STRUCT Group Paper Reading

LM4LV: AFrozen Large Language Model for Low-level Vision Tasks

arXiv 2024

Boyang Zheng *
Shanghai Jiao Tong University
bytetriper@sjtu.edu.cn

Jinjin Gu
Shanghai AI Laboratory
jinjin.gu@sydney.edu.au

Shijun Li
Nanjing University
shijun_lee@outlook.com

Chao Dong
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
Shanghai AI Laboratory
chao.dong@siat.ac.cn

Presented by Zejia Fan
2023.6.2

Author

Boyang Zheng

- Junior-year student, Shanghai Jiao Tong University
- Shanghai AI Lab, advised by Chao Dong.

Author

Jinjin Gu

- Education Experience
  - Ph.D. from the University of Sydney
  - Supervised by Prof. Wanli Ouyang and Prof. Luping Zhou
- Research Interest
  - computer vision and image processing
- Cite 7866 (ESRGAN 四作)

Author

Byung Hyun Lee

- Education Experience
  - Junior student in Nanjing University
  - A research intern at XPixel group

Author

Chao Dong

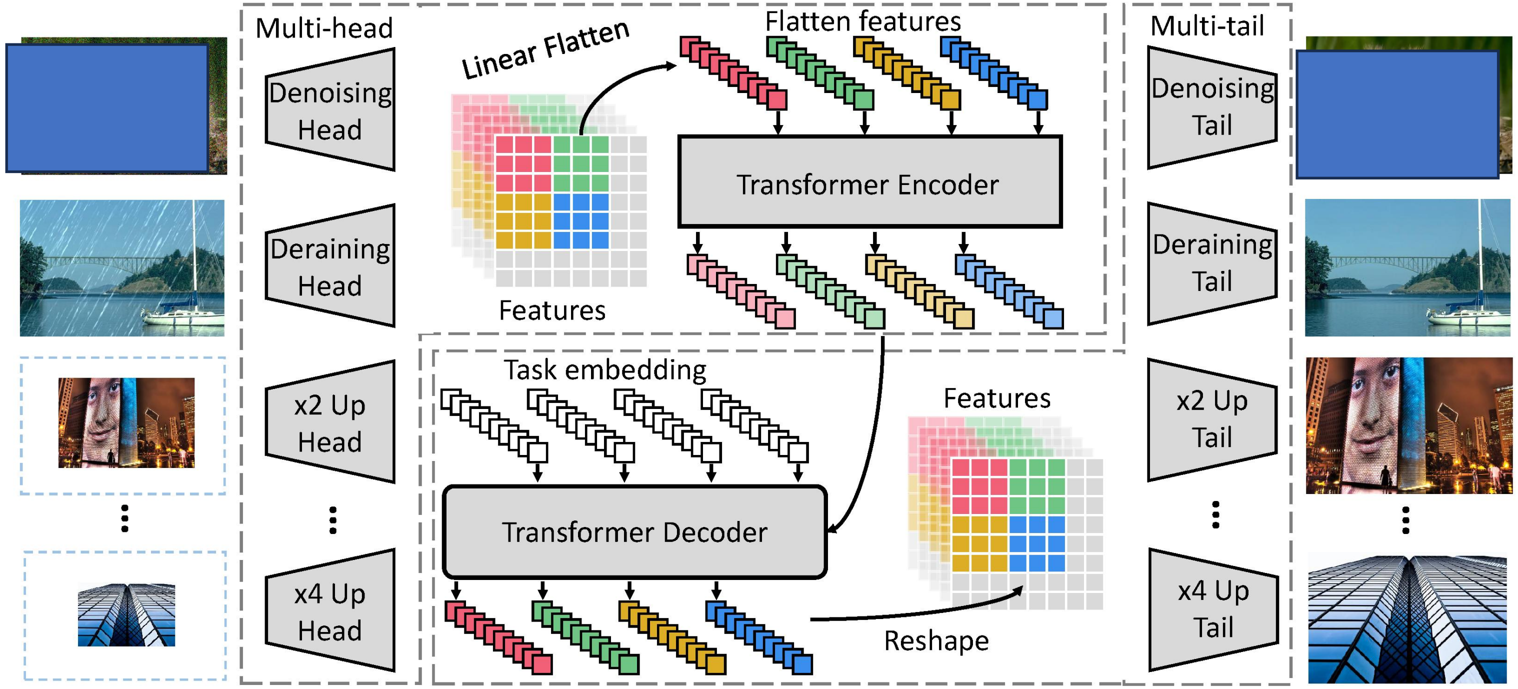
- Work Experience
  - Professor in Shenzhen Institute of Advanced Technology
- Education Experience
  - Ph.D. degree from The Chinese University of Hong Kong
- Research Interest
  - low-level vision problems
- Cite 35631

