Multimodal Prompt Perceiver: Empower Adaptiveness, Generalizability and Fidelity for All-in-One Image Restoration

Yuang Ai[1,2] Huaibo Huang[1,2]✉ Xiaoqiang Zhou[1,3] Jiexiang Wang[1,3] Ran He[1,2]
[1]MAIS & CRIPAC, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]University of Science and Technology of China, Hefei, China
shallowdream555@gmail.com, huaibo.huang@cripac.ia.ac.cn,
{xq525,jiexiang}@mail.ustc.edu.cn, rhe@nlpr.ia.ac.cn

CVPR 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Introduction

- Novel multimodal prompt learning approach that **harnesses Stable Diffusion (SD) priors** to enhance adaptiveness, generalizability and fidelity for all-in-one image restoration.

- Developing a **dual-branch module** to master two types of SD prompts, **textual for holistic** representation and **visual for multiscale detail** representation. Both prompts are **dynamically adjusted** by degradation predictions.

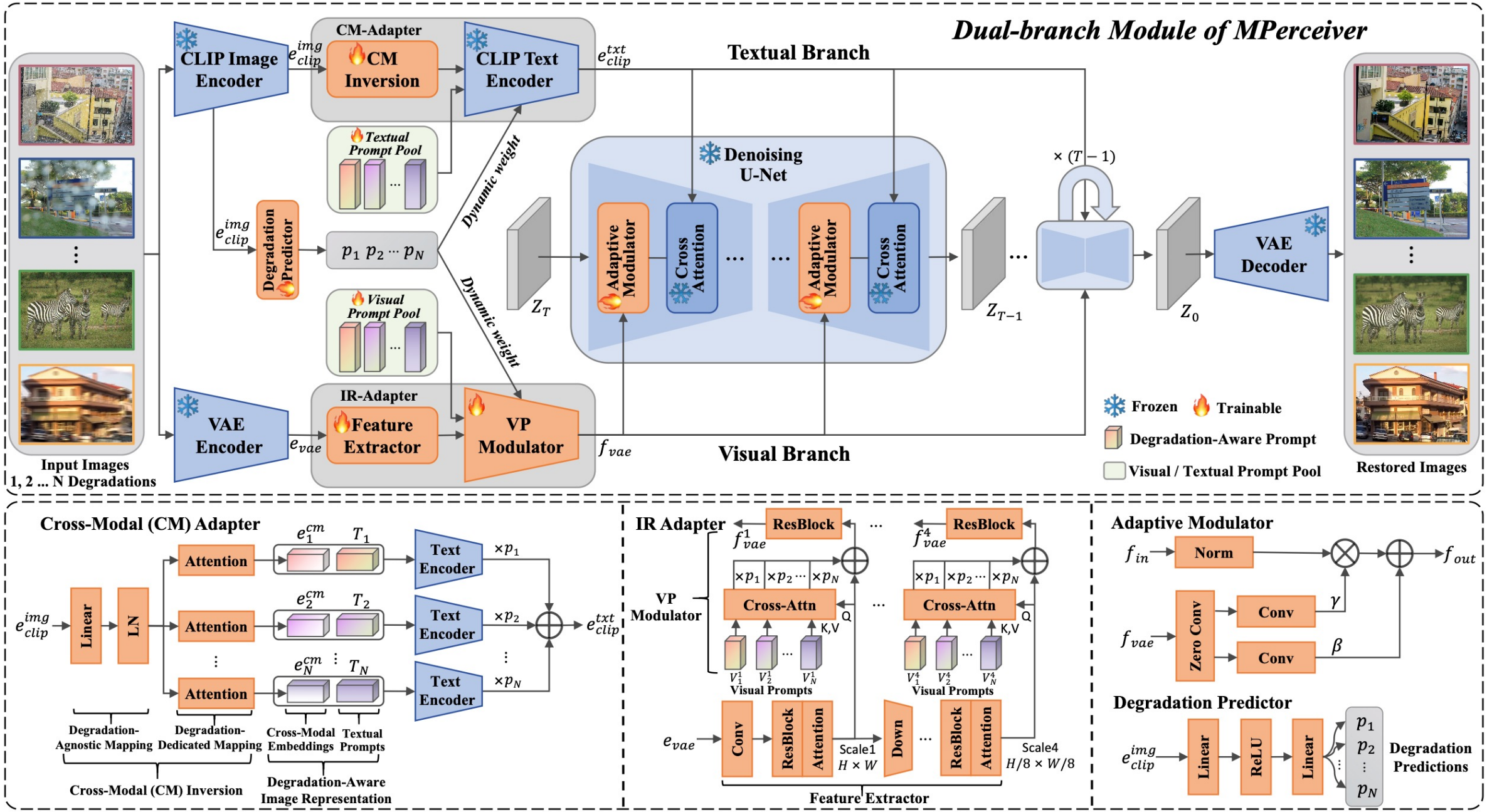
- Extensive experiments on 16 IR tasks including all-in-one, few-shot and zero-shot underscore the superiority of MPerceiver in terms of adaptiveness, generalizability and fidelity.

Introduction

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Framework


Dual-branch Module of MPerceiver
CLIP Image Encoder
CM-Adapter
CM Inversion
CLIP Text Encoder
Textual Branch
Textual Prompt Pool
Degradation Predictor
Dynamic weight
Denoising U-Net
Adaptive Modulator
Cross Attention
× (T − 1)
VAE Decoder
Visual Prompt Pool
VAE Encoder
IR-Adapter
Feature Extractor
VP Modulator
Visual Branch
Frozen
Trainable
Degradation-Aware Prompt
Visual / Textual Prompt Pool
Input Images
1, 2 ... N Degradations
Restored Images
Cross-Modal (CM) Adapter
Linear
LN
Attention
Text Encoder
Degradation-Agnostic Mapping
Degradation-Dedicated Mapping
Cross-Modal Embeddings
Textual Prompts
Cross-Modal (CM) Inversion
Degradation-Aware Image Representation
IR Adapter
ResBlock
Cross-Attn
Visual Prompts
Conv
Down
Scale1 H × W
Scale4 H/8 × W/8
Feature Extractor
Adaptive Modulator
Norm
Zero Conv
Degradation Predictor
ReLU
Degradation Predictions

