Learning A Low-Level Vision Generalist via Visual Task Prompt

ACM MM 24

Xiangyu Chen
University of Macau; Shanghai AI Laboratory; Shenzhen Institute of Advanced Technology, CAS

Yihao Liu
Shanghai AI Laboratory; Shenzhen Institute of Advanced Technology, CAS

Yuandong Pu
Shanghai Jiao Tong University, Shanghai AI Laboratory

Wenlong Zhang
Shanghai AI Laboratory, The Hong Kong Polytechnic University

Jiantao Zhou*
State Key Laboratory of Internet of Things for Smart City, University of Macau

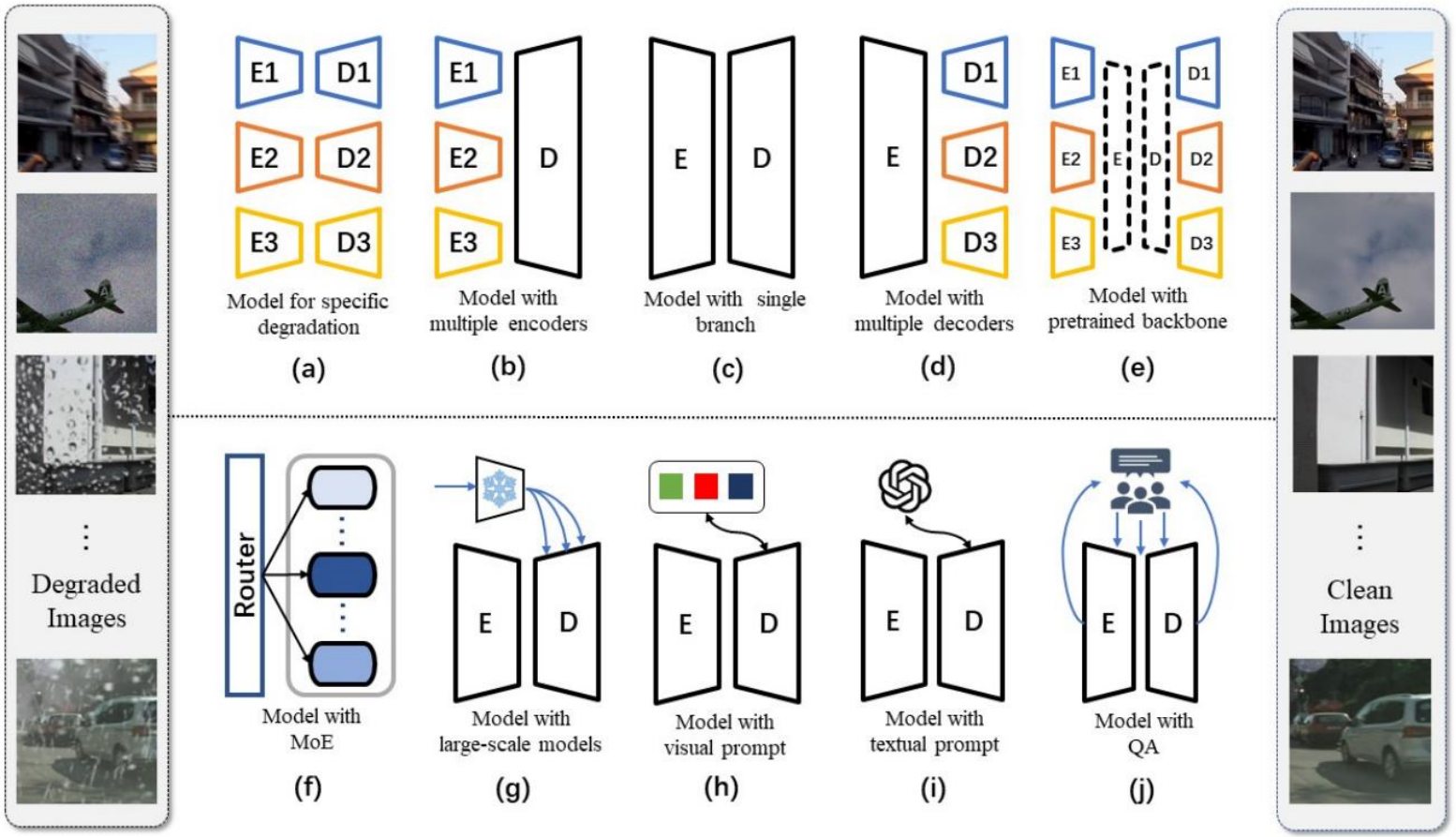
Yu Qiao
Shanghai AI Laboratory; Shenzhen Institute of Advanced Technology, CAS

Chao Dong*
Shanghai AI Laboratory; Shenzhen Institute of Advanced Technology, CAS; Shenzhen University of Advanced Technology