Author

XPixel Metaverse
Our mission is to make the world look clearer and better !
Single-Image SR
SRCNN
ECCV 14, PAMI 15
1st SR Deep Model
FSRCNN
ECCV 16
ARCNN
ICCV 15
RL-Restore
CVPR 18
Path-Restore
PAMI 21
SFTGAN
CVPR 18
ESRGAN
ECCVW 18
PIRM 18 Perceptual SR Champion
RankSRGAN
ICCV 19 Oral, PAMI 21
Image Super Resolution (SR)
Blind SR
CinCGAN
CVPRW 18
1st Blind SR Model
Blind
FAIG
NeurIPS 21 Spotlight
SRGA
1st SR Generalization Index
DDR
Dropout in SR
CVPR 22
Interpretation
LAM
CVPR 21
1st SR Interpretation Method
Interpretation
Low-Level Vision
Modulation
AdaFM
CVPR 19 Oral
1st Restoration Modulation Model
DNI
CVPR 18
CSRNet
ECCV 20
CResMD
ECCV 20, PAMI 21
IKC
CVPR 19
Interactive
Modulation
Blind SR Survey
CUGAN
CVPRW 21
ECycleGAN
TMM 22
Real-ESRGAN
ICCVW 21
Video SR
EDVR
CVPRW 19
NTIRE 19 Video SR Champion
EDVI
ECCVW 20
AIM 20 Video-Temporal SR champion
VFHQ
CVPRW 22
BasicVSR
CVPR 21
NTIRE 21 Video SR Champion
Video Processing
PAN
ECCVW 20
Open Source
BasicSR
The Most Widely Used SR Open Source
PIPAL
ECCV 20
NTIRE 21, 22 IQA Challenge
GCFSR
CVPR 22
RepSR
ClassSR
CVPR 21
HAT
SR Transformer SOTA
BSRN
CVPRW 22
NTIRE 22 Light-Weight SR Champion
BALANCE
平衡
FOCUS
专注
LOVE
奉献
v22.06.06 by Jin Li
Director: Chao Dong