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

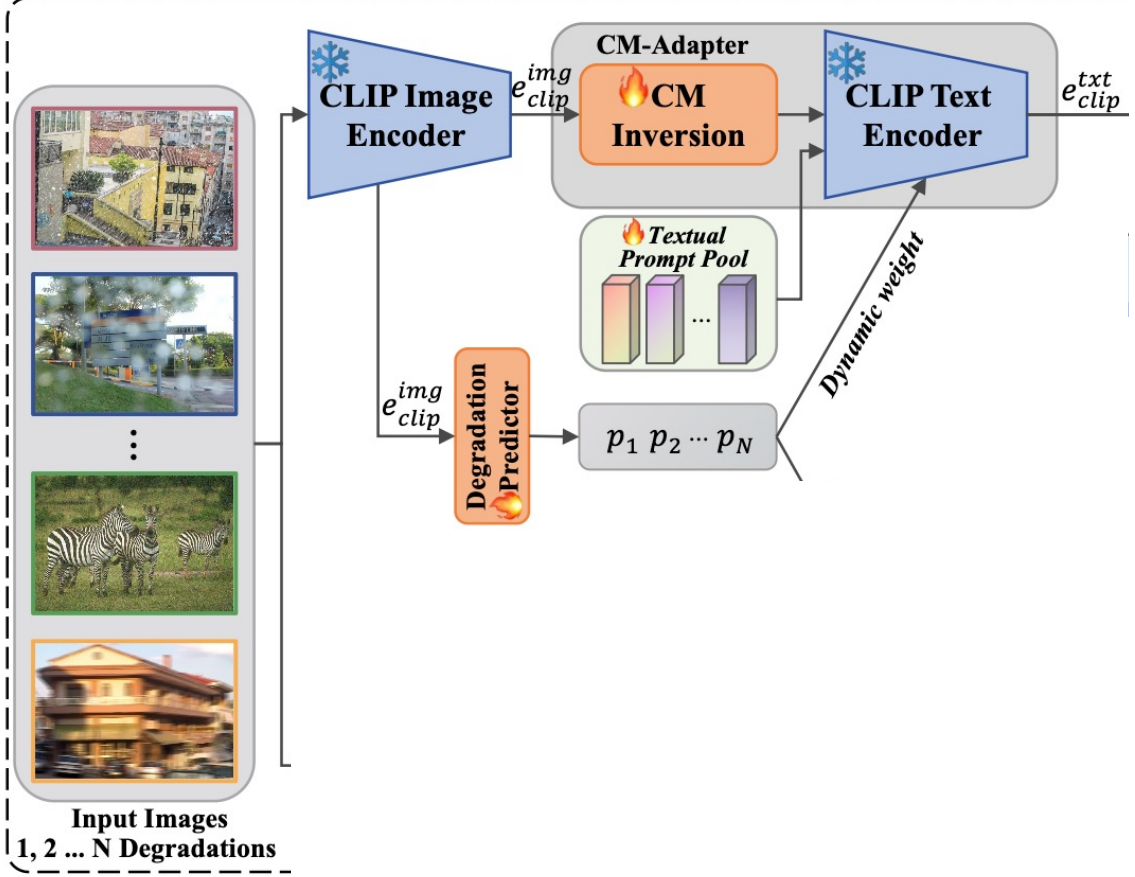
CM-Adapter

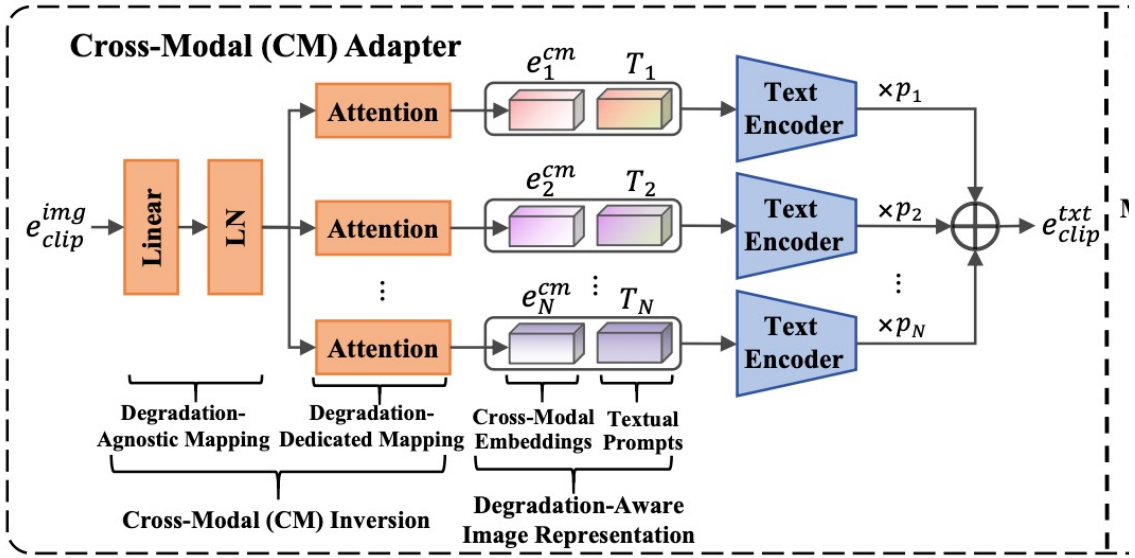
$$e_i^{cm} = \text{Attn}_i(\text{LN}(\text{FC}(e_{clip}^{img}))), i \in \{1, \cdots, N\},$$

$$T = \{T_1, \cdots, T_N\} \in \mathbb{R}^{N \times L \times C_{clip}}$$

$$P = \{p_1, \cdots, p_N\} \in \mathbb{R}^N$$

$$e_{clip}^{txt} = \sum_{i=1}^{N} p_i \mathcal{E}_{clip}^{txt}(\{e_i^{cm}, T_i\}).$$