Presented by Zejia Fan
2025.5.18

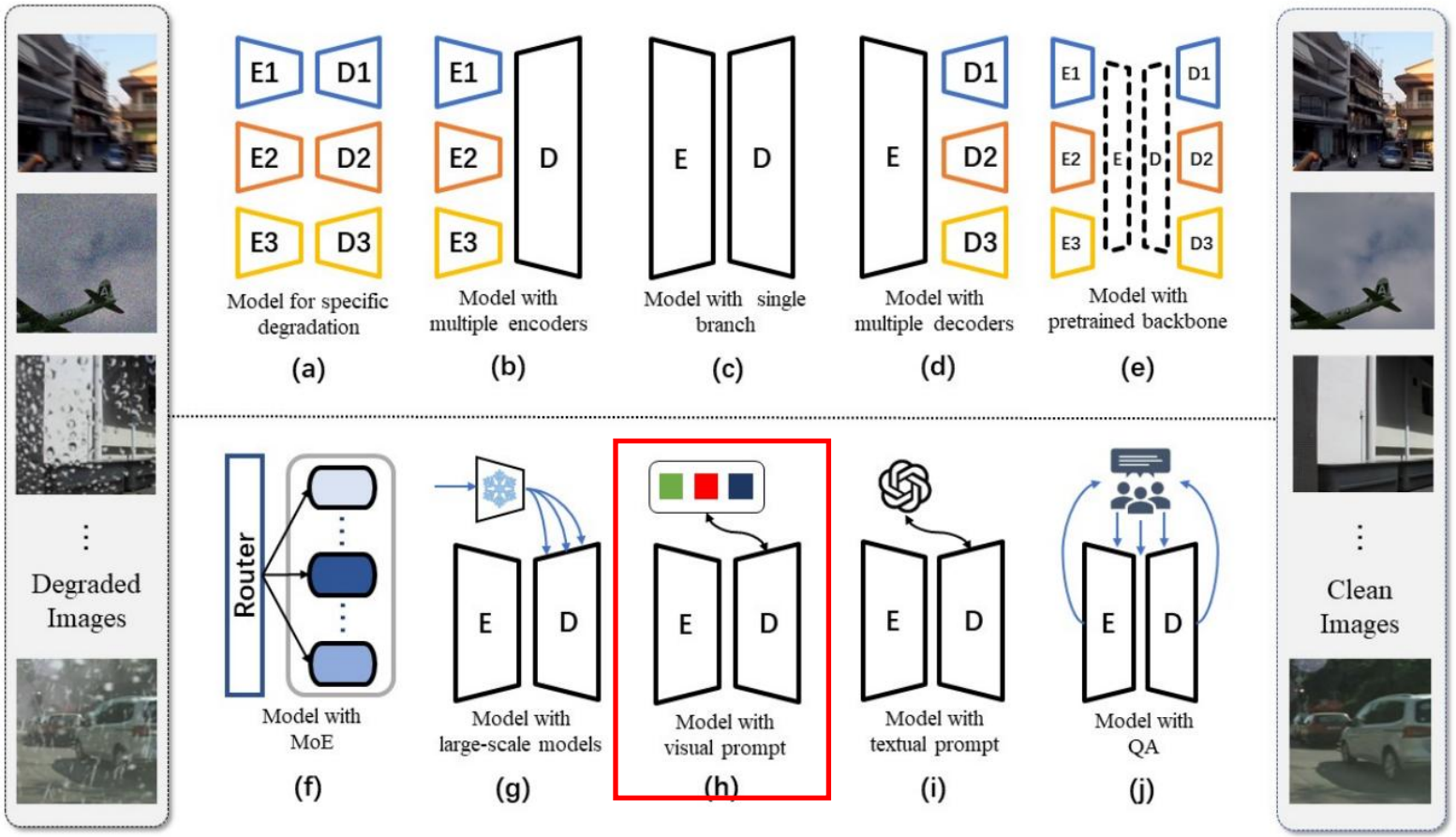
Background

- All-in-one restoration



Background

- All-in-one restoration

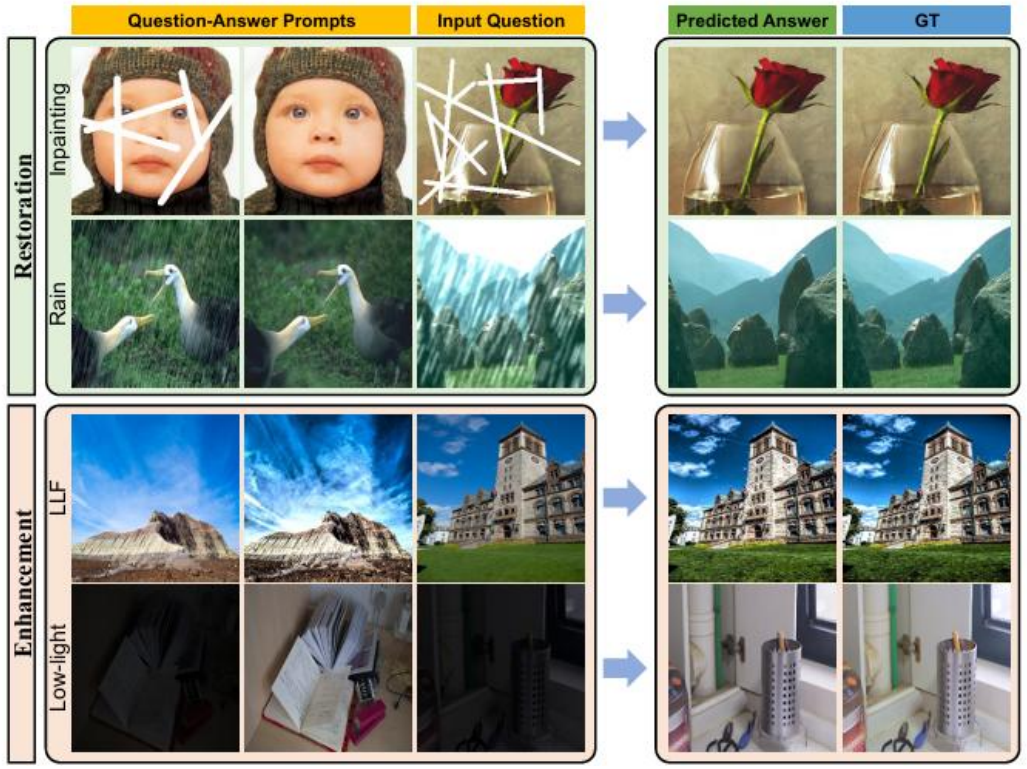


Background

- In context prompting in CV?

Background

- In context prompting in CV?



Background

- In context prompting in CV?
- 怀旧一下

23/02/19	Dezhao Wang	NeurIPS22	Visual Prompting via Image Inpainting	PDF

Visual Prompting via Image Inpainting

Amir Bar* [1,2], **Yossi Gandelsman*** [1], **Trevor Darrell**[1], **Amir Globerson**[2], **Alexei A. Efros**[1]

[1]UC Berkeley [2]Tel Aviv University

Background-Prompting

✅ **高精度**：通过参数更新充分适配任务
❌ **依赖标注数据**：需大量任务特定标注
❌ **计算成本高**：每个任务需独立微调

- Pretrain-finetune
  - Finetuning the pretrained model with downstream data&labels
- Pretrain-prompt-predict
  - No longer need finetuning
  - Using prompting to predict
    - E.g. Sentiment prediction. “I love this moive.”
    - Prompt: “I love this moive. Overall, it was a [Z] movie.” We predict sentiment by[Z]

Background-Prompting

✅ **少样本适配**：仅需少量示例调整提示
✅ **参数高效**：模型主体参数冻结，仅优化提示部分
❌ **提示设计敏感**：依赖模板质量或提示学习算法

- Learned prompt
  - Human-designed prefix is time-consuming/not robust
  - Learning a prefix with few-shot data
    - “I love this moive. Overall, it was a [Z] moive.”
    - “I love this moive.[P][Z].” [P] is tuned on the downstream data.

Background

- In context prompting

Task: Sentiment Analysis

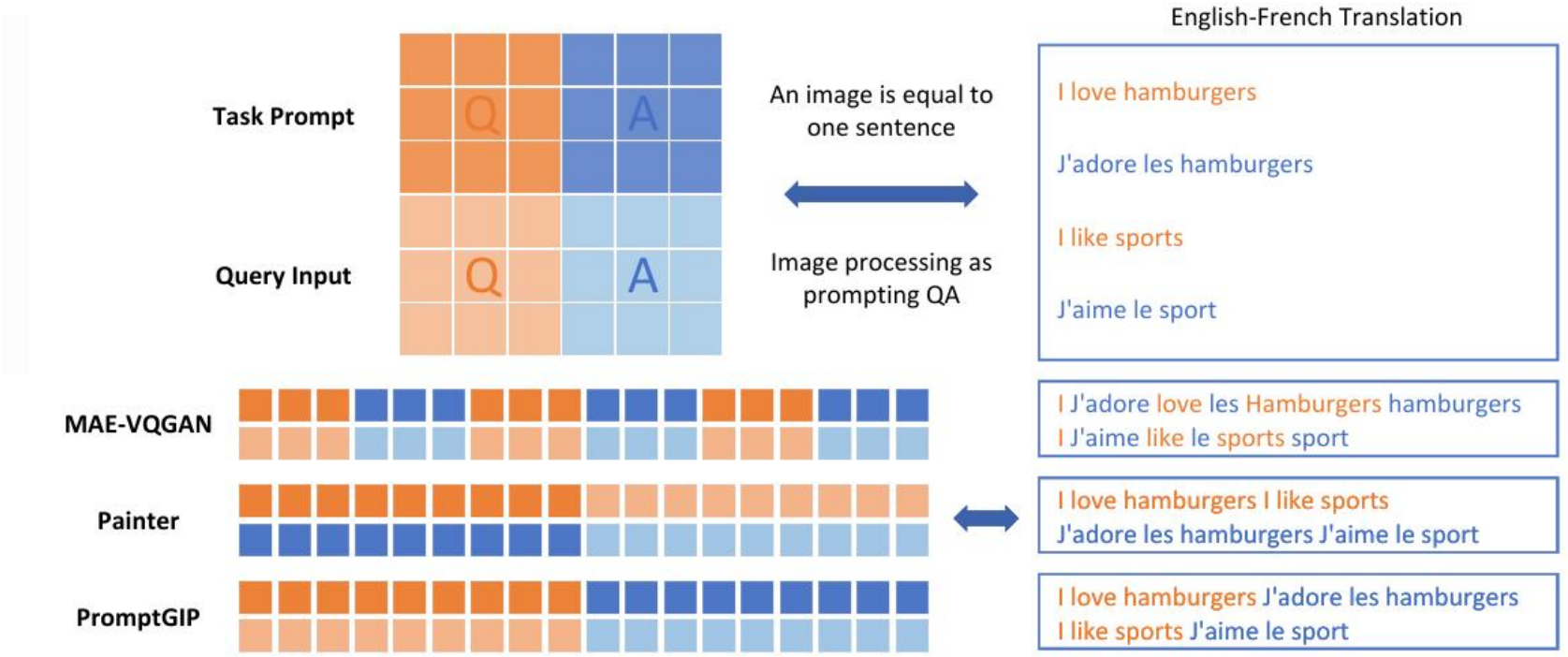
Input: "This movie is fantastic!" → **Output**: Positive

Input: "The service was terrible." → **Output**: Negative

New Input: "The visuals are stunning but the plot is dull." → **Output**: ?