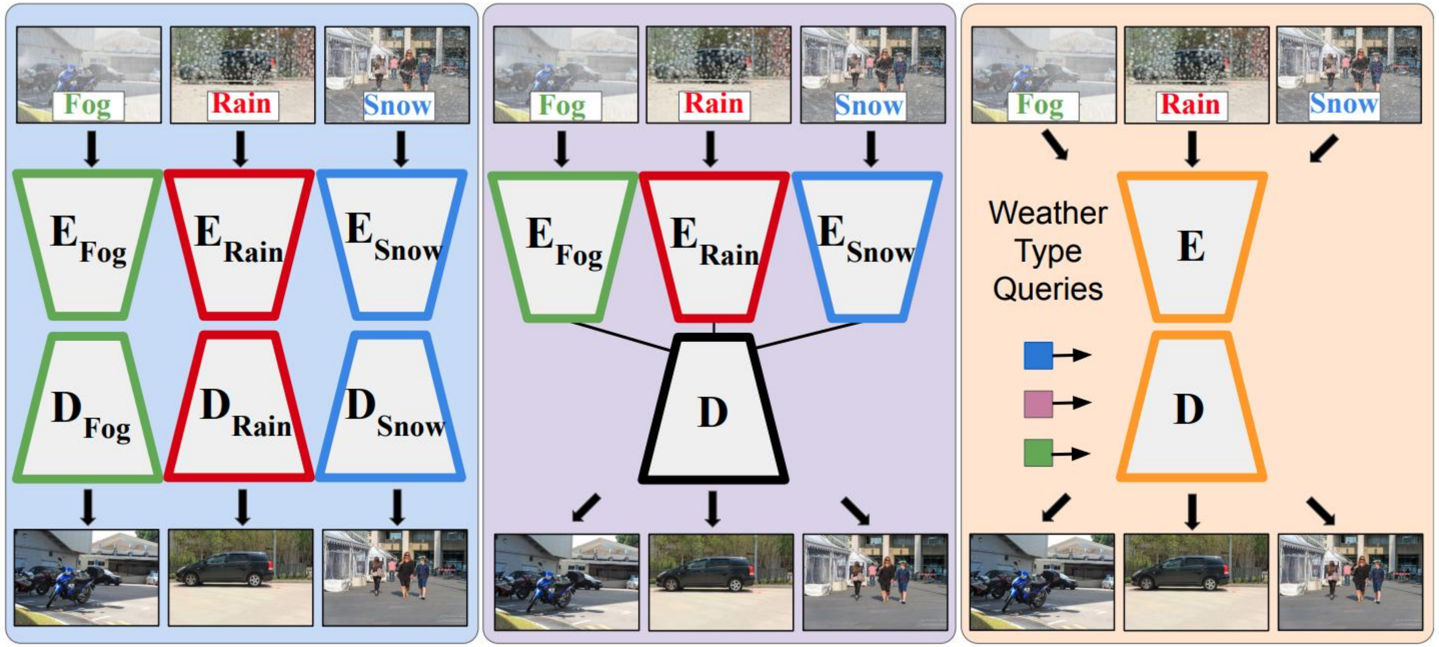
Background

- Pre-Trained Image Processing Transformer
- CVPR 2021



Background

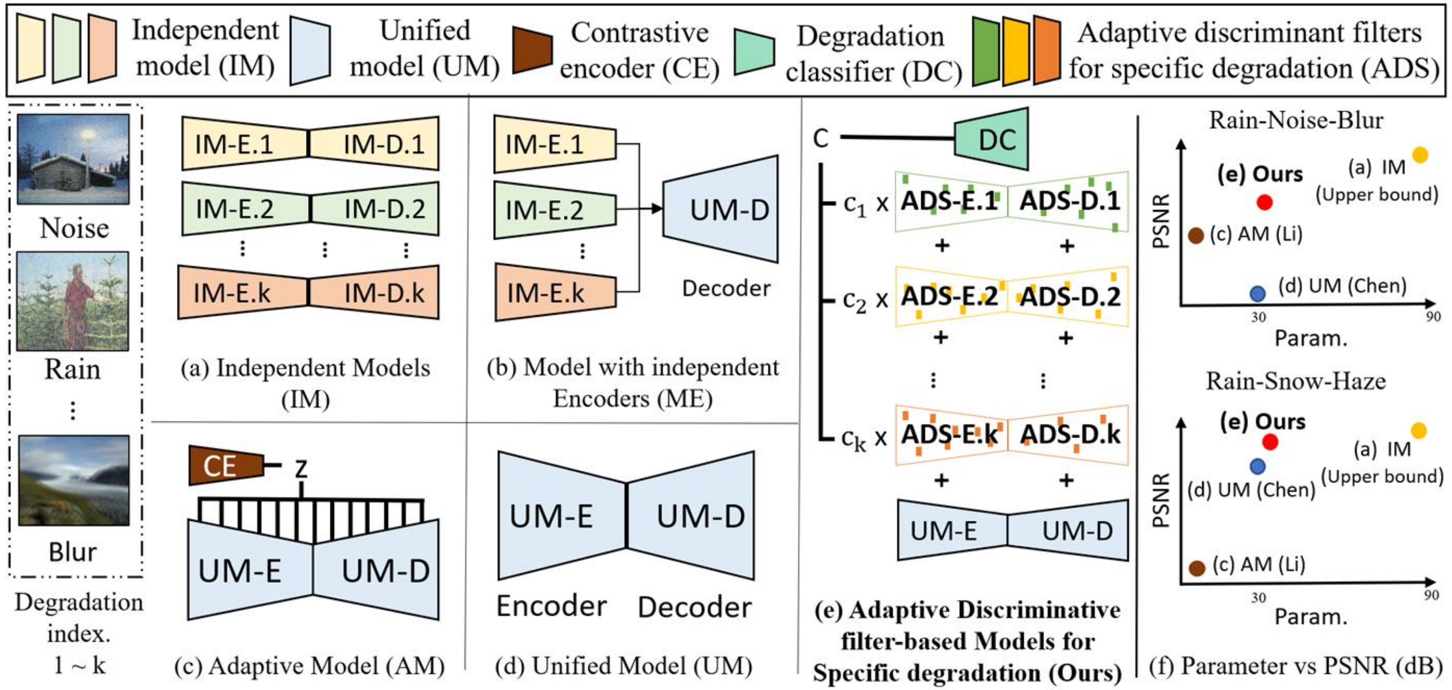
- Transweather: Transformer-based restoration of images degraded by adverse weather conditions
- CVPR 2022



(a) Task Specific Networks (b) All-in-One Network (c) TransWeather (Ours)