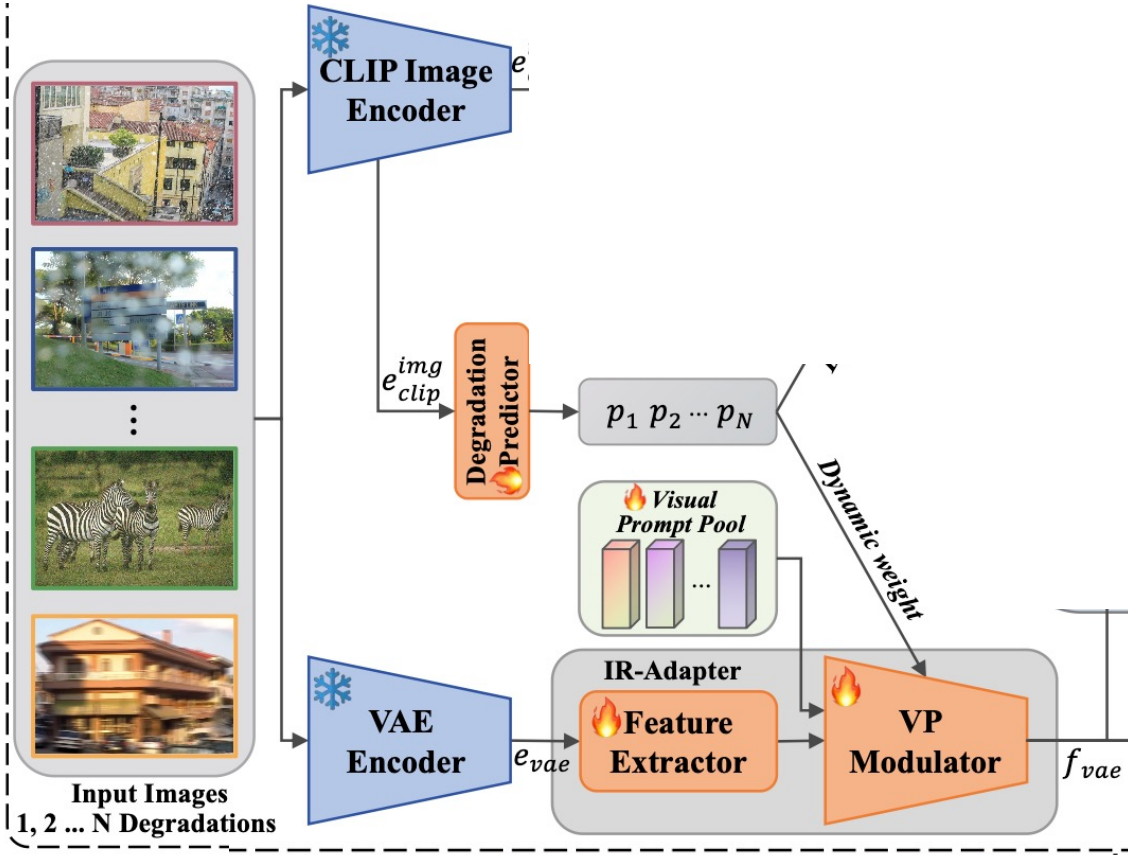
IR-Adapter

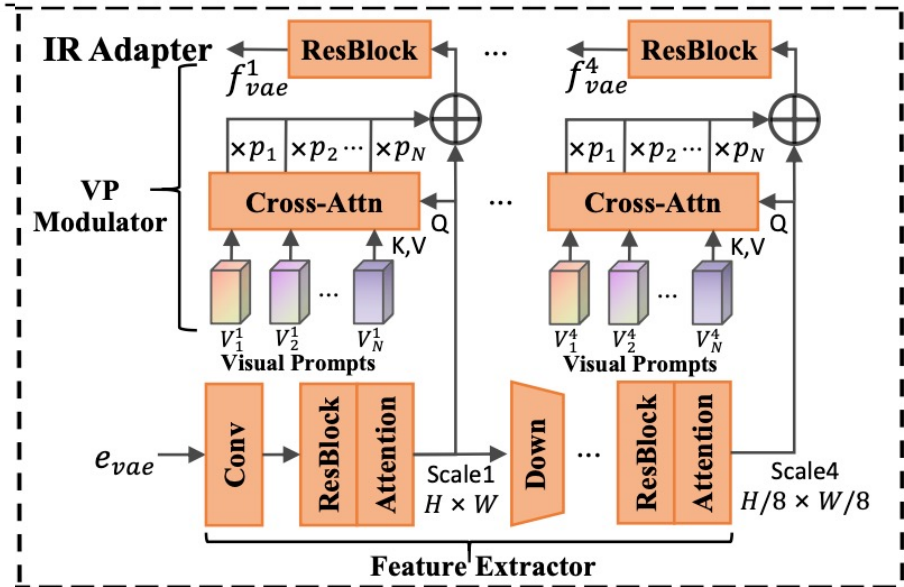
$$e_{vae} \in \mathbb{R}^{H \times W \times 4}$$

