Ours

Unified Image Restoration

Table 3: Comparison of the average results over ten different datasets on the *unified* image restoration task.

Method	Distortion		Perceptual	
	PSNR↑	SSIM↑	LPIPS↓	FID↓
NAFNet	26.34	0.847	0.159	55.68
NAFNet + Degradation	27.02	0.856	0.146	48.27
NAFNet + DA-CLIP	**27.22**	**0.861**	0.145	47.94
Restormer	26.43	0.850	0.157	54.03
AirNet	25.62	0.844	0.182	64.86
PromptIR	27.14	0.859	0.147	48.26
IR-SDE	23.64	0.754	0.167	49.18
Ours	27.01	0.794	**0.127**	**34.89**

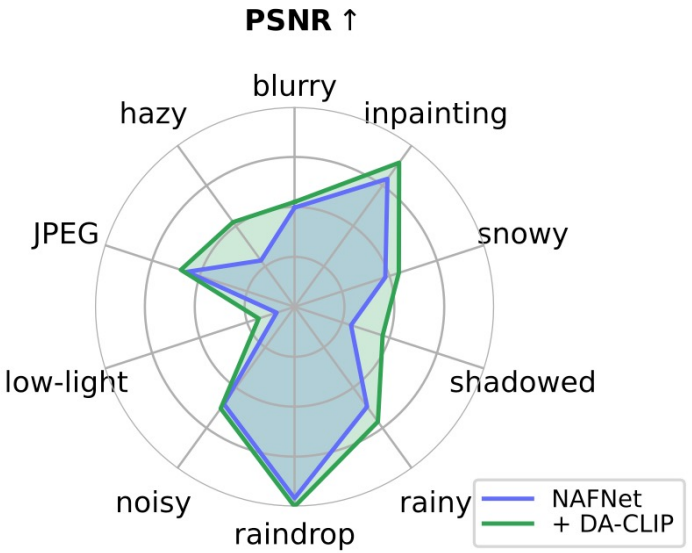


Figure 7: NAFNet with DA-CLIP for *unified* image restoration.

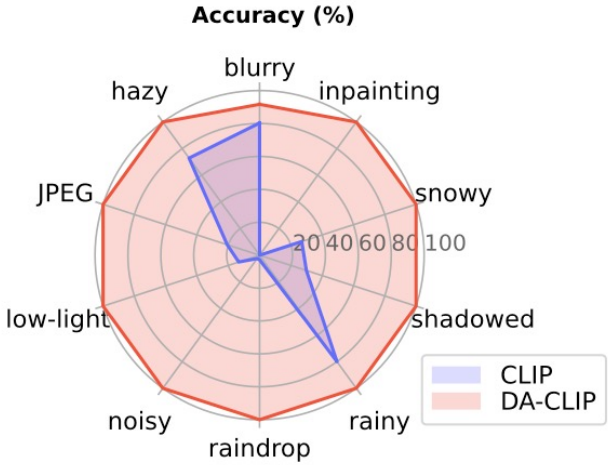
Discussion and Analysis

Image dehazing

Image deblurring

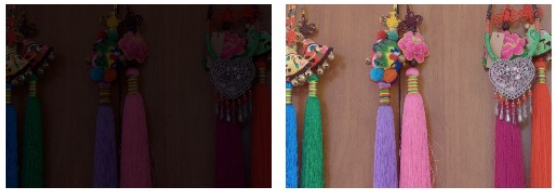
Face inpainting

JPEG artifact deduction



Degradation prediction

Image desnowing



Low-light image enhancement

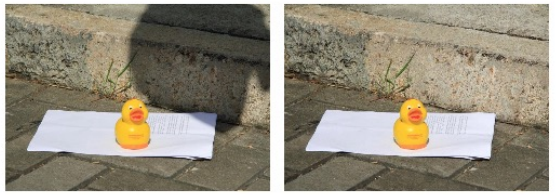


Image shadow removal

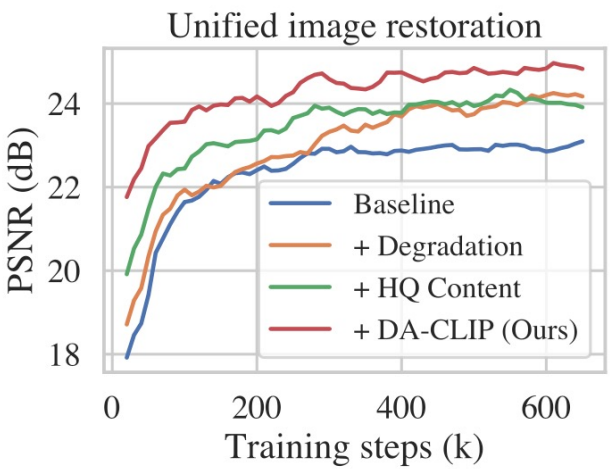
Image denoising

Image raindrop removal

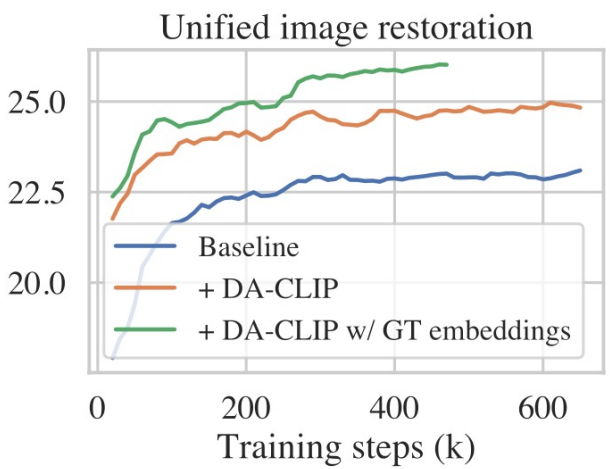
Image deraining

- prompt is set to “a [degradation type] photo”