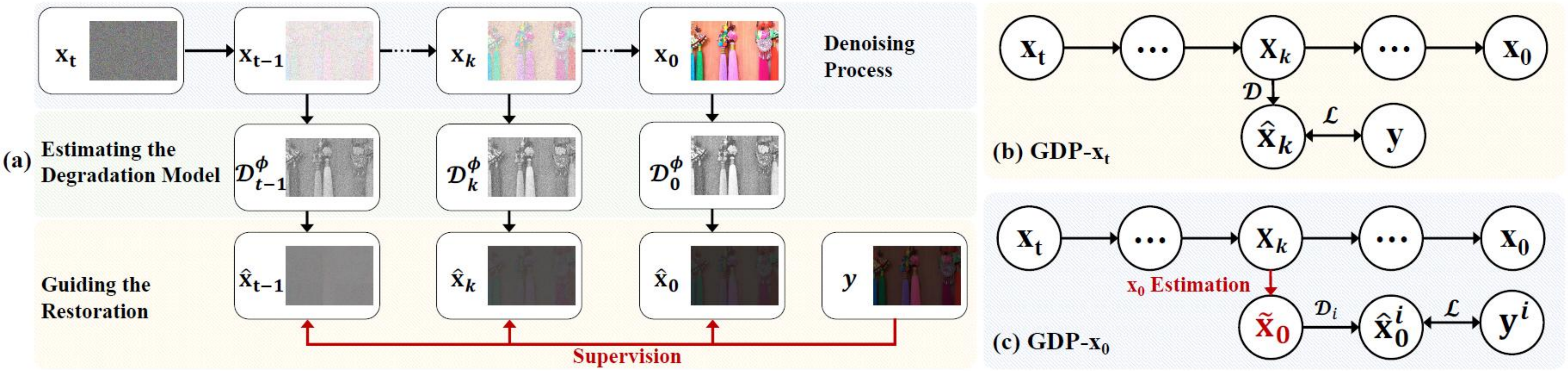
Background

- All-in-one Image Restoration for Unknown Degradations Using Adaptive Discriminative Filters for Specific Degradations
- CVPR 2023



Unified Image Restoration and Enhancement

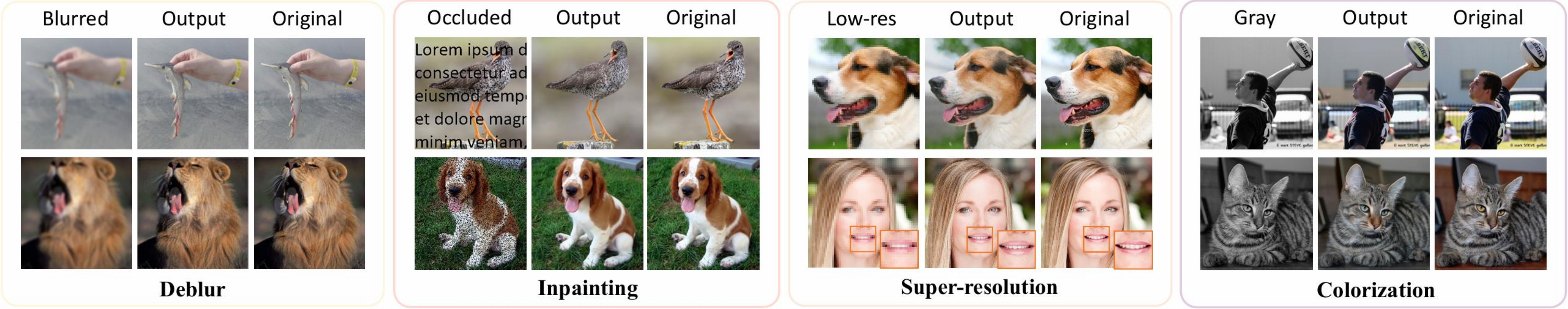
- Generative Diffusion Prior (GDP)
  - A unified framework for multiple restoration and enhancement tasks.
  - Use a pretrained unconditional image synthesis diffusion model as prior.



[7] Ben Fei, et al. Generative Diffusion Prior for Unified Image Restoration and Enhancement, CVPR, 2023

Unified Image Restoration and Enhancement

- Generative Diffusion Prior (GDP)
  - A unified framework for multiple restoration and enhancement tasks.
  - Use a pretrained unconditional image synthesis diffusion model as prior.
  - Different degradation models learned during the sampling process.