$$f_{vae} = \{f_{vae}^1, f_{vae}^2, f_{vae}^3, f_{vae}^4\}$$

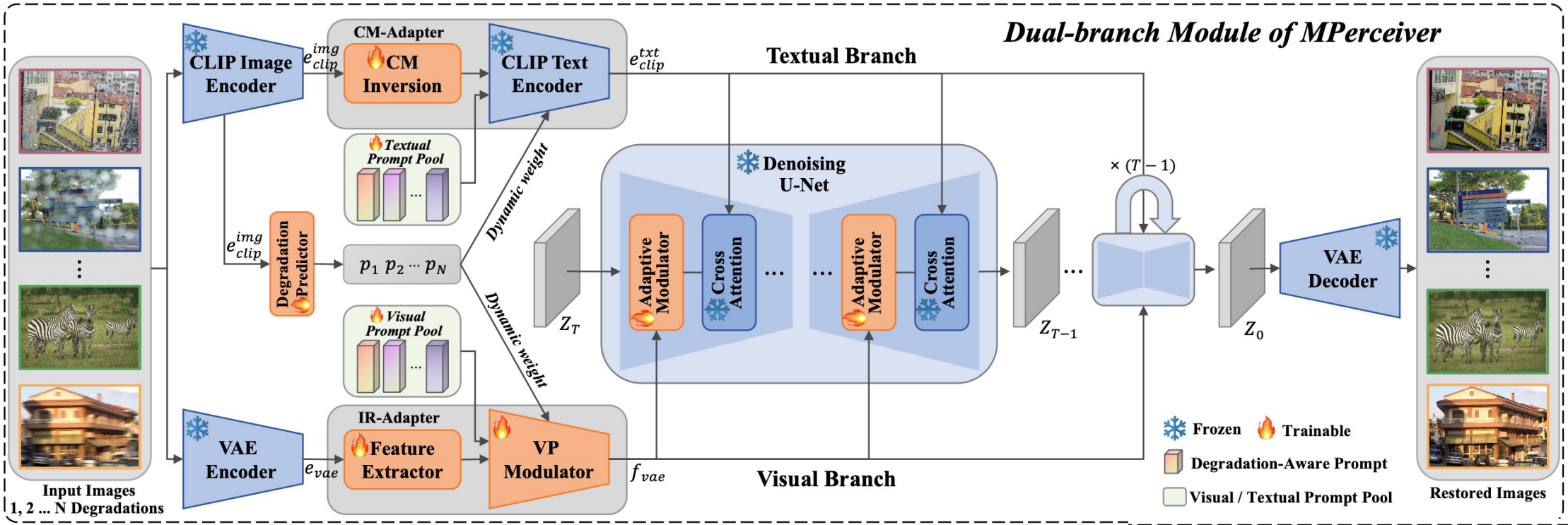
$$V = \{V^k, | k \in \{1, 2, 3, 4\}\}$$

$$f_{vae}^k = \mathrm{RB}_k(f_{vae}^k + \sum_{i=1}^{N} p_i \mathrm{MHCA}_k(f_{vae}^k, V_i^k, V_i^k))$$