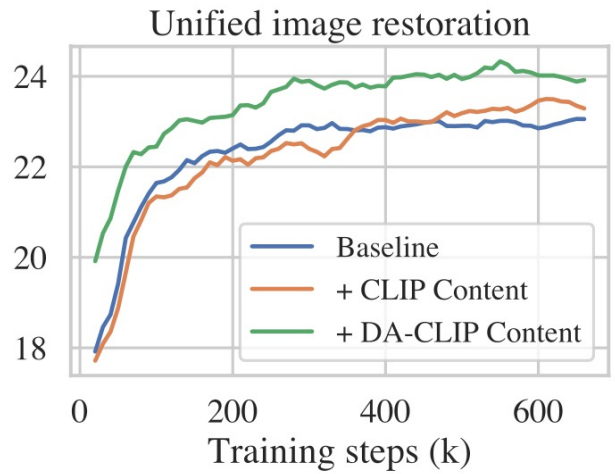
Discussion and Analysis



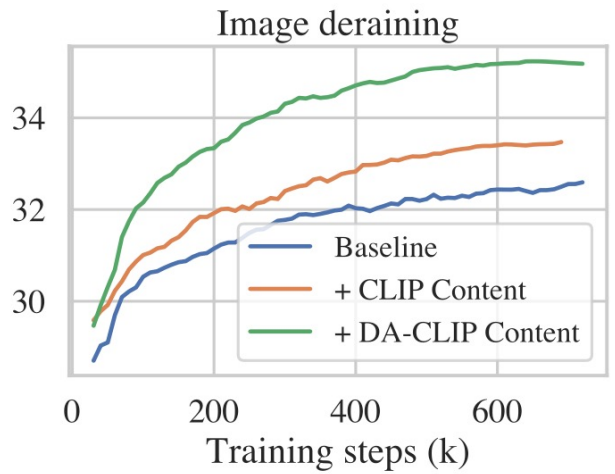
(a)



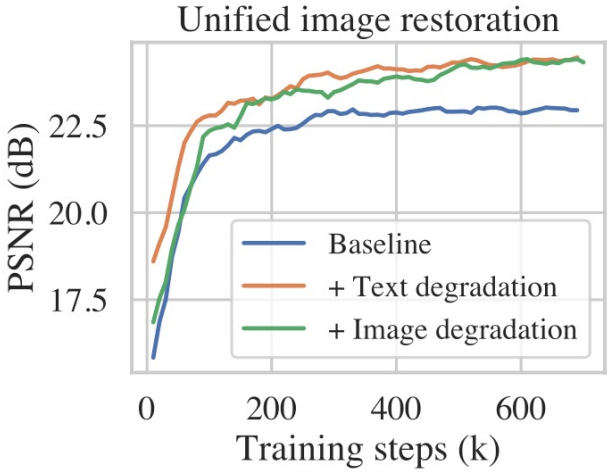
(b)



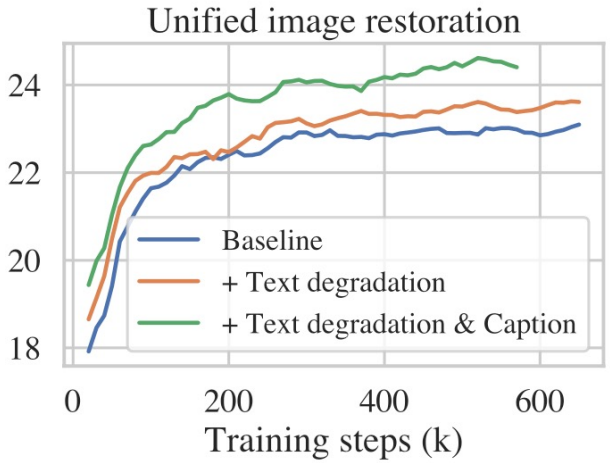
(c)



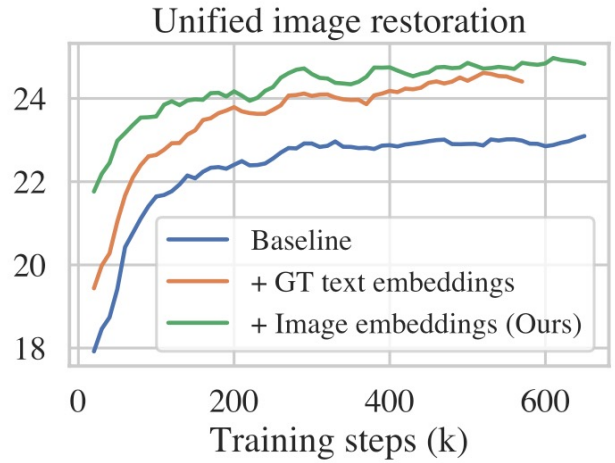
(d)



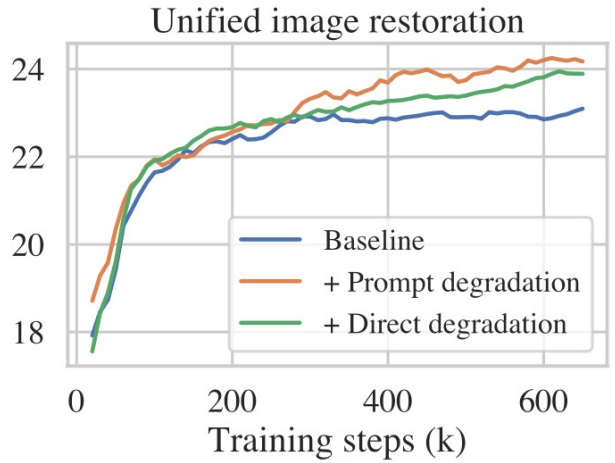
(e)



(f)