[7] Ben Fei, et al. Generative Diffusion Prior for Unified Image Restoration and Enhancement, CVPR, 2023

Unified Image Restoration and Enhancement

- Generative Diffusion Prior (GDP)

[7] Ben Fei, et al. Generative Diffusion Prior for Unified Image Restoration and Enhancement, CVPR, 2023

LM4LV: AFrozen Large Language Model for Low-level Vision Tasks

arXiv 2024

Boyang Zheng *
Shanghai Jiao Tong University
bytetriper@sjtu.edu.cn

Jinjin Gu
Shanghai AI Laboratory
jinjin.gu@sydney.edu.au

Shijun Li
Nanjing University
shijun_lee@outlook.com

Chao Dong
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
Shanghai AI Laboratory
chao.dong@siat.ac.cn

Presented by Zejia Fan
2023.6.2

Strength

- Take advantage of a frozen LLM for low-level vision
- No multi-modality data needed

Strength

- Multi-modal LLM (MLLM)
  - those that require an additional text-to-image module
  - those that do not
    - Structure like VQGAN, every modal into tokens
    - Training on massive multi-modal data
    - Unified as next-token prediction
    - Failing to provide a clear understanding of the capability of a LLM in processing visual features
  - Only discuss the former

Inspiration

- Current MLLMs are BLIND to Low-level Features
- Vision module in MLLMs often tend to capture high-level semantics but fail to maintain low-level details

Figure 1: Reconstruction results of the vision modules in different MLLMs. Emu2 provides highly semantic consistent images but fails to maintain low-level details, while MAE can reconstruct images with precise low-level details.