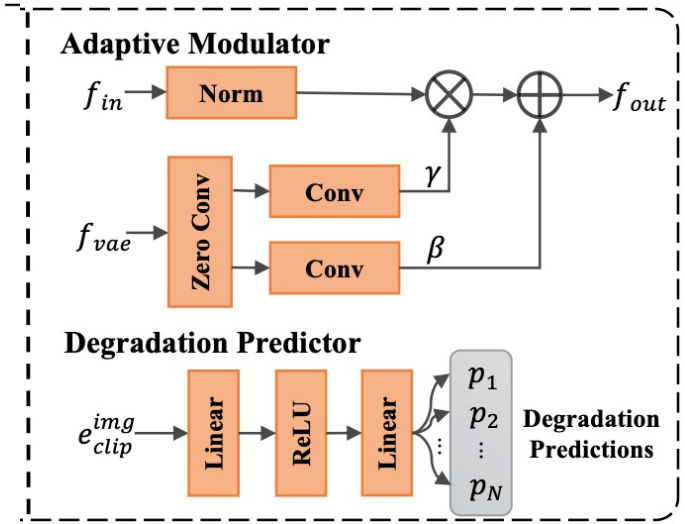




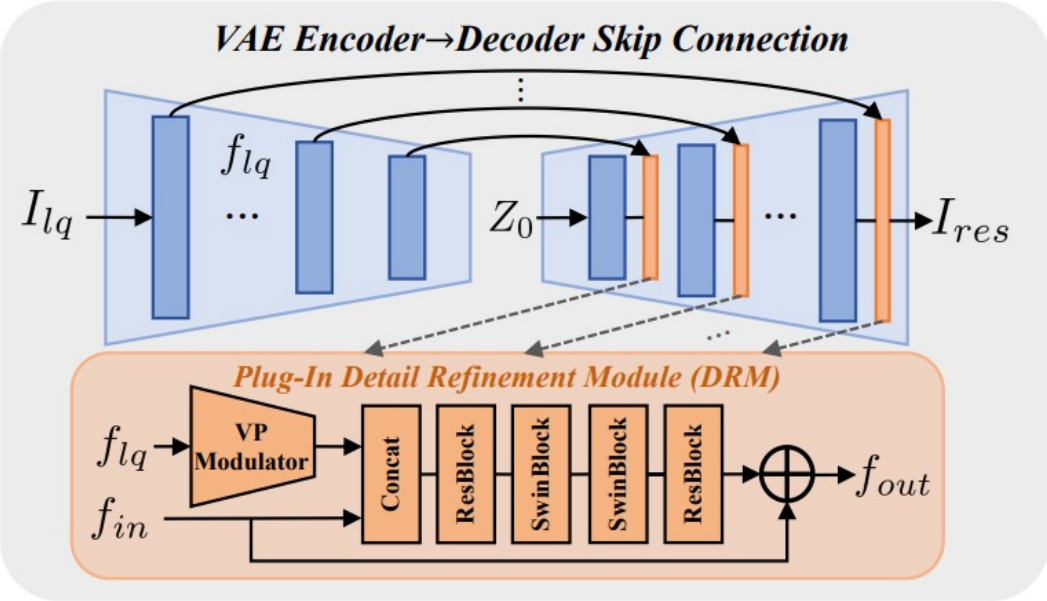
Adaptive Modulator



$$f_{out} = \gamma(f_{vae}) \odot \mathrm{Norm}(f_{in}) + \beta(f_{vae})$$



Detail Refinement Module

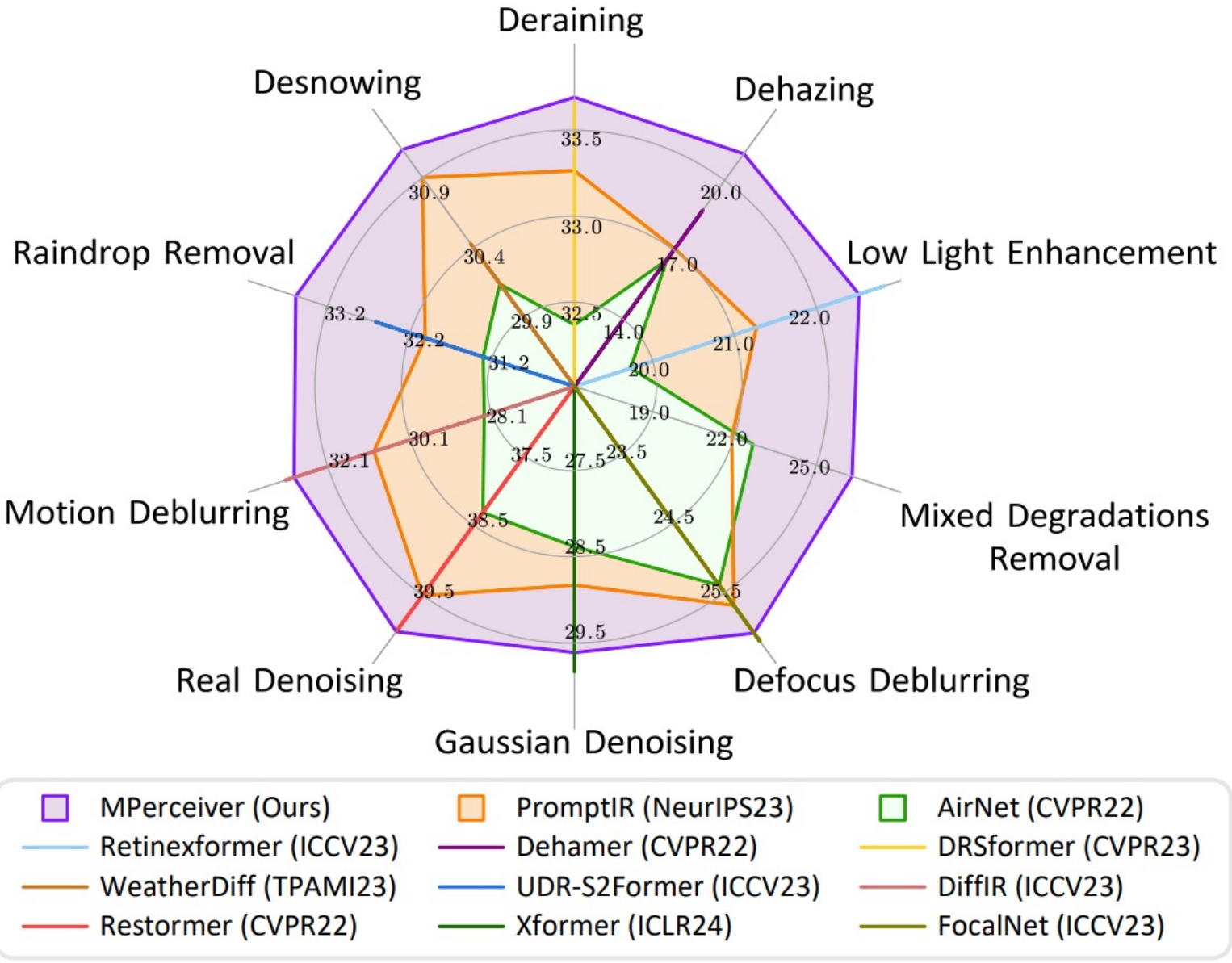


(a) Input (b) w/o DRM (c) w/ DRM

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

All-in-one Results

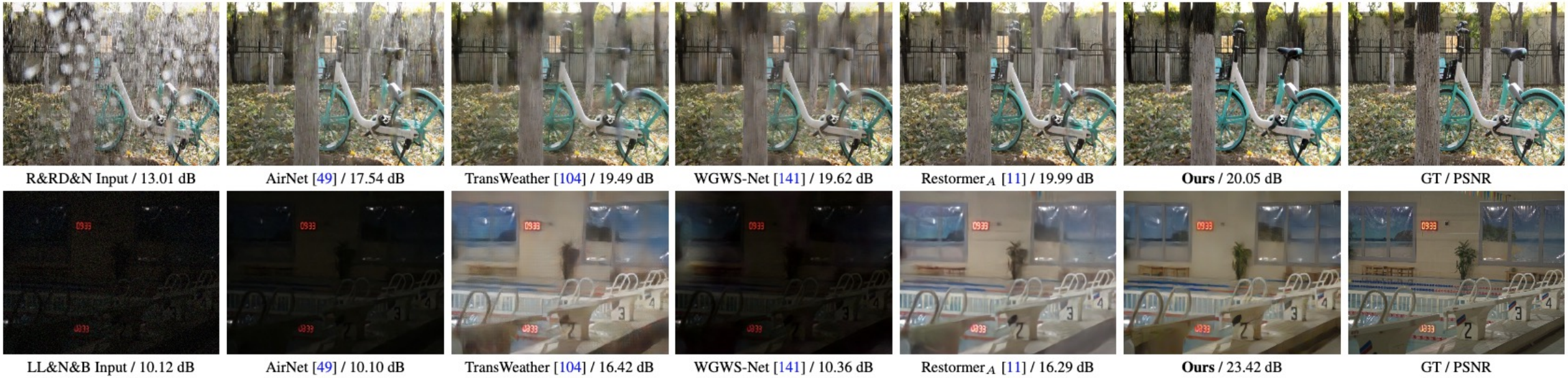


Type	Method	***Deraining*** (Rain1400)		Method	***Dehazing*** (Average)		Method	***Desnowing*** (Snow100K-L)	
		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓
Task Specific	Uformer [109]	32.84 / 0.931	23.31 / 0.061	AECRNet [113]	17.84 / 0.546	225.8 / 0.526	DesnowNet [68]	27.17/ 0.898	- / -
	Restormer [122]	33.68 / 0.939	20.33 / 0.050	SGID [7]	14.36 / 0.562	342.9 / 0.580	DDMSNet [129]	28.85 / 0.877	3.24 / 0.096
	DRSformer [14]	33.66 / 0.939	20.06 / 0.050	DeHamer [29]	18.64 / 0.622	241.6 / 0.488	DRT [59]	29.56 / 0.892	8.15 / 0.135
	UDR-S^2 [12]	33.08 / 0.930	19.89 / 0.053	MB-Taylor [85]	17.94 / 0.602	250.8 / 0.499	WeatherDiff [81]	30.43 / 0.915	2.81 / 0.100
All in One	AirNet [49]	32.36 / 0.928	22.38 / 0.058	AirNet [49]	16.48 / 0.589	219.9 / 0.479	AirNet [49]	30.14 / 0.907	3.92 / 0.105
	PromptIR [83]	33.26 / 0.935	22.59 / 0.058	PromptIR [83]	16.97 / 0.595	231.9 / 0.471	PromptIR [83]	30.91 / 0.913	3.79 / 0.100
	DA-CLIP [70]	29.67 / 0.851	35.01 / 0.116	DA-CLIP [70]	15.01 / 0.544	224.6 / 0.468	DA-CLIP [70]	28.31 / 0.862	3.11 / 0.098
	Restormer$_A$ [122]	33.09 / 0.933	24.14 / 0.061	Restormer$_A$ [122]	15.86 / 0.584	221.4 / 0.477	Restormer$_A$ [122]	30.98 / 0.914	4.54 / 0.104
	NAFNet$_A$ [11]	33.27 / 0.936	22.39 / 0.050	NAFNet$_A$ [11]	15.97 / 0.597	228.6 / 0.454	NAFNet$_A$ [11]	31.42 / 0.920	2.72 / 0.091
	MPerceiver(Ours)	33.40 / 0.937	17.82 / 0.049	**MPerceiver(Ours)**	20.95 / 0.644	196.8 / 0.437	**MPerceiver(Ours)**	31.02 / 0.916	2.31 / 0.087
	MPerceiver+(Ours)	33.69 / 0.940	17.36 / 0.047	**MPerceiver+(Ours)**	21.08 / 0.651	190.1 / 0.422	**MPerceiver+(Ours)**	31.11 / 0.918	2.14 / 0.085