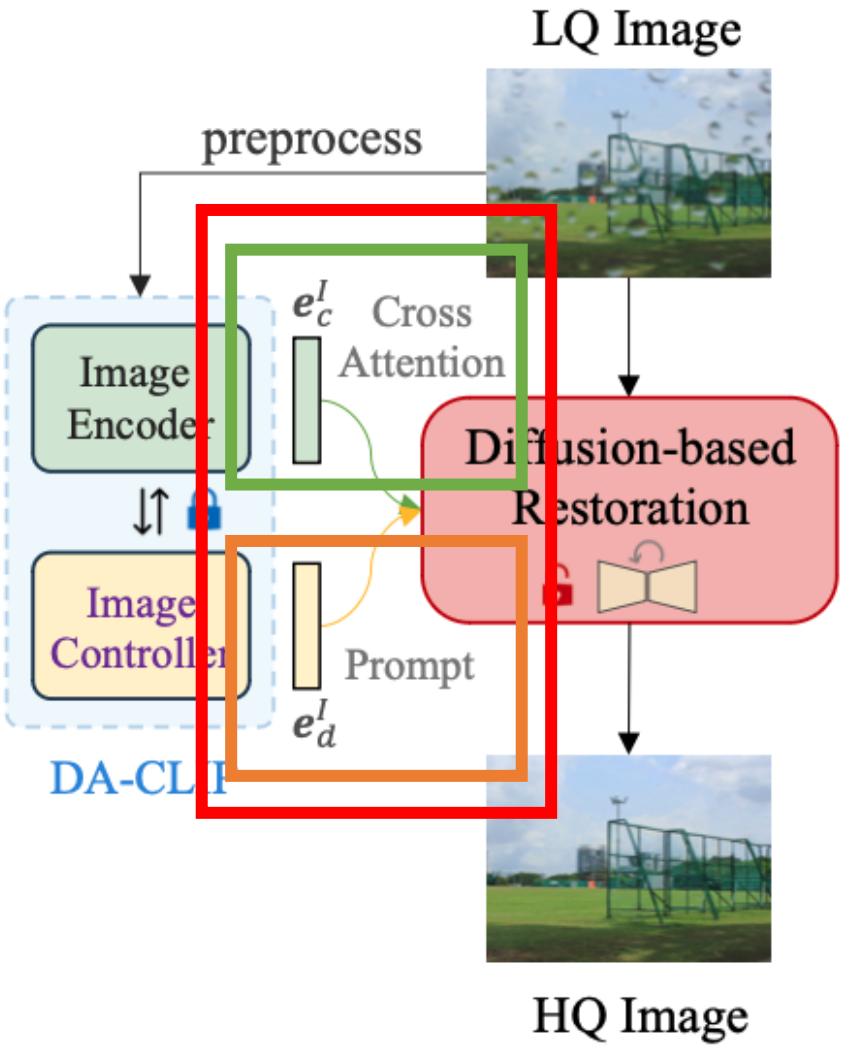


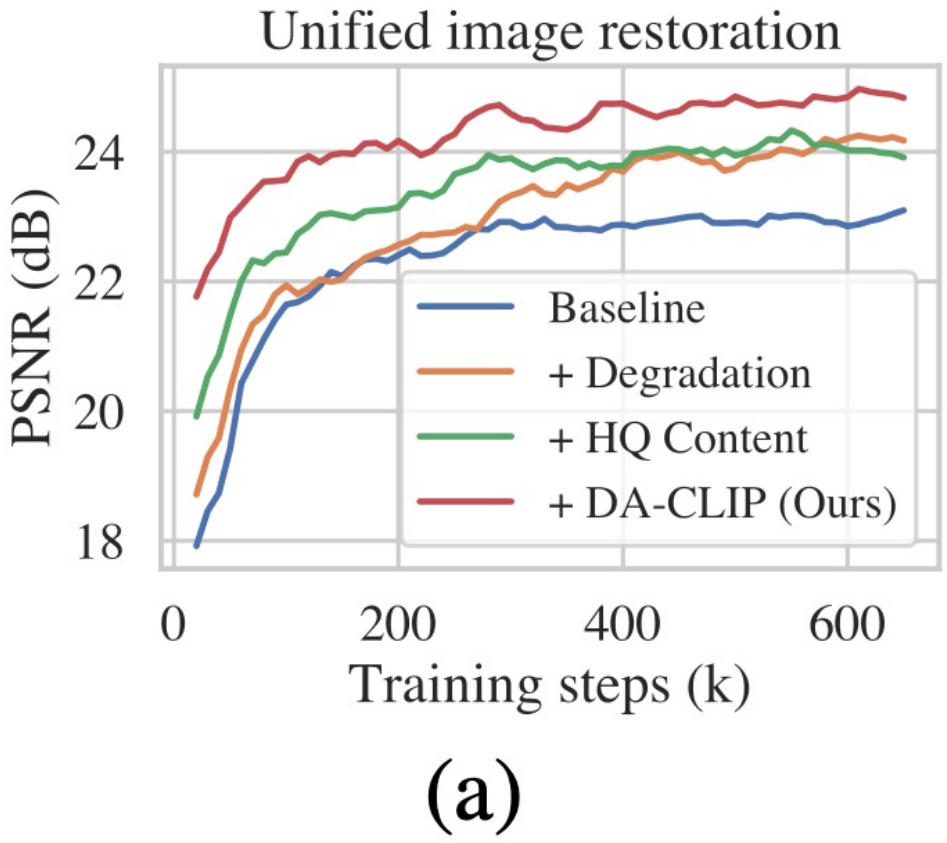
(g)



(h)