Framework

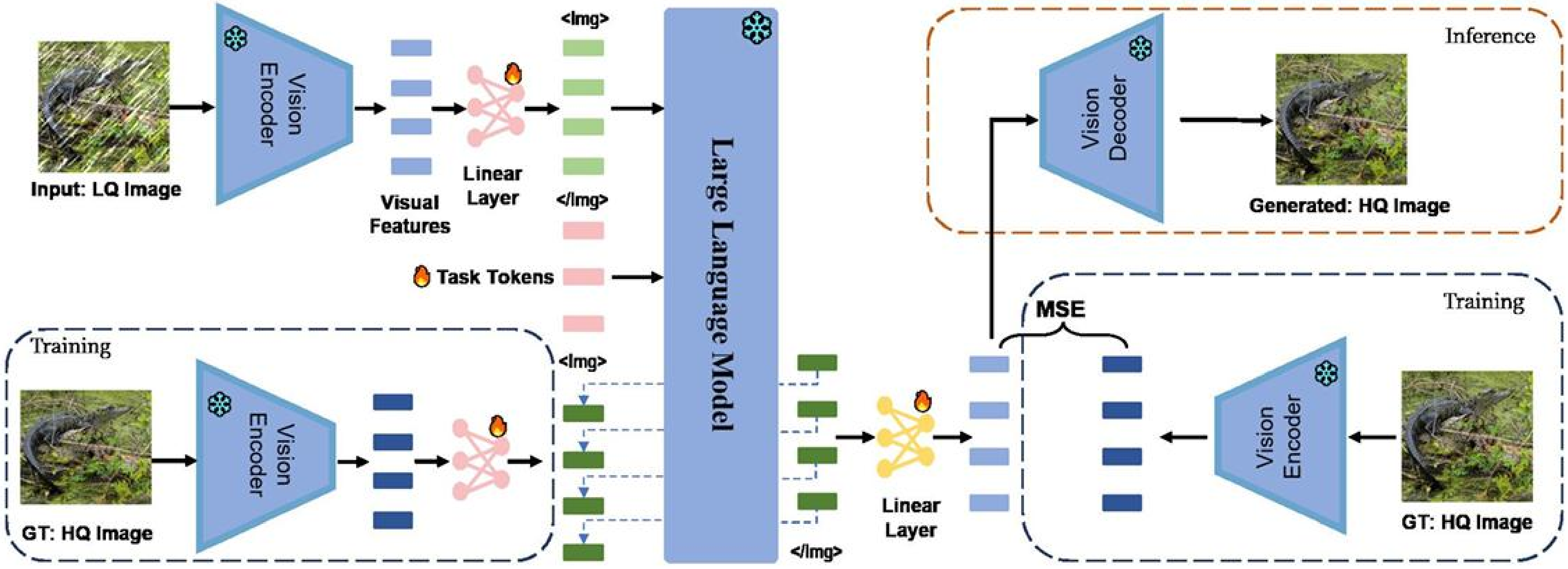


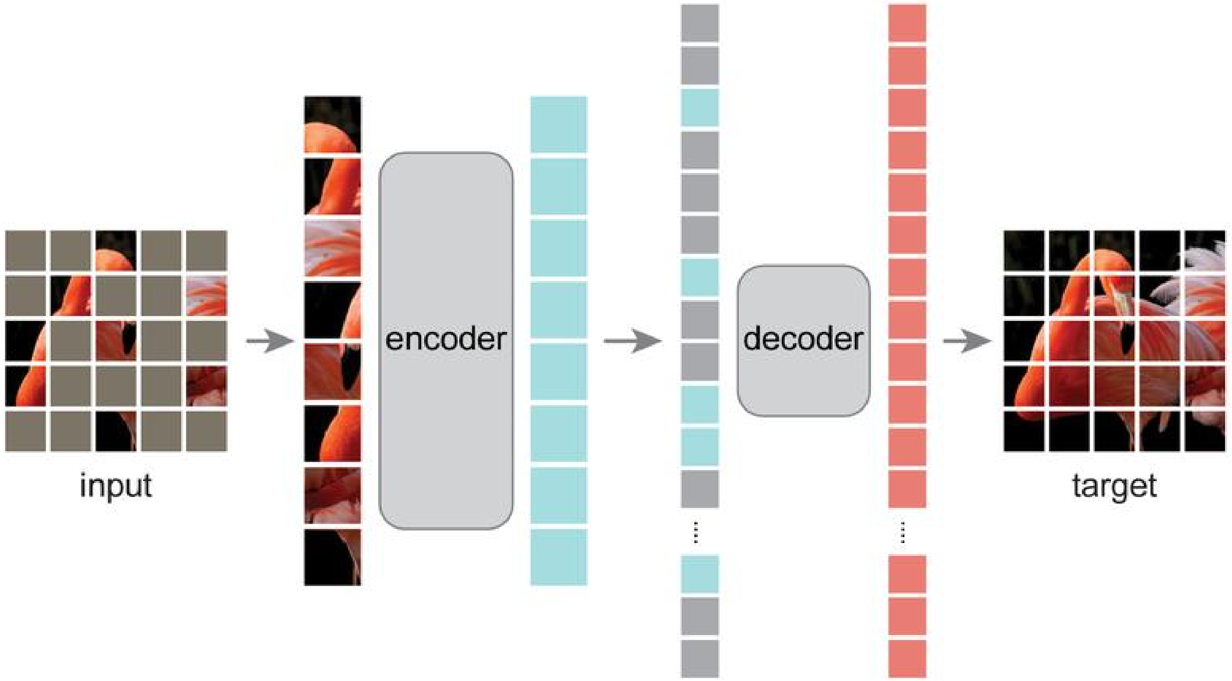
Figure 2: Network structure of our design. In the training phase, the visual tokens and the task tokens learns to prompt the LLM to generate next visual/text tokens. In the inference phase, the LLM generates visual tokens and text tokens in an auto-regressive manner. The visual tokens are then decoded into images.

Vision module choice

1. The training objective of the vision module should be reconstruction
   - the encoded feature can be decoded back to pixel space
2. Trained in an unsupervised manner to avoid any multi-modal training
   - If the encoder transformed image into text-like features, it becomes unclear whether the LLMis leveraging its powerful text processing abilities or it inherently has the capability to process other modalities (visual).

Vision module choice

Masked Autoencoder (MAE)



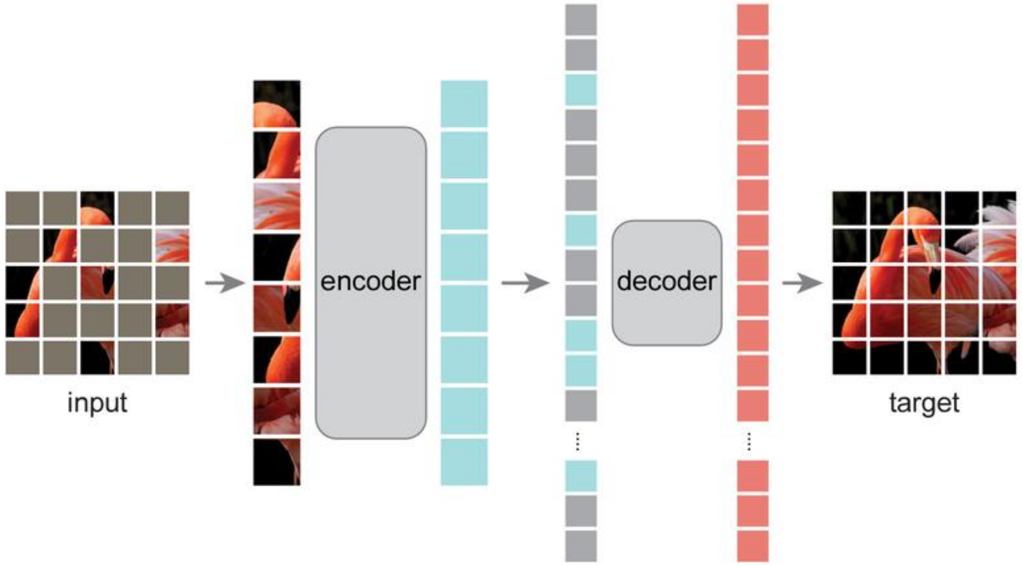
Vision module choice

Masked Autoencoder (MAE)