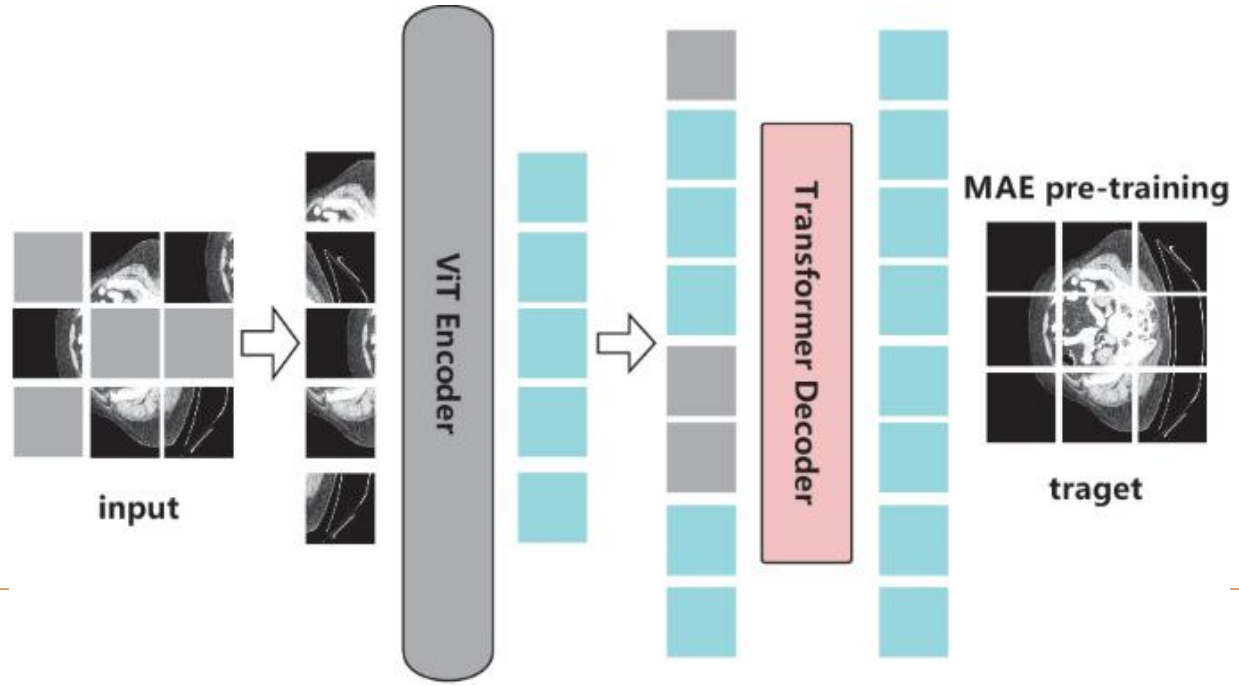
Background

- In context prompting

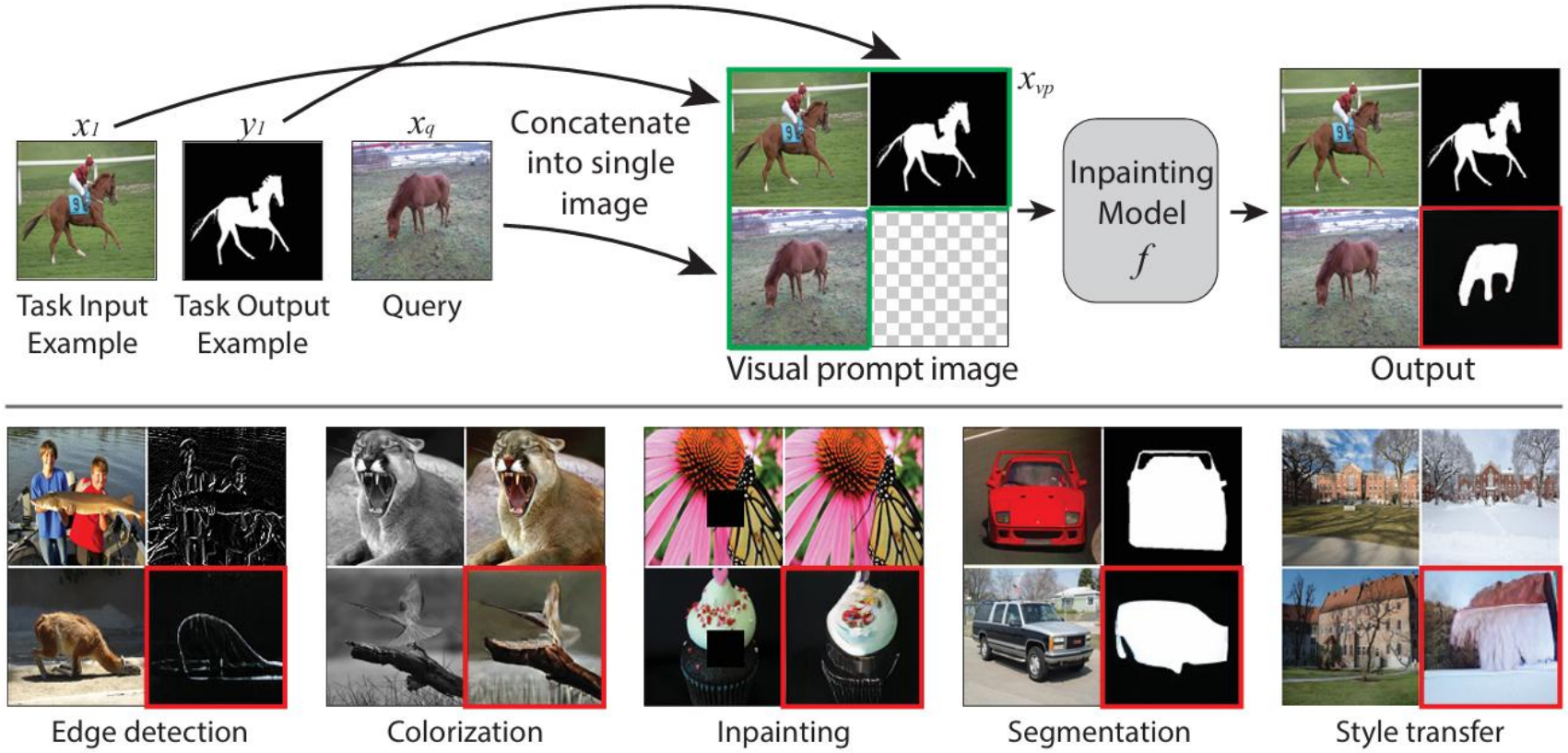


Background-MAE

- reconstruct the masked portions by randomly masking input image patches, thereby learning powerful visual representations.



Visual Prompting via Image Inpainting (NIPS 2022)



Visual Prompting via Image Inpainting (NIPS 2022)