Type	Method	***Raindrop Removal*** (RainDrop)		Method	***Low-light Enhance.*** (LOL-v2-Real)		Method	***Motion Deblur*** (GoPro)	
		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓
Task Specific	AttentGAN [84]	31.59 / 0.917	33.33 / 0.056	SNR [117]	21.48 / 0.849	58.76 / 0.159	MPRNet [121]	32.66 / 0.959	10.98 / 0.091
	Quan *et al.* [87]	31.37 / 0.918	30.56 / 0.065	SNR-SKF [114]	21.93 / 0.842	73.70 / 0.160	Restormer [122]	32.92 / 0.961	10.63 / 0.086
	IDT [116]	31.87 / 0.931	25.54 / 0.059	RQ-LLIE [67]	22.37 / 0.854	56.92 / 0.143	Stripformer [103]	33.08 / 0.962	9.03 / 0.079
	UDR-S^2 [12]	32.64 / 0.942	27.17 / 0.064	Retinexformer [9]	22.80 / 0.840	62.45 / 0.169	DiffIR [115]	33.20 / 0.963	9.65 / 0.081
All in One	AirNet [49]	31.32 / 0.925	33.34 / 0.073	AirNet [49]	19.69 / 0.821	55.43 / 0.151	AirNet [49]	28.31 / 0.910	15.31 / 0.122
	PromptIR [83]	32.03 / 0.938	35.75 / 0.073	PromptIR [83]	21.23 / 0.860	53.92 / 0.145	PromptIR [83]	31.02 / 0.938	17.54 / 0.131
	DA-CLIP [70]	30.44 / 0.880	29.38 / 0.078	DA-CLIP [70]	21.76 / 0.762	48.23 / 0.134	DA-CLIP [70]	27.12 / 0.823	16.81 / 0.136
	Restormer$_A$ [122]	31.75 / 0.936	38.22 / 0.075	Restormer$_A$ [122]	20.77 / 0.851	57.04 / 0.155	Restormer$_A$ [122]	30.59 / 0.934	14.56 / 0.115
	NAFNet$_A$ [11]	32.79 / 0.943	29.80 / 0.063	NAFNet$_A$ [11]	18.04 / 0.827	54.25 / 0.147	NAFNet$_A$ [11]	32.01 / 0.953	13.42 / 0.101
	MPerceiver(Ours)	33.21 / 0.929	21.27 / 0.051	**MPerceiver(Ours)**	22.16 / 0.848	45.90 / 0.130	**MPerceiver(Ours)**	32.49 / 0.959	10.69 / 0.089
	MPerceiver+(Ours)	33.62 / 0.930	19.37 / 0.044	**MPerceiver+(Ours)**	22.49 / 0.854	45.29 / 0.129	**MPerceiver+(Ours)**	32.98 / 0.961	10.51 / 0.087

Type	Method	***Defocus Deblur*** (DPDD)		Method	***Gaussian Denoising*** (Average)		Method	***Real Denoising*** (SIDD)	
		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓		PSNR / SSIM ↑	FID / LPIPS ↓
Task Specific	DRBNet [94]	25.73 / 0.791	49.04 / 0.183	SwinIR [58]	29.60 / 0.842	58.99 / 0.146	MPRNet [121]	39.71 / 0.958	49.54 / 0.200
	Restormer [122]	25.98 / 0.811	44.55 / 0.178	Restormer [122]	29.73 / 0.845	57.56 / 0.148	Uformer [109]	39.89 / 0.960	47.18 / 0.198
	NRKNet [88]	26.11 / 0.810	55.23 / 0.210	ART [126]	29.79 / 0.845	58.50 / 0.141	Restormer [122]	40.02 / 0.960	47.28 / 0.195
	FocalNet [17]	26.18 / 0.808	48.82 / 0.210	Xformer [125]	29.83 / 0.847	55.22 / 0.144	ART [126]	39.99 / 0.960	42.38 / 0.189
All in One	AirNet [49]	25.37 / 0.770	58.82 / 0.193	AirNet [49]	28.37 / 0.801	69.36 / 0.181	AirNet [49]	38.32 / 0.945	51.20 / 0.134
	PromptIR [83]	25.66 / 0.791	52.64 / 0.197	PromptIR [83]	28.82 / 0.816	63.76 / 0.170	PromptIR [83]	39.52 / 0.954	50.52 / 0.198
	DA-CLIP [70]	24.91 / 0.749	57.43 / 0.201	DA-CLIP [70]	25.13 / 0.692	59.82 / 0.235	DA-CLIP [70]	34.04 / 0.824	34.56 / 0.186
	Restormer$_A$ [122]	25.74 / 0.795	54.74 / 0.213	Restormer$_A$ [122]	28.65 / 0.812	63.48 / 0.172	Restormer$_A$ [122]	39.48 / 0.954	51.75 / 0.190
	NAFNet$_A$ [11]	25.85 / 0.803	48.45 / 0.191	NAFNet$_A$ [11]	29.21 / 0.829	60.84 / 0.163	NAFNet$_A$ [11]	39.76 / 0.957	45.54 / 0.197
	MPerceiver(Ours)	25.88 / 0.803	48.22 / 0.190	**MPerceiver(Ours)**	29.57 / 0.838	61.44 / 0.158	**MPerceiver(Ours)**	39.96 / 0.959	41.11 / 0.191
	MPerceiver+(Ours)	26.06 / 0.805	46.07 / 0.190	**MPerceiver+(Ours)**	29.61 / 0.839	60.91 / 0.156	**MPerceiver+(Ours)**	40.05 / 0.960	41.46 / 0.190