- Encoder frozen, finetune decoder
- Originally calculate the reconstruction loss solely on masked tokens

Table 1: Reconstruction FID (rFID), precision, recall and PSNR on the validation set of ImageNet. MAE-L1 indicates to use L1 loss for fine-tuning MAE's decoder. MAE* is the version tuned by a combination of L1 loss and LPIPS Loss. Best results are bolded.

Model	rFID↓	prec(%)↑	recall(%)↑	PSNR↑
MAE	84.22	13.35	45.78	19.15
MAE-L1	9.96	88.46	97.57	**29.21**
VQGAN	1.49	94.90	99.67	22.61
MAE*	**1.24**	**99.94**	**99.97**	28.96



Framework

- An auto-regressive manner
- Trainable task token
- Two linear adaptation modules

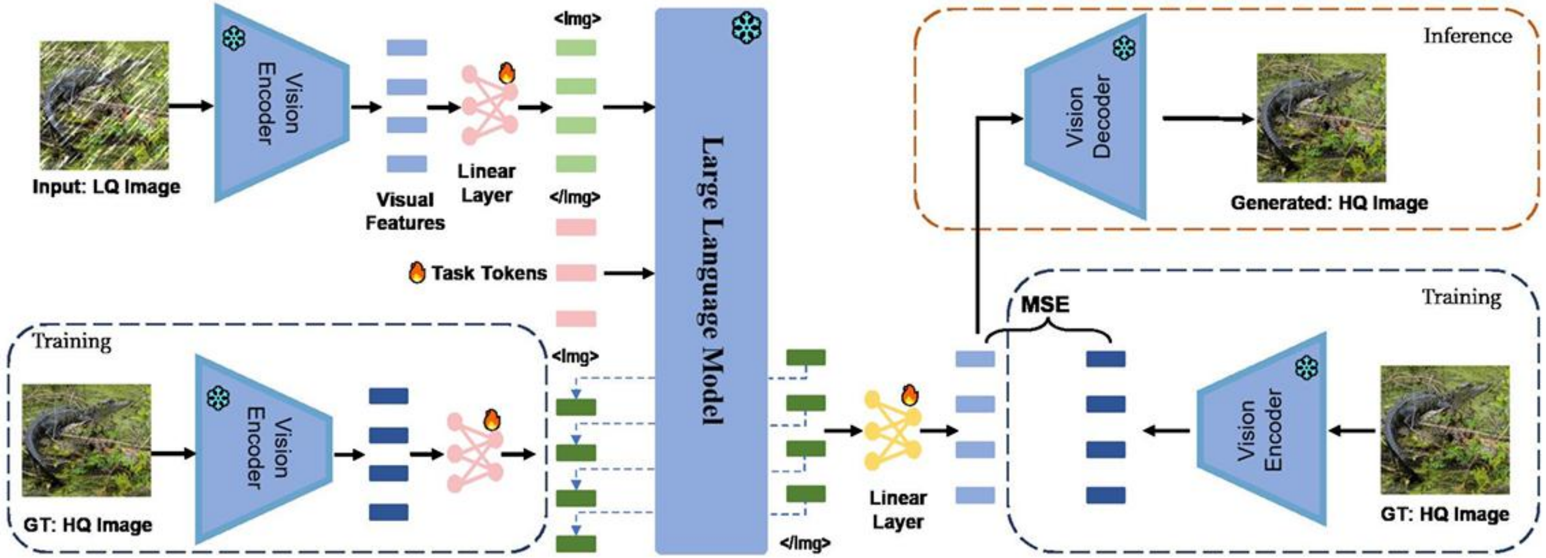


Figure 2: Network structure of our design. In the training phase, the visual tokens and the task tokens learns to prompt the LLM to generate next visual/text tokens. In the inference phase, the LLM generates visual tokens and text tokens in an auto-regressive manner. The visual tokens are then decoded into images.