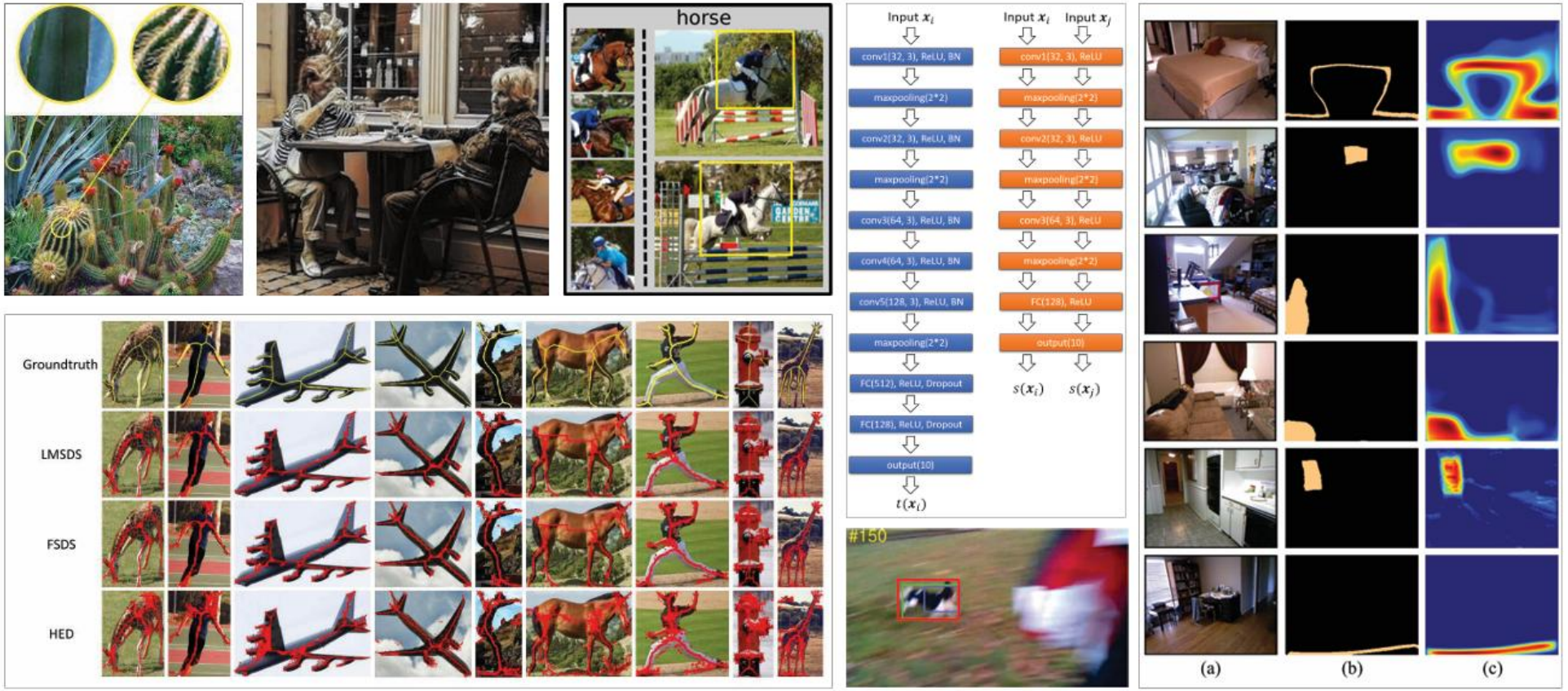
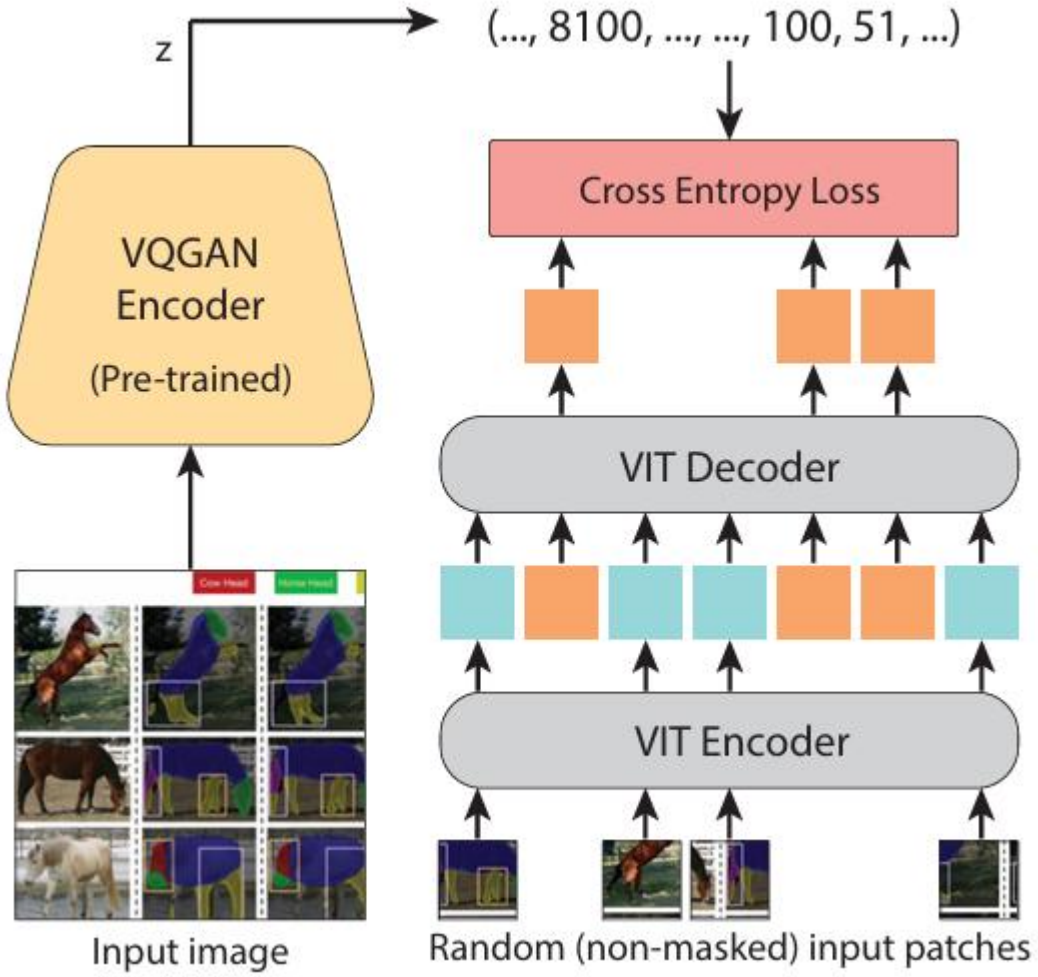


Figure 3: **Random images from our Computer Vision Figures dataset**. We curated a dataset of 88k unlabeled figures from Computer Vision academic papers. During training, we randomly sample crops from these figures, without any additional parsing.

Visual Prompting via Image Inpainting (NIPS 2022)



Visual Prompting via Image Inpainting (NIPS 2022)

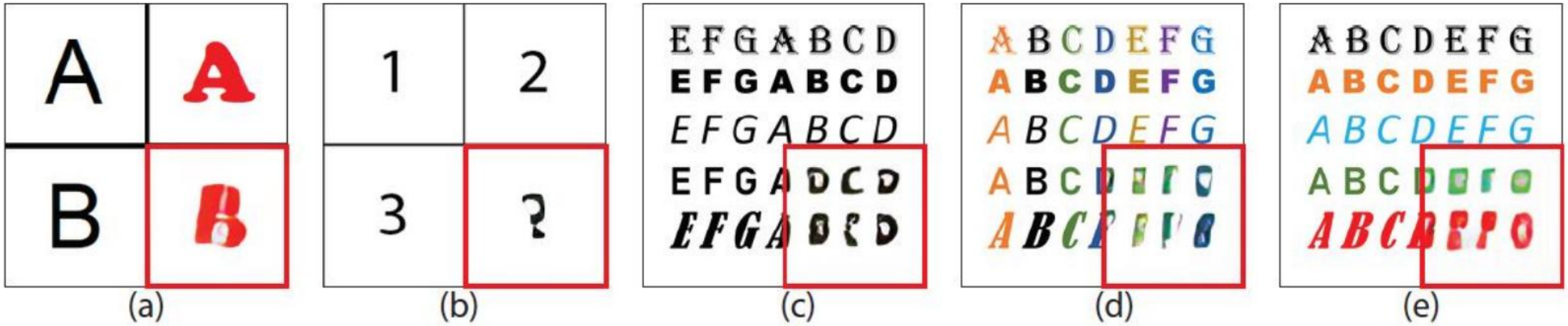
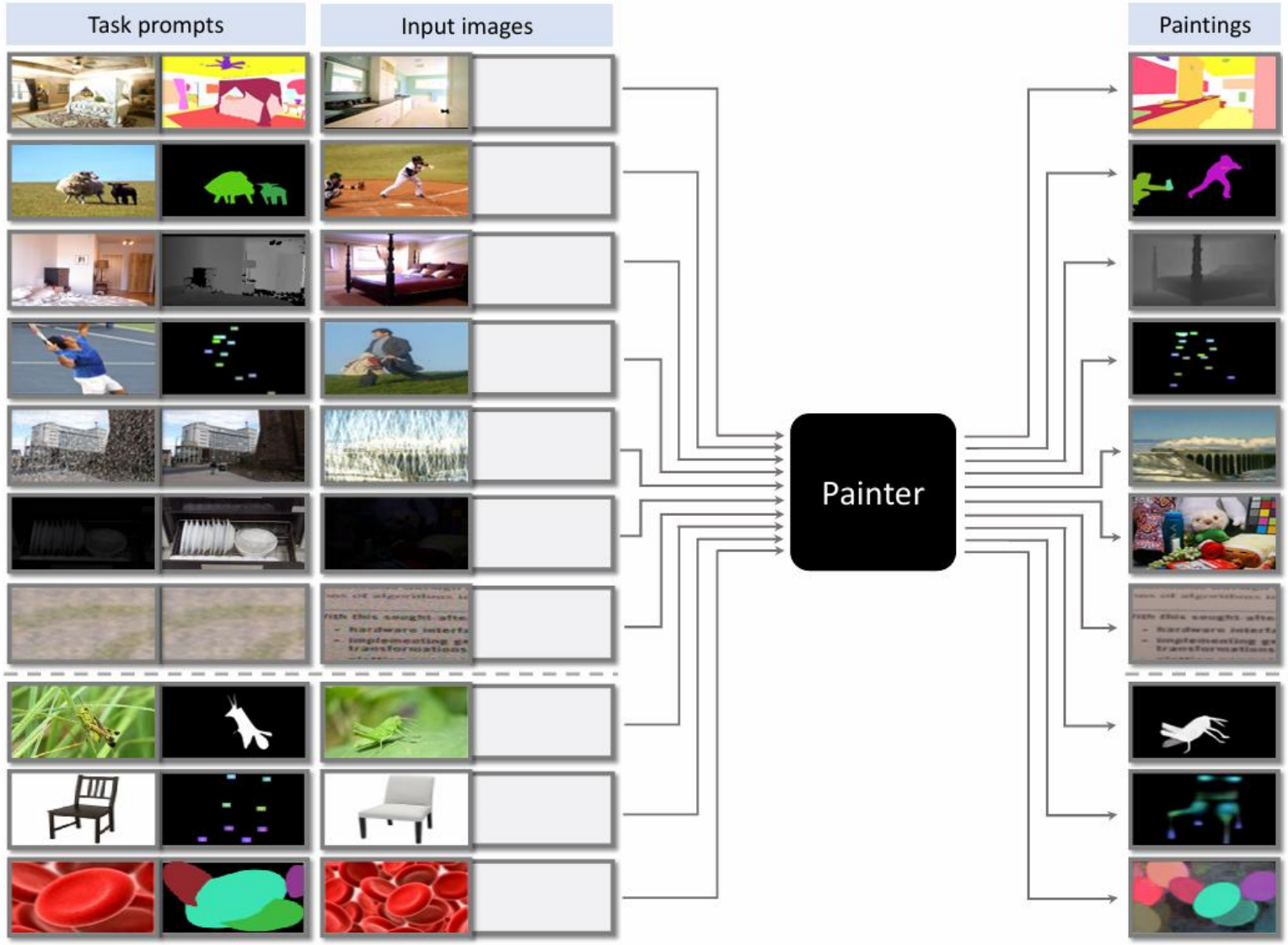