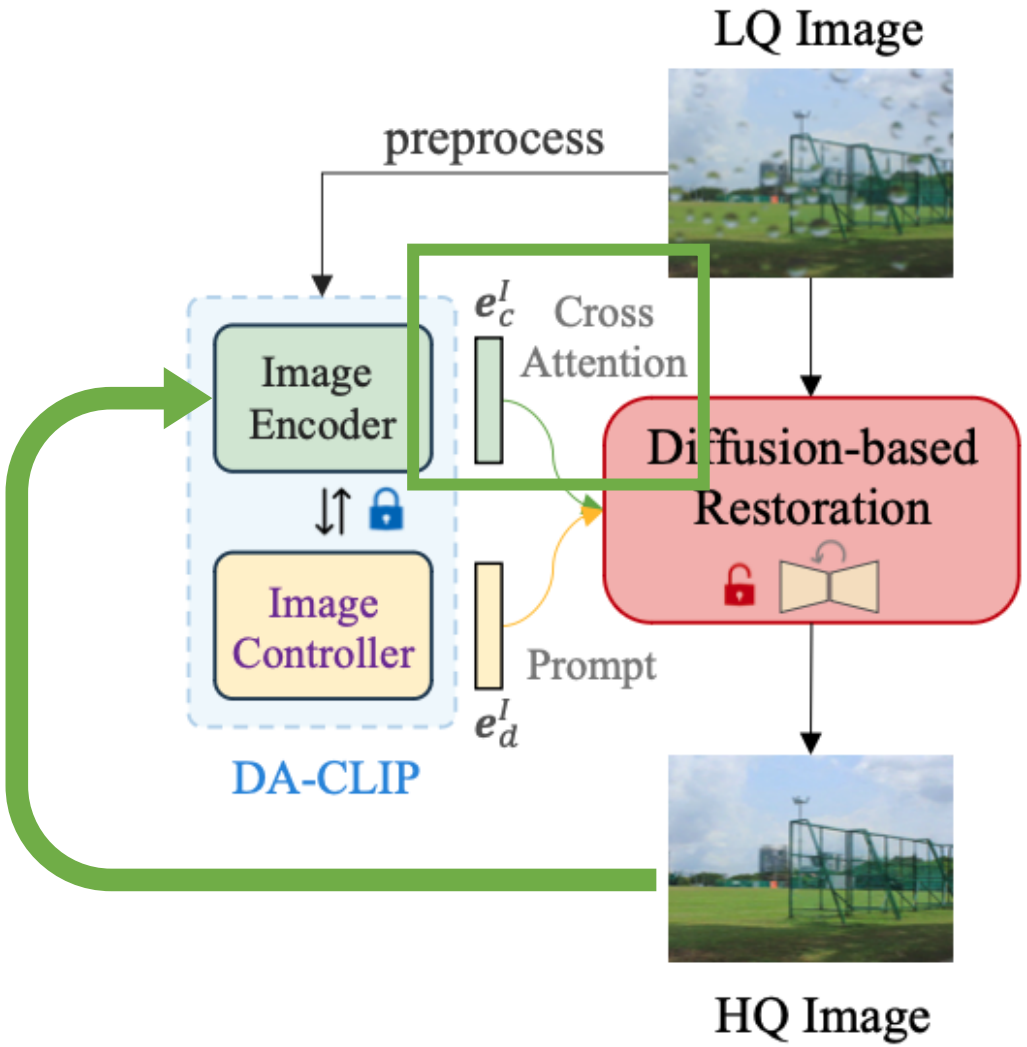
Discussion and Analysis

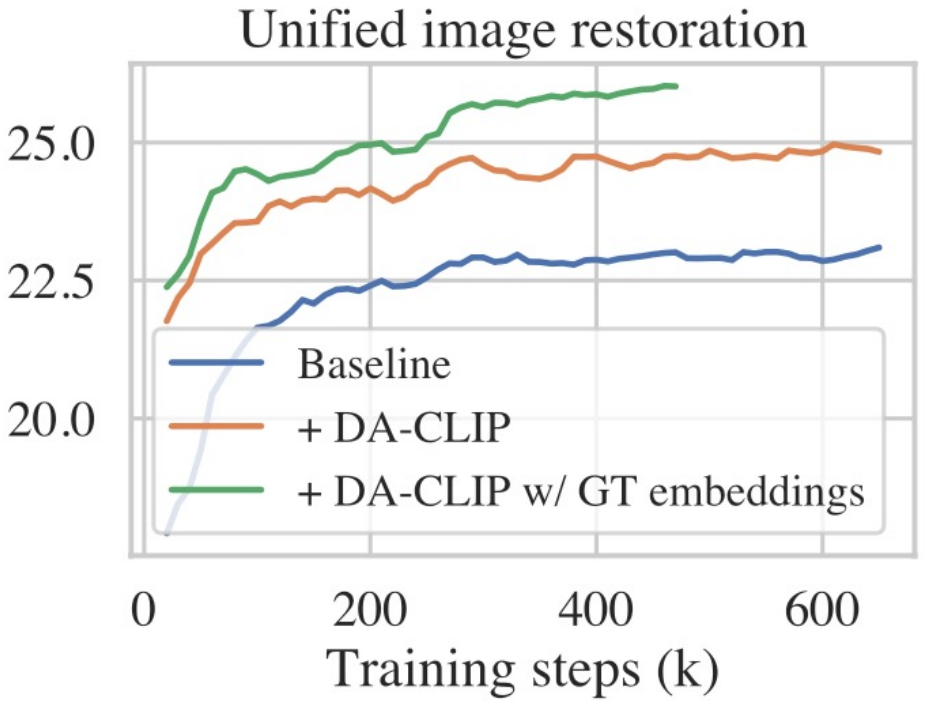




(a)