Mixed degradation benchmark MID6

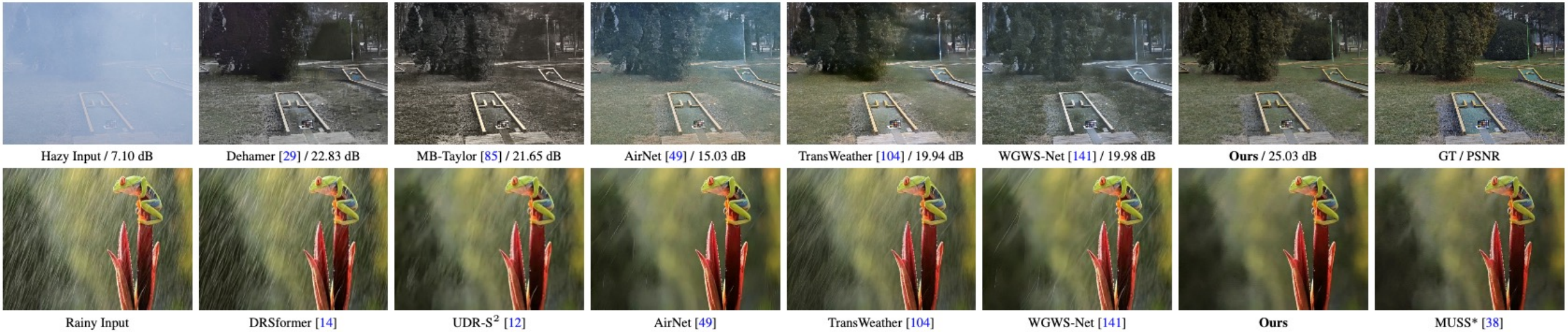


R&RD&N Input / 13.01 dB | AirNet [49] / 17.54 dB | TransWeather [104] / 19.49 dB | WGWS-Net [141] / 19.62 dB | Restormer$_A$ [11] / 19.99 dB | **Ours** / 20.05 dB | GT / PSNR

LL&N&B Input / 10.12 dB | AirNet [49] / 10.10 dB | TransWeather [104] / 16.42 dB | WGWS-Net [141] / 10.36 dB | Restormer$_A$ [11] / 16.29 dB | **Ours** / 23.42 dB | GT / PSNR

Method	Haze &Noise &Blur			Lowlight &Noise &Blur			Rain &Noise &Blur			Rain &Raindrop &Noise			Raindrop &Noise &Blur			Snow &Noise &Blur		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
AirNet [49]	17.51	0.613	0.440	17.09	0.552	0.527	24.31	0.601	0.364	21.46	0.570	0.471	26.44	0.745	0.353	22.62	0.566	0.454
TransWeather [104]	25.11	0.739	0.241	20.36	0.626	0.408	25.01	0.683	0.294	21.59	0.541	0.412	27.76	0.737	0.229	23.90	0.672	0.306
WGWS-Net [141]	17.66	0.617	0.394	17.57	0.570	0.448	22.10	0.600	0.353	20.12	0.522	0.446	25.67	0.718	0.307	20.03	0.569	0.401
PromptIR [83]	18.41	0.631	0.437	20.95	0.649	0.413	23.75	0.647	0.313	21.31	0.556	0.461	25.41	0.721	0.327	20.92	0.601	0.362
Restormer$_A$ [122]	17.03	0.602	0.470	16.49	0.541	0.498	23.22	0.611	0.332	20.39	0.561	0.493	24.48	0.697	0.401	21.39	0.606	0.392
NAFNet$_A$ [11]	16.59	0.548	0.541	15.72	0.605	0.520	23.48	0.563	0.346	22.72	0.599	0.439	27.20	0.769	0.307	20.28	0.535	0.484
MPerceiver (Ours)	26.19	0.782	0.211	23.84	0.671	0.343	26.00	0.762	0.193	22.35	0.525	0.268	28.49	0.771	0.127	24.36	0.719	0.263

Real-world datasets results



Type	*Deraining* (LHP [31])			*Deraining* (SSID [38])			*Desnowing* (RealSnow [141])			*Motion Deblur* (RealBlur-J [92]))		
	Method	PSNR ↑	SSIM ↑	Method	NIQE ↓	BRISQUE ↓	Method	PSNR ↑	SSIM ↑	Method	PSNR ↑	SSIM ↑
Task Specific	MUSS* [38]	30.02	0.886	MUSS* [38]	3.43	28.97	MIRNetv2 [123]	31.39	0.916	MPRNet [121]	28.70	0.873
	Restormer [122]	29.72	0.889	Restormer [122]	4.12	33.29	ART [126]	31.05	0.913	Restormer [122]	28.96	0.879
	DRSformer [14]	30.04	0.895	DRSformer [14]	4.19	35.52	Restormer [122]	31.38	0.923	Stripformer [103]	28.82	0.876
	UDR-S^2 [12]	28.59	0.884	UDR-S^2 [12]	3.77	35.86	NAFNet [11]	31.44	0.919	DiffIR [115]	29.06	0.882
All in One	AirNet [49]	31.73	0.889	AirNet [49]	3.69	30.91	AirNet [49]	31.02	0.923	AirNet [49]	27.91	0.834
	TransWeather [104]	29.87	0.867	TransWeather [104]	3.96	30.94	TransWeather [104]	31.13	0.922	TransWeather [104]	28.03	0.837
	WGWS-Net [141]	30.77	0.885	WGWS-Net [141]	3.71	30.79	WGWS-Net [141]	31.37	0.919	WGWS-Net [141]	28.10	0.838
	MPerceiver (Ours)	32.07	0.889	**MPerceiver** (Ours)	3.60	30.77	**MPerceiver** (Ours)	31.45	0.924	**MPerceiver** (Ours)	29.13	0.881

Zero-shot and Few-shot

- under-display camera IR and underwater IR

Table 4. **[Zero-shot]** ***UDC IR*** (TOLED / POLED) results.