Figure 10: **Style and content extrapolation using MAE-VQGAN**. The model can extrapolate the style of a new content (a), but fails to predict a new content (b). The model struggles to extrapolate new style and content of longer sequences (c-e).

Images Speak in Images [CVPR 2023]



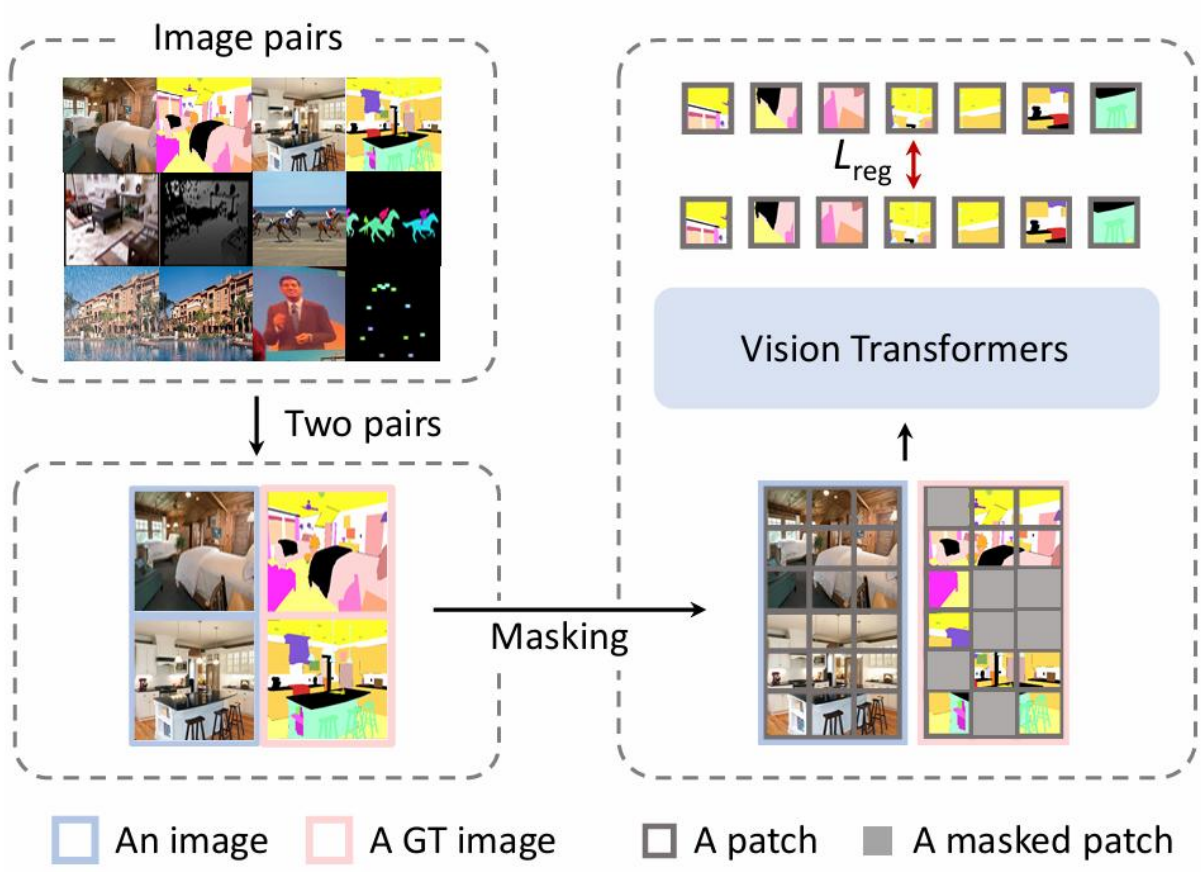
Images Speak in Images: A Generalist Painter for In-Context Visual Learning

Xinlong Wang[1*] Wen Wang[2*] Yue Cao[1*] Chunhua Shen[2] Tiejun Huang[1,3]

[1] Beijing Academy of Artificial Intelligence [2] Zhejiang University [3] Peking University

out-of-domain vision tasks

Images Speak in Images [CVPR 2023]



Images Speak in Images

- Multi-task dataset

任务类型	数据集	数据规模
深度估计	NYUv2	训练集：24K图像 测试集：654图像（215场景）
语义分割	ADE20K	训练集：20K 验证集：2K 测试集：3K
全景分割	MS-COCO	训练集：118K 验证集：5K
人体关键点检测	COCO-Person	训练集：15K 验证集：标准划分

Tasks	Deraining							Enhance.	Denoising
Datasets	Rain14000 [15]	Rain1800 [49]	Rain800 [53]	Rain100H [49]	Rain100L [49]	Rain1200 [52]	Rain12 [29]	LoL [45]	SIDD [1]
Train Samples	11200	1800	700	0	0	0	12	485	320
Test Samples	2800	0	100	100	100	1200	0	15	40
Testset Rename	Test2800	-	Test100	Rain100H	Rain100L	Test1200	-	-	-

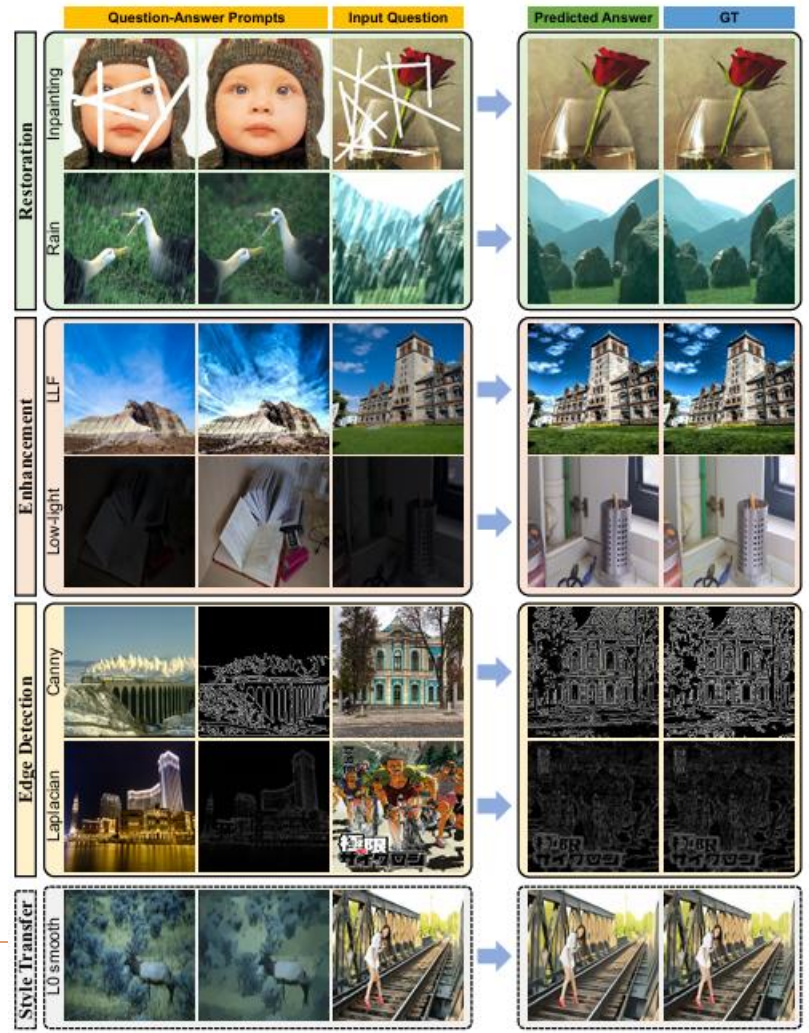
Table S1. Dataset description for image restoration tasks.