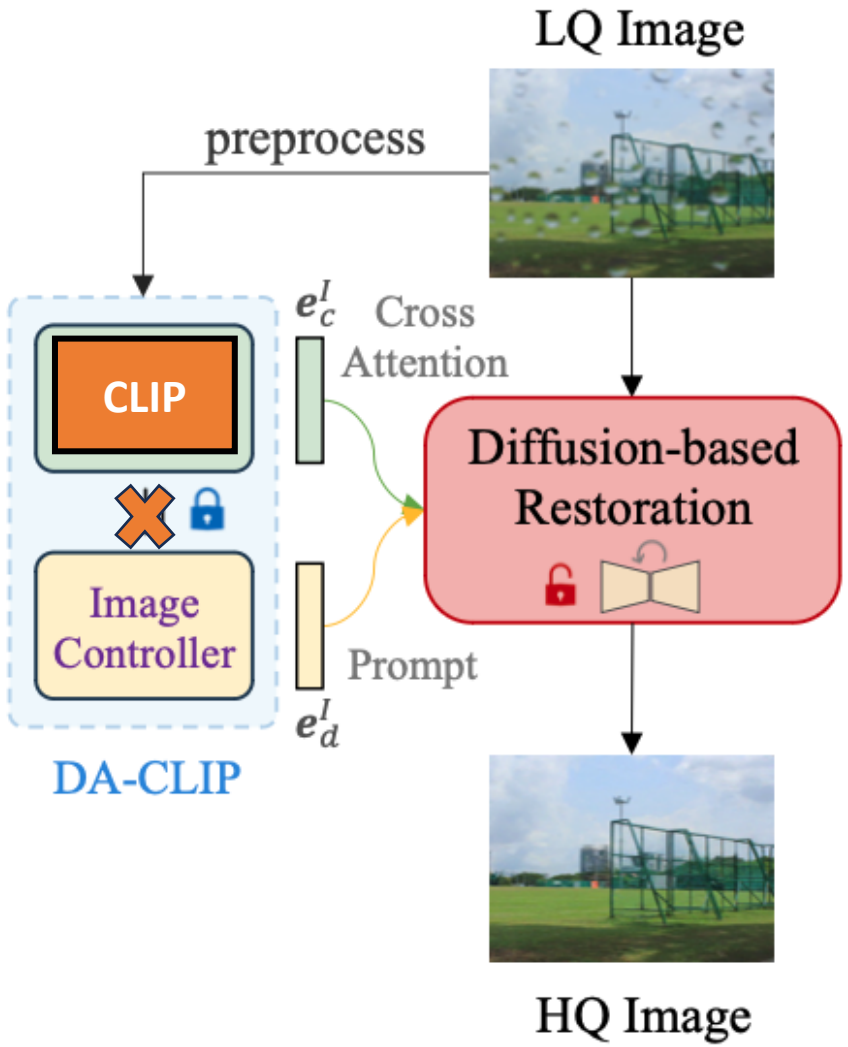
Discussion and Analysis

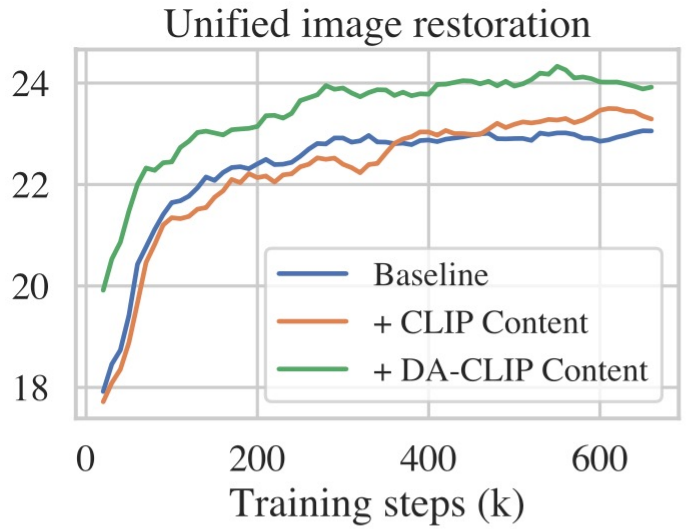




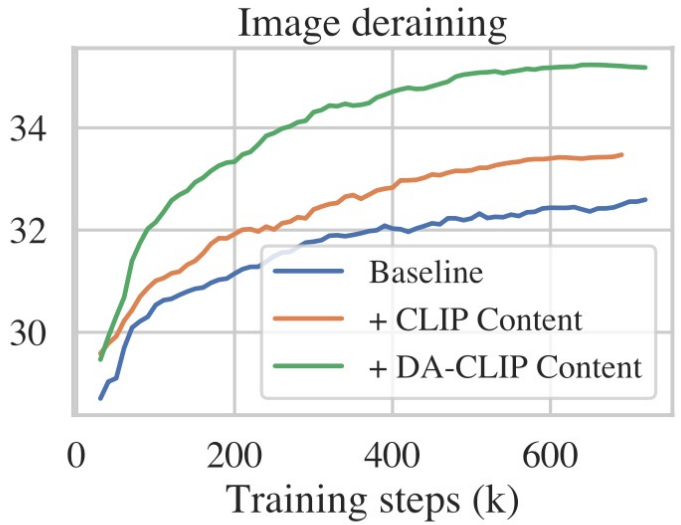
(b)