Method	TOLED [140]			POLED [140]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
AirNet [49]	26.76	0.799	0.307	13.49	0.522	0.696
TransWeather [104]	27.58	0.810	0.316	15.86	0.590	0.707
WGWS-Net [141]	22.11	0.731	0.374	10.96	0.429	0.776
Restormer$_A$ [122]	27.74	0.841	0.294	13.94	0.528	0.681
NAFNet$_A$ [11]	27.90	0.848	0.320	10.68	0.555	0.713
MPerceiver (Ours)	32.92	0.863	0.161	20.41	0.650	0.445

Table 5. **[Zero-shot]** ***Underwater IR*** results.

Method	UIEB [50]			UWCNN[51]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
AirNet [49]	17.09	0.761	0.304	13.54	0.737	0.403
TransWeather [104]	17.17	0.754	0.303	13.59	0.731	0.408
WGWS-Net [141]	16.99	0.745	0.340	13.83	0.740	0.410
Restormer$_A$ [122]	17.34	0.770	0.300	13.49	0.737	0.401
NAFNet$_A$ [11]	17.31	0.736	0.307	13.62	0.736	0.405
MPerceiver (Ours)	22.69	0.902	0.150	14.77	0.774	0.299

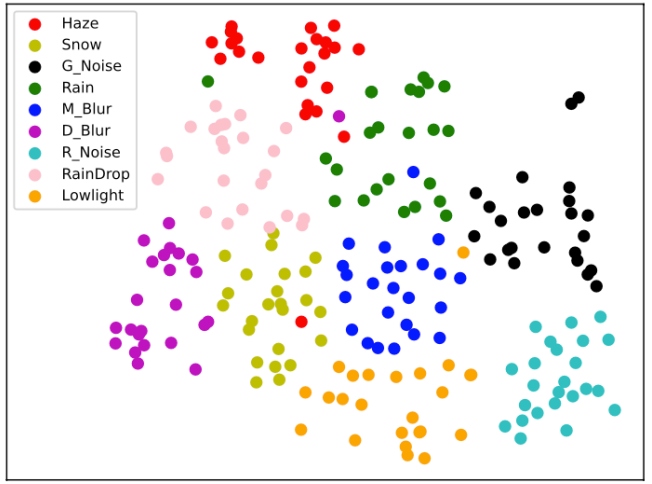
- 3%-5%
- JPEG compression artifact removal, demosaicking, demoireing

Type	Method	LIVE1 [98]		BSD500 [74]	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Task Specific	QGAC [20]	27.62	0.804	27.74	0.802
	FBCNN [44]	27.77	0.803	27.85	0.799
All-in-One	AirNet [49]	27.47	0.797	27.60	0.788
	TransWeather [104]	26.45	0.755	26.68	0.785
	WGWS-Net [141]	26.50	0.750	26.60	0.741
	MPerceiver (Ours)	27.79	0.804	27.88	0.795

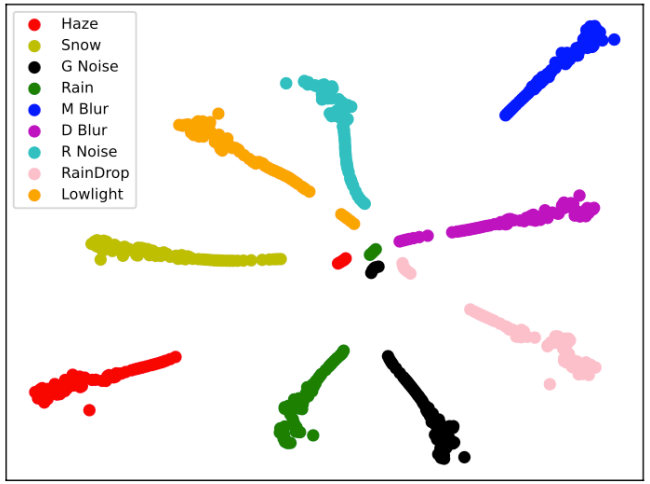
Datasets	RLDD [30]	RNAN [133]	DRUNet [130]	AirNet [49]	TransWeather [104]	WGWS-Net [141]	**MPerceiver** (Ours)
Kodak [23]	42.49	43.16	42.68	40.55	39.58	41.22	43.06
McMaster [131]	39.25	39.70	39.39	37.36	36.68	38.06	39.68

Type	Method	Venue	PSNR ↑	SSIM ↑
Task Specific	MBCNN [135]	*CVPR ' 20*	30.03	0.893
	FHDe^{2}Net [32]	*ECCV ' 20*	27.78	0.896
	WDNet [62]	*ECCV ' 20*	28.08	0.904
	ESDNet [120]	*ECCV ' 22*	30.11	0.920
All-in-One	AirNet [49]	*CVPR ' 22*	28.59	0.866
	TransWeather [104]	*CVPR ' 22*	27.68	0.848
	WGWS-Net [141]	*CVPR ' 23*	28.13	0.861
	MPerceiver (Ours)	-	30.19	0.885

Ablation Study



(a) t-SNE from visual prompt



(b) t-SNE from CLIP textual embeddings

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline (Stable Diffusion)	16.49	0.481	0.538
+Textual Branch	19.19	0.557	0.447
+Textual Branch w/o TP Pool	18.94	0.553	0.457
+Visual Branch	25.05	0.763	0.213
+Visual Branch w/o VP Pool	24.94	0.760	0.216
+(Visual & Textual) Branch	25.31	0.770	0.199
+(Visual & Textual) Branch w/o TP Pool	25.20	0.768	0.206
+(Visual & Textual) Branch w/o VP Pool	25.13	0.767	0.204
+(Visual & Textual) Branch + DRM (Full Model)	29.17	0.842	0.162

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- Introduce multimodal **prompt learning approach** utilizing **Stable Diffusion** priors for enhanced adaptiveness, generalizability, and fidelity in all-in-one image restoration.

- Propose novel **dual-branch module**, comprising the cross-modal adapter and image restoration adapter, learns **holistic and multiscale detail** representations.

- Across 16 image restoration tasks, including all-in-one, zero-shot, and few-shot scenarios, MPerceiver demonstrates superior adaptiveness, generalizability, and fidelity.