PromptGIP [ArXiv 2023]

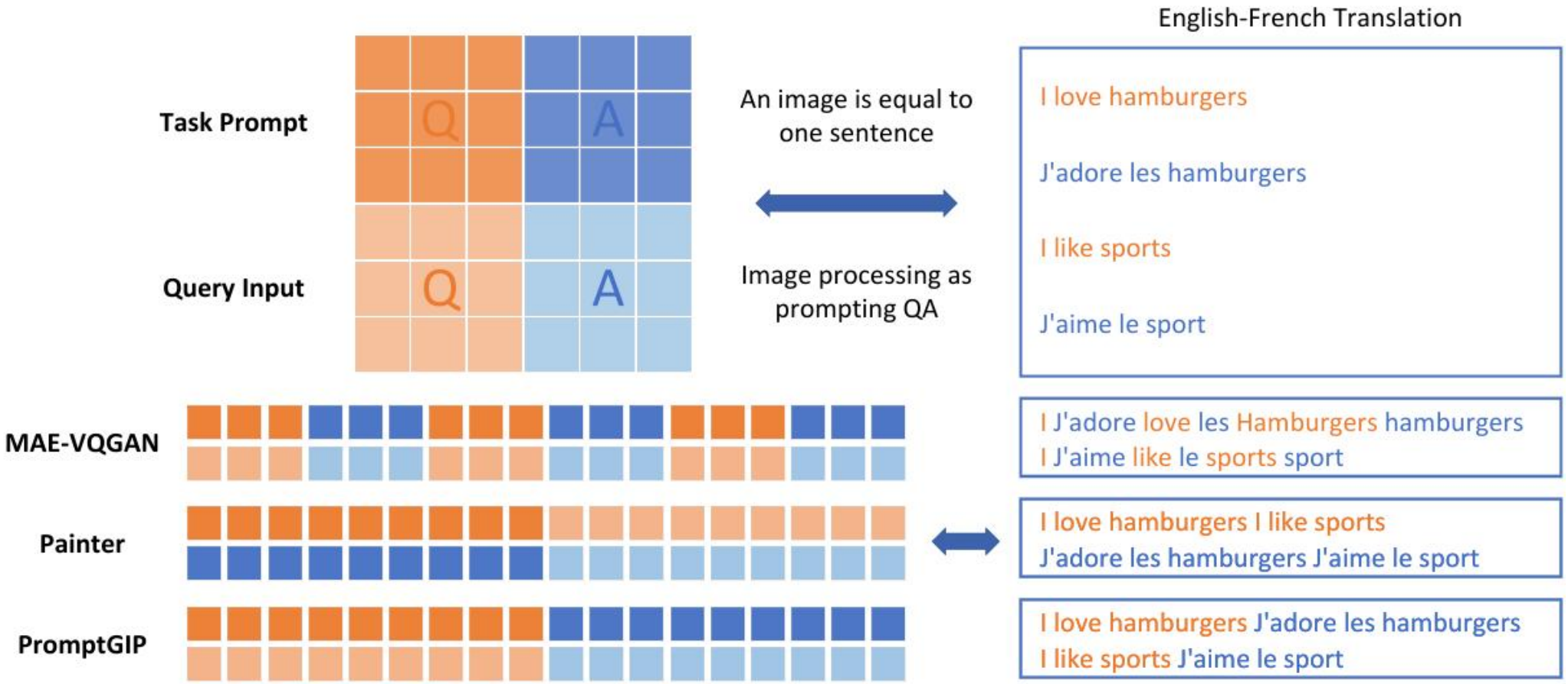
Unifying Image Processing as Visual Prompting Question Answering

Yihao Liu [*1,2] Xiangyu Chen [*1,2,3] Xianzheng Ma [*1] Xintao Wang [4] Jiantao Zhou [3] Yu Qiao [1,2] Chao Dong [2,1]

- 15 image processing tasks

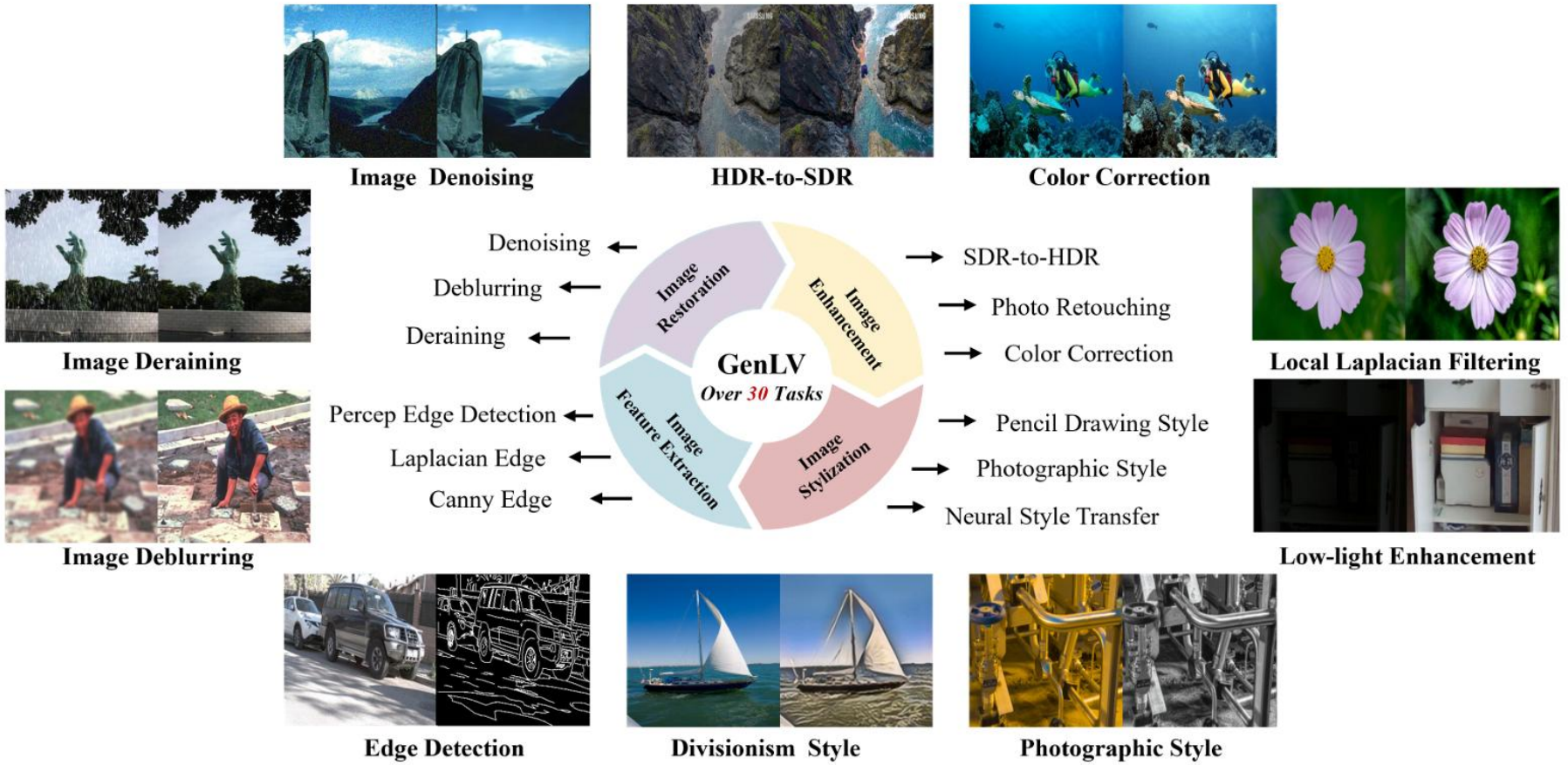


PromptGIP [ArXiv 2023]



GenLV [ACM MM 2024]

- Over 30 tasks



GenLV [ACM MM 2024]