Discussion and Analysis



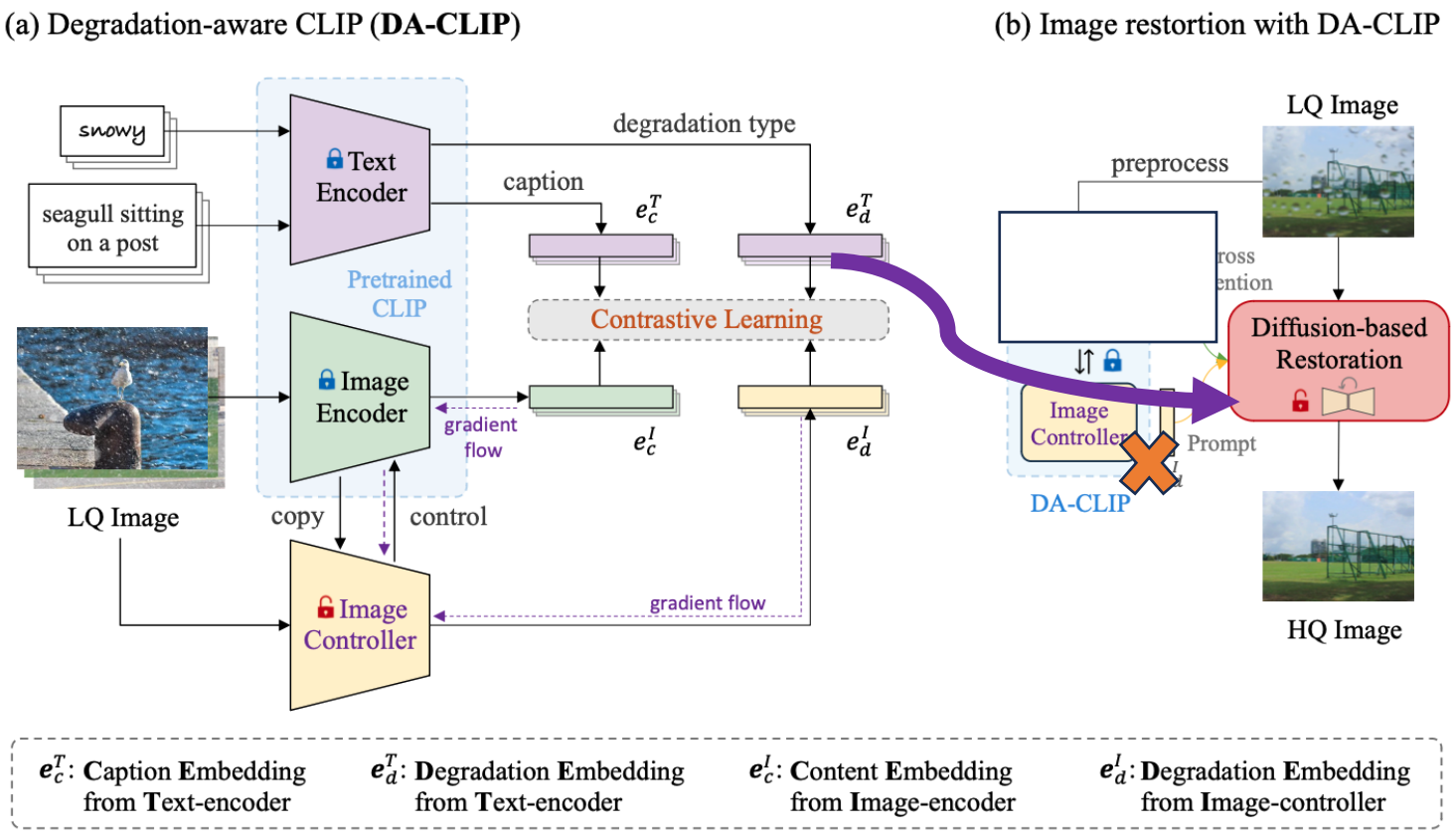


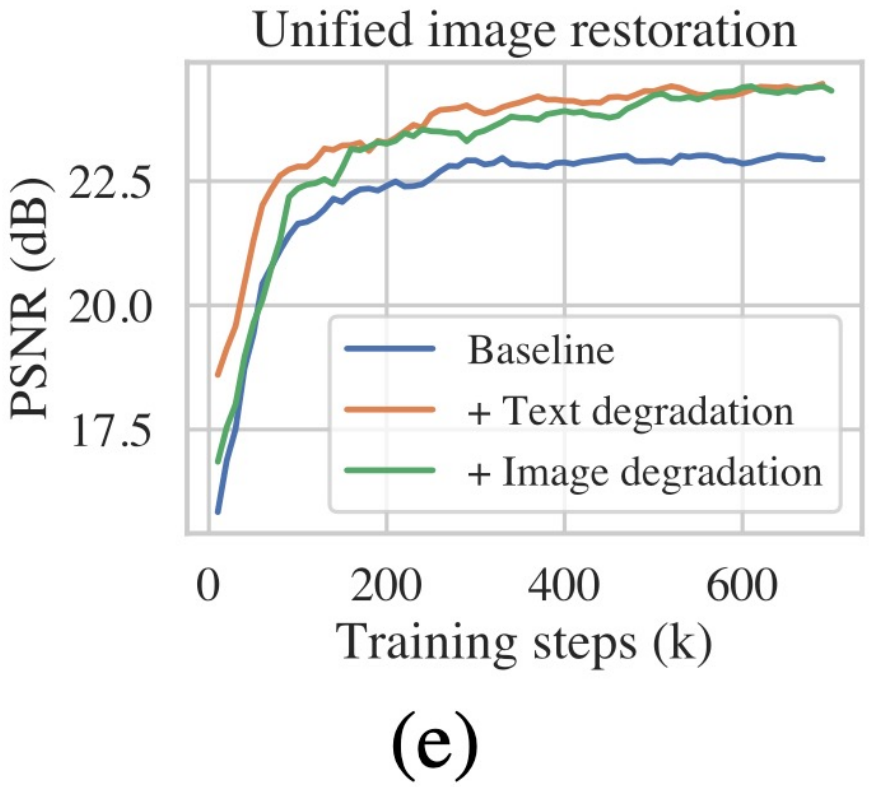
(c)



(d)

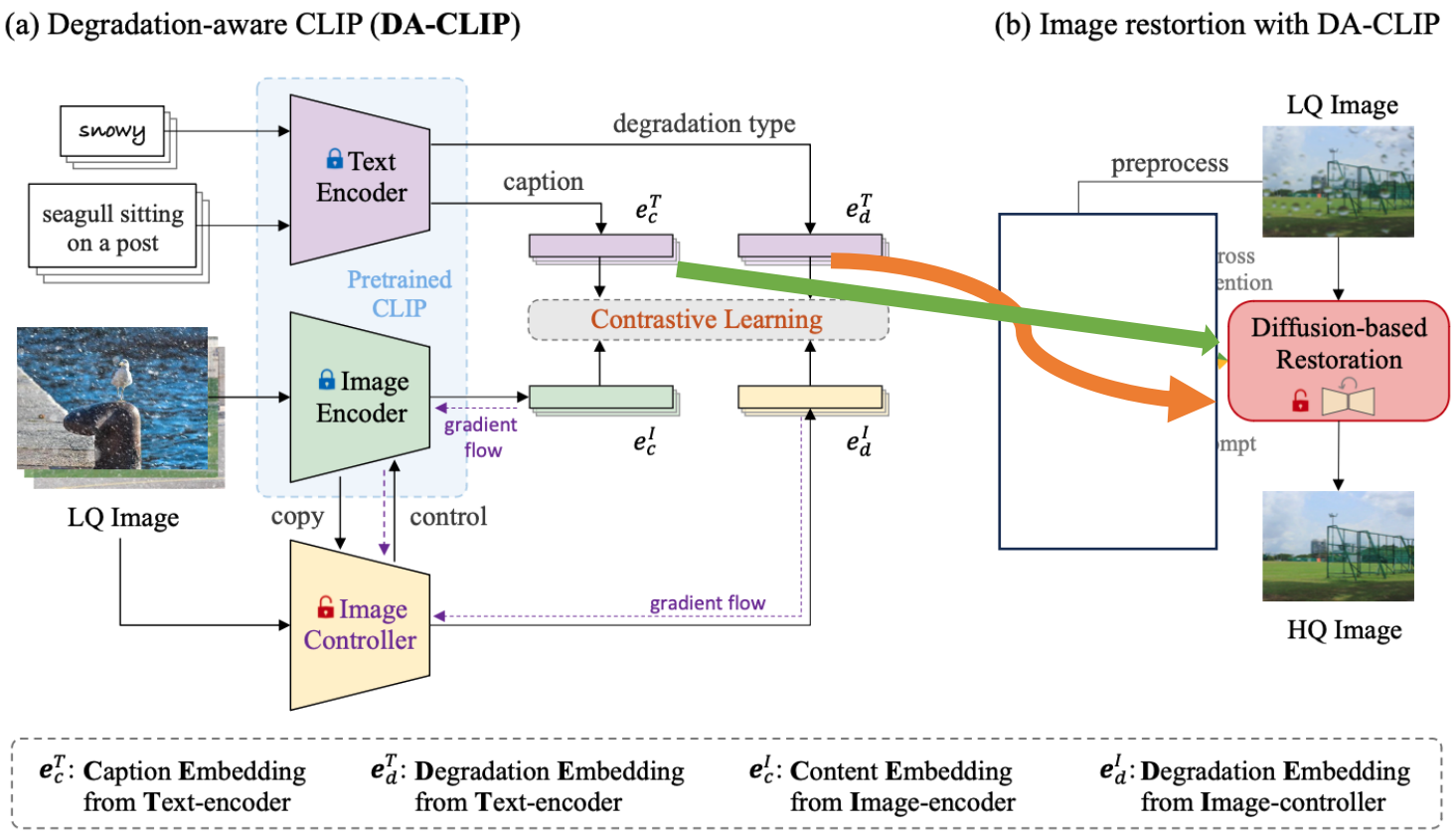
Discussion and Analysis

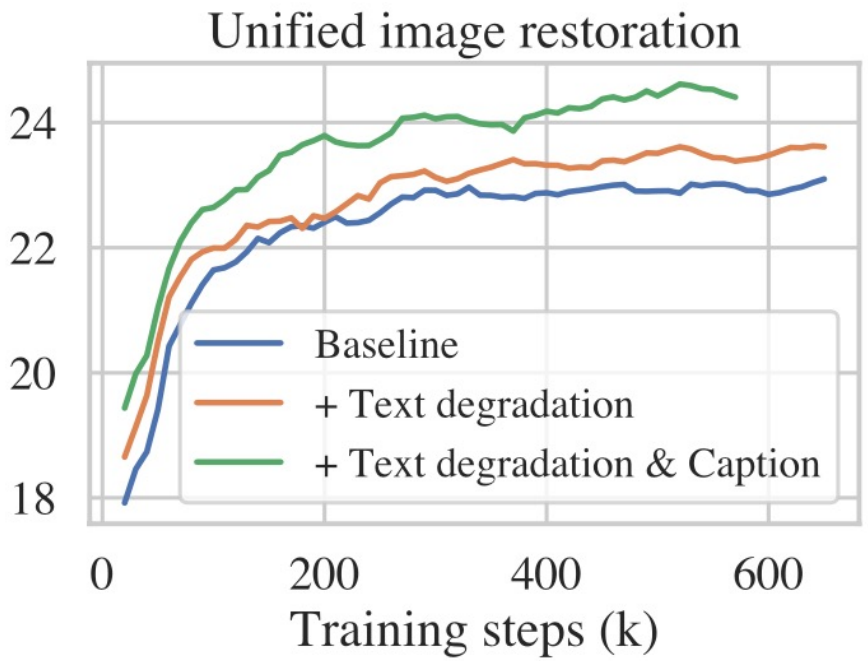




(e)