- 现有图像复原方法无法处理边缘检测或风格化等跨域任务
- 基于MAE、ViT的模型可以处理跨域任务，但对低频信息处理不佳

GenLV [ACM MM 2024]

- 网络结构：
  - Unet形式, prompt作为control information
  - Transposed Attention等，不局限于MAE与ViT结构

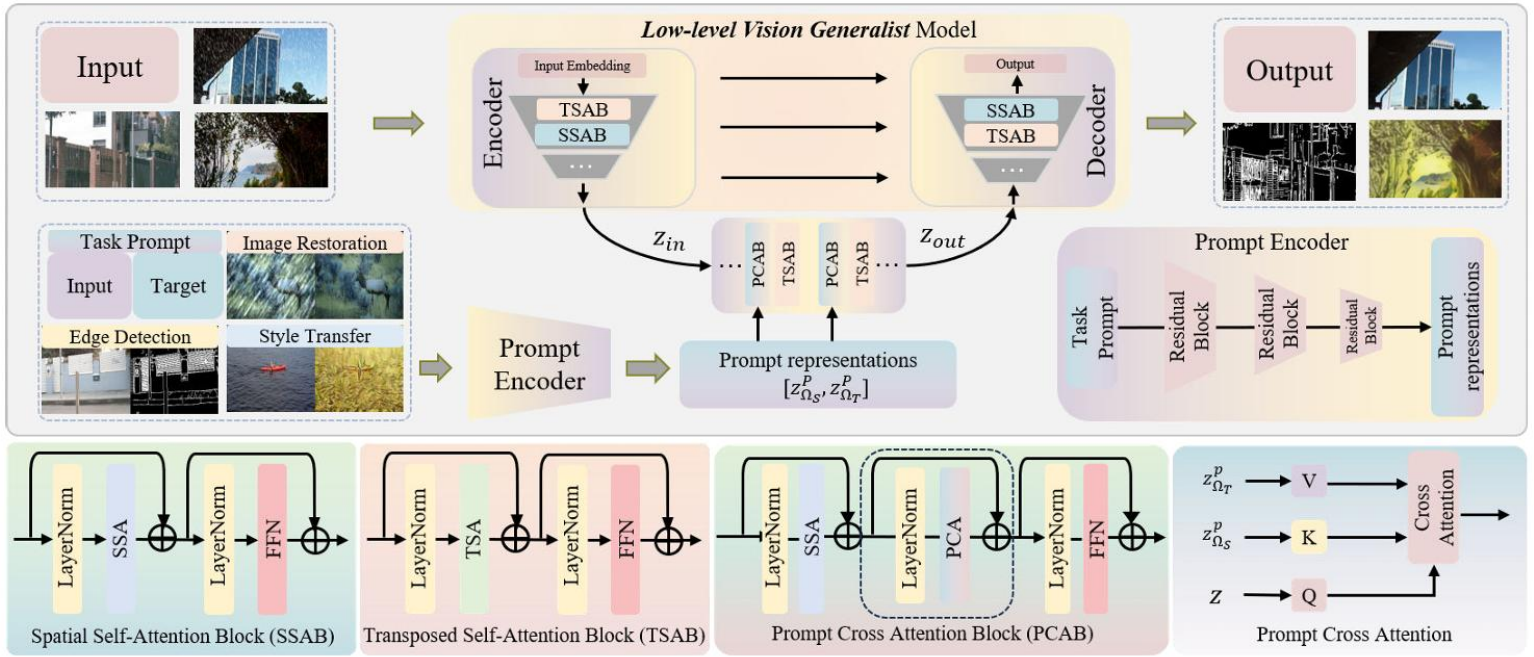
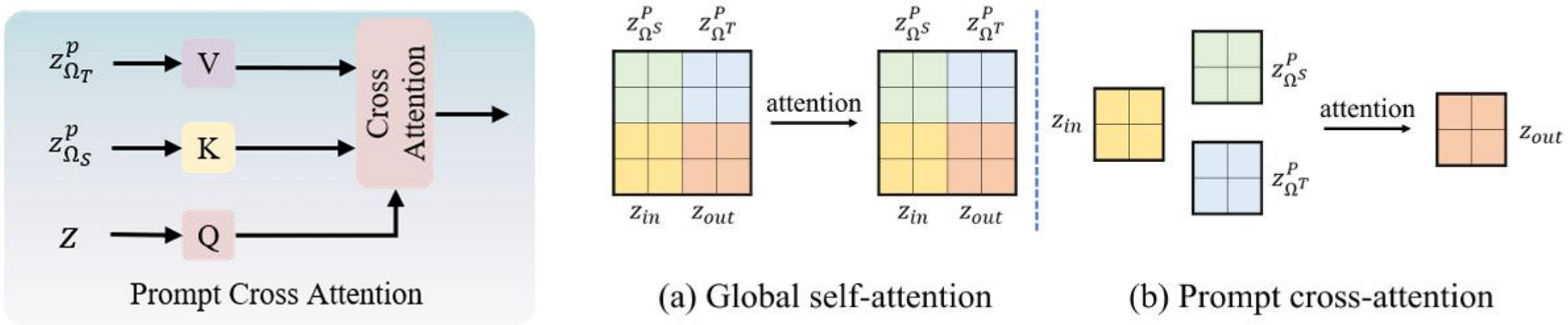


Figure 3: Overall approach of our low-level vision generalist model, GenLV.

Prompt Cross Attention



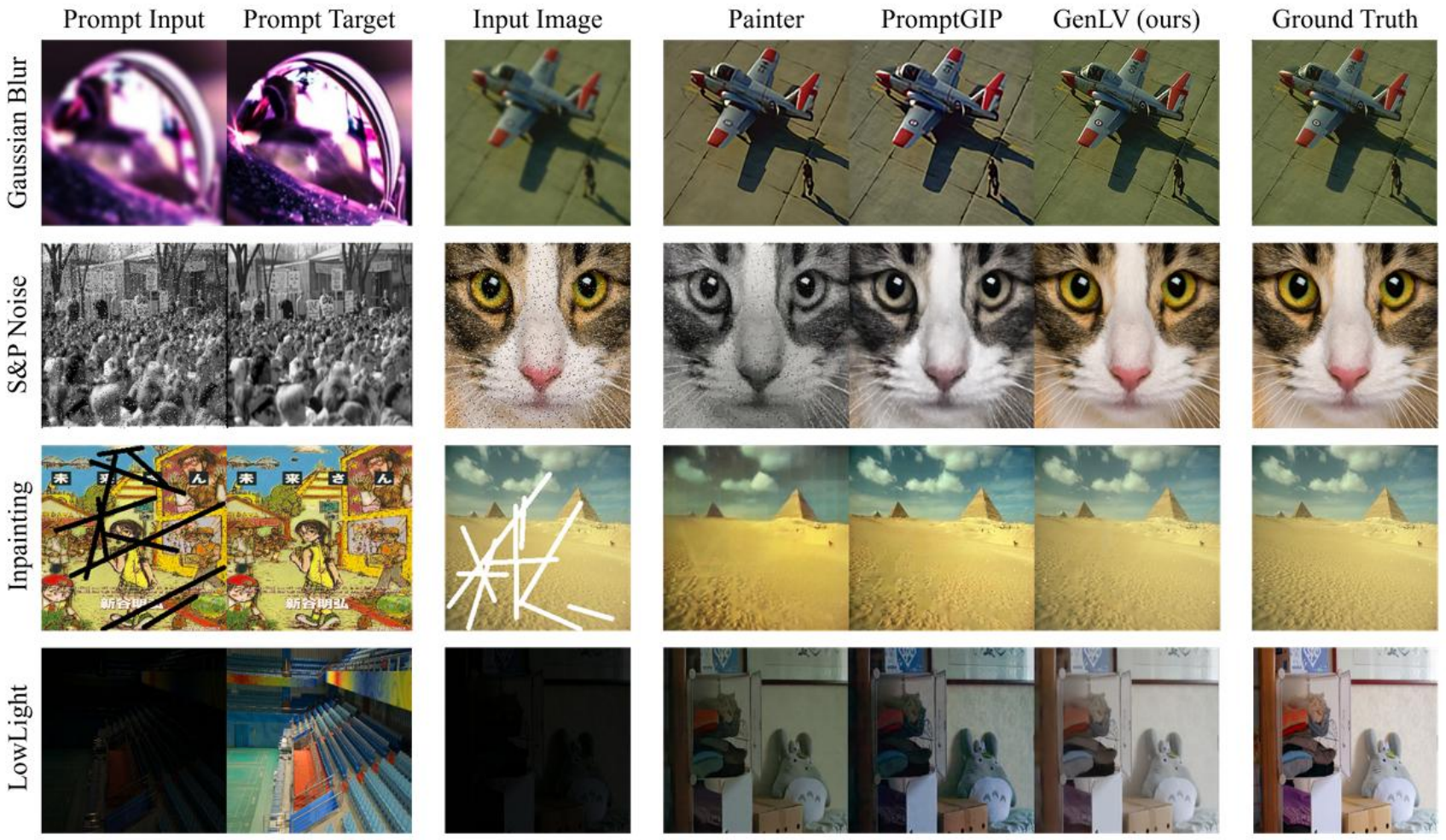
Experiment

Table 1: Quantitative results on image restoration tasks. #: public released model. ⋆: trained with only restoration tasks. †: trained with all 30 low-level vision tasks. GN: Gaussian noise. PN: Poisson noise. ViT-VPIP: ViT backbone adopted in the VPIP framework. Our GenLV can also be represented as X-Restormer-VPIP. PSNR↑ (dB) is calculated as the quantitative metric.

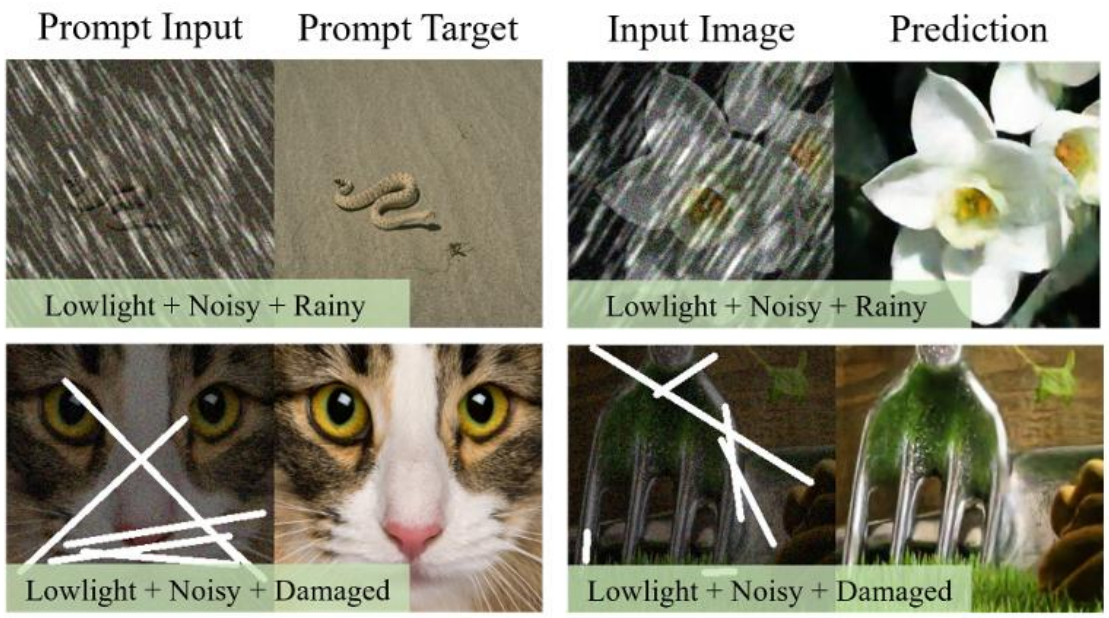
	GN	PN	S&P Noise	GB	JPEG	Ringing	R-L	Inpainting	SimpleRain	ComplexRain	Haze
Real-ESRGAN#	25.38	26.57	21.50	21.49	25.21	24.64	21.71	14.06	16.10	21.01	11.86
PromptIR⋆	28.86	31.48	36.45	24.56	26.77	**27.85**	31.31	**28.11**	30.76	24.08	16.85
PromptGIP⋆	26.48	27.76	28.08	22.88	25.86	25.69	27.05	25.28	25.79	24.33	24.55
ViT⋆	24.67	25.39	23.71	22.17	24.76	23.89	24.09	23.11	23.21	23.04	24.91
ViT-VPIP⋆	26.14	27.20	25.43	24.13	26.19	25.98	26.98	25.03	25.51	**24.79**	24.06
X-Restormer⋆	28.70	31.36	35.33	24.13	26.68	26.88	30.01	27.68	29.65	24.39	16.73
GenLV⋆ (ours)	**28.99**	**31.69**	**36.63**	**24.58**	**26.91**	27.74	**31.50**	**28.11**	**31.10**	24.71	**28.91**
Painter†	24.28	24.41	24.93	21.55	22.30	23.58	24.36	22.52	22.42	23.14	20.20
PromptGIP†	23.63	23.98	25.05	20.84	22.21	23.86	24.94	22.11	23.16	21.79	21.90
ViT-VPIP†	25.30	26.15	24.41	22.74	25.35	24.62	25.24	23.73	24.00	23.70	24.04
GenLV† (ours)	**28.49**	**31.05**	**34.20**	**23.39**	**26.21**	**25.78**	**28.21**	**27.17**	**28.18**	**25.11**	**29.70**

Experiment

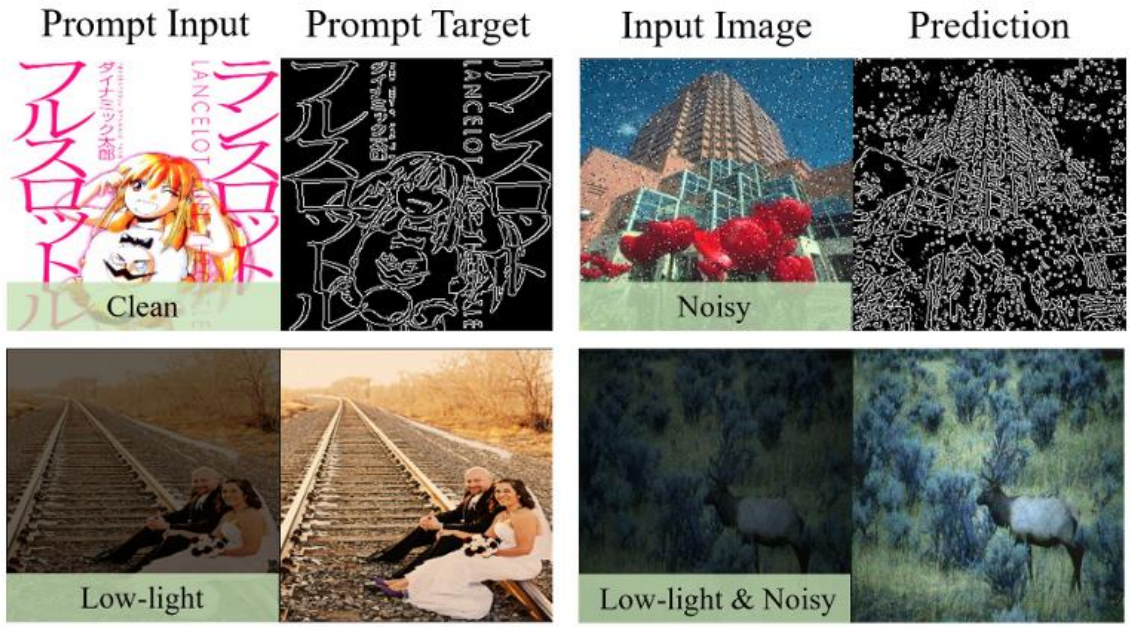
- 对比方法
  - 受promp影响大
  - 低频不清晰
  - 颜色异常



Experiment



(a) Results for mixed degraded images.



(b) Results for images based on cross-domain prompts.

Experiment

(c) Results for mixed degraded images on single-task prompts.

Experiment

Table 4: Standard deviation of the performance computed based on 20 different prompt images. PSNR (dB) is calculated as the quantitative metrics.

	GN	GB	LowLight	ICC	PencilDraw	RTV
Painter†	2.3930	1.8845	1.8865	1.9573	1.1820	2.6163
PromptGIP†	3.1035	2.2893	0.6766	0.6311	1.4200	1.3130
GenLV†	**0.1033**	**0.0208**	**0.0399**	**0.0512**	**0.5518**	**0.0195**

Experiment

Table 5: Computation cost of different parts of our model.

Component	Params	MACs
Main Network	27.6M	166.9G
Prompt Encoder & Fusion	5.1M	35.7G

总结

- In-Context Prompting for vision
- GenLV:
  - Model Structure
  - Prompt Cross Attention
  - Lack: ability for unseen

Thanks for your listening!