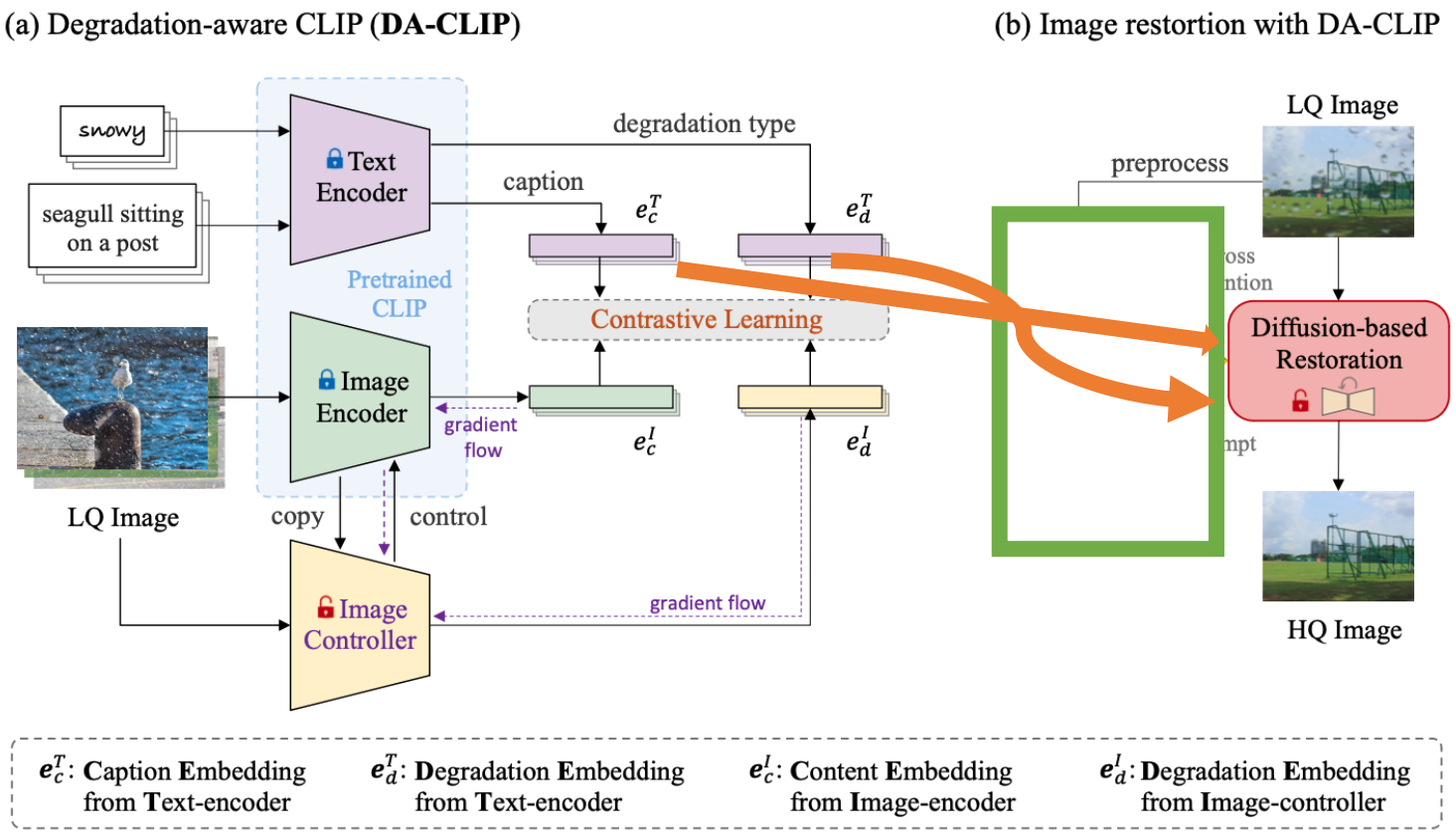
Discussion and Analysis

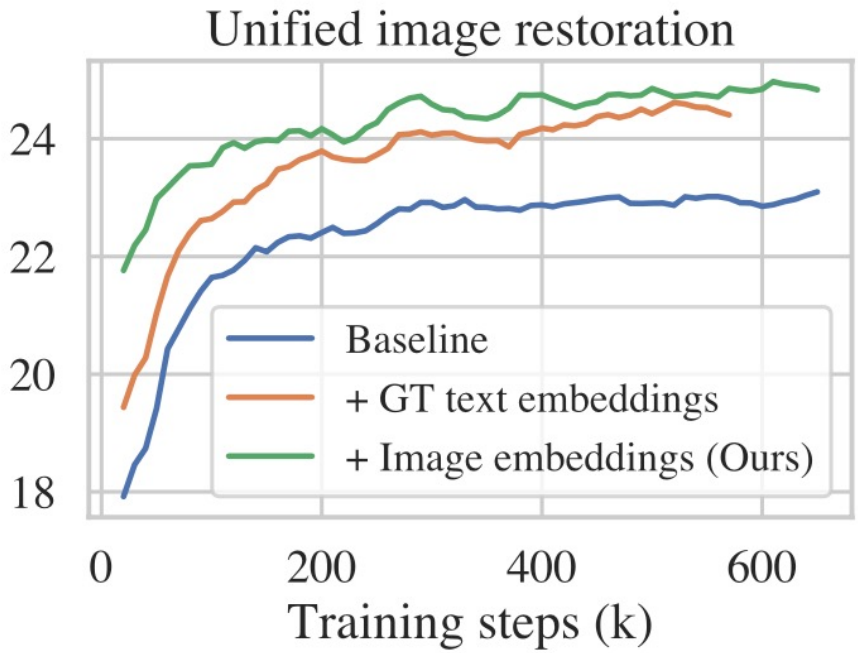




(f)

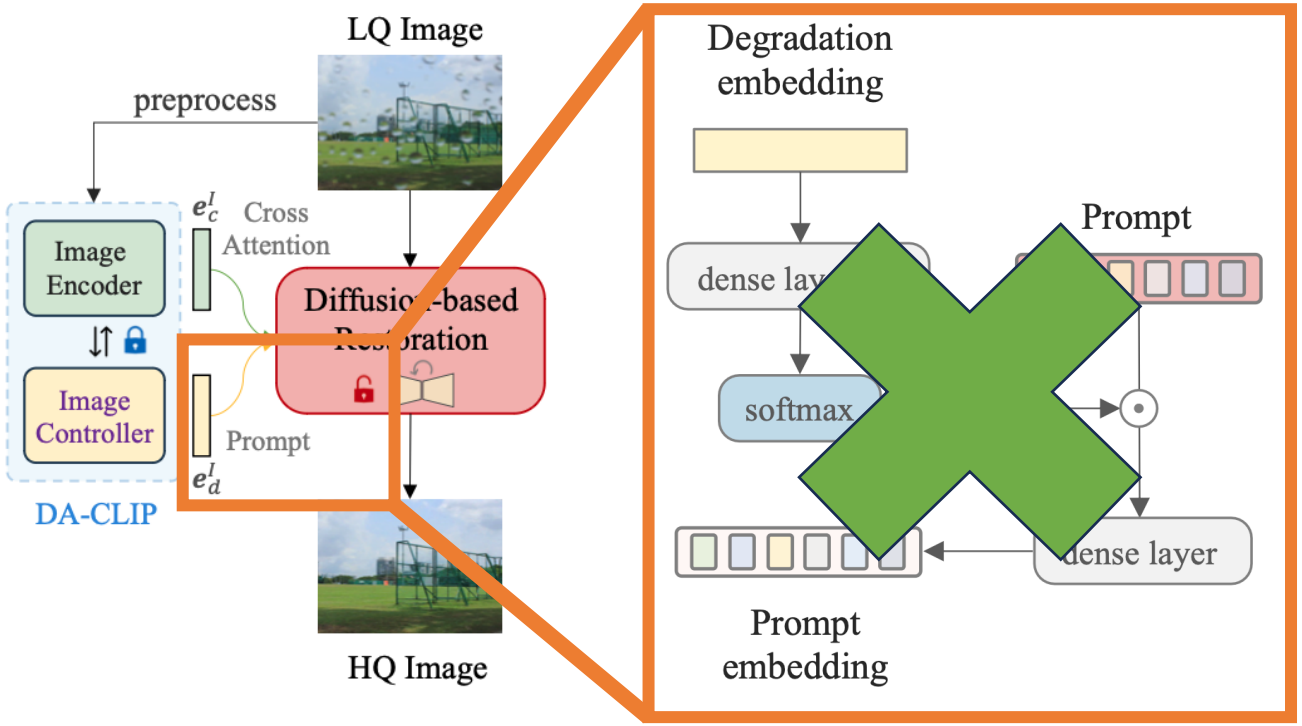
Discussion and Analysis

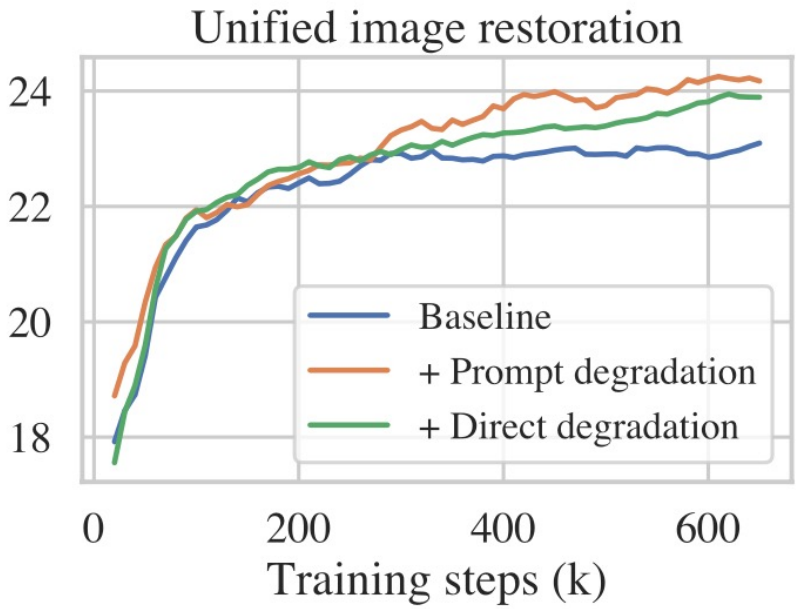




(g)

Discussion and Analysis





(h)

Outline

- Preliminary
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- Present DA-CLIP to leverage **large-scale pretrained vision-language models** as a framework for image restoration. Image controller that predicts the **degradation** and **HQ content embeddings** from corrupted inputs.

- Use **cross-attention** to integrate the content embedding into restoration networks and **prompt learning module** to utilize the degradation context.

- Demonstrate the effectiveness of DA-CLIP by applying it to image restoration models for both degradation-specific and unified image restoration across all ten degradation types.