Framework

- An auto-regressive manner
- Trainable task token

```
Human:  <Img><LQ-image></Img> <task> Assistant:  <Img><HQ-image></Img>
```

Experiment

- LLaMA2-7B instruct as base LLM for all experiments
- MAE-large for vision module
- LLAVA595K for degradation generation
- Main tasks: denoising, deblurring , pepper noise removal, deraining, mask removal
- MAE-r as removing LLM

Experiment

Table 2: Results of LM4LV on various low-level vision tasks. The top five tasks are image restoration tasks, the bottom two tasks do not require restoration, but involve large-scale spatial operations.

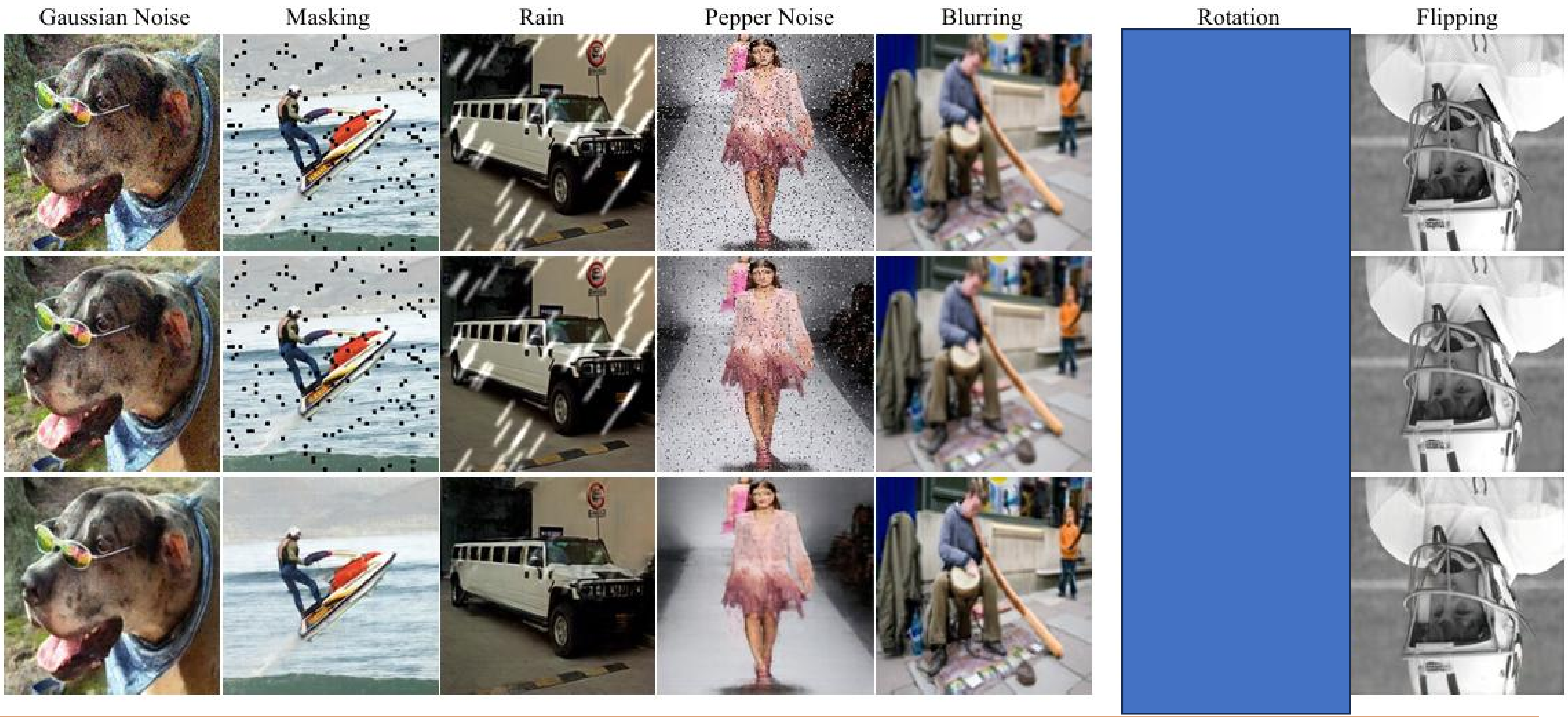
Tasks	Degraded		MAE-r		LM4LV		
	PSNR ↑	SSIM ↑	PSNR↑	SSIM ↑	PSNR ↑	SSIM ↑	$\Delta_{\text{PSNR/SSIM}}$
Denoising	23.11dB	0.49	19.96dB	0.65	26.77dB	0.80	+6.81dB/+0.15
Deblurring	30.88dB	0.83	26.14dB	0.78	26.23dB	0.79	+0.09dB/+0.01
Deraining	20.52dB	0.84	19.96dB	0.74	24.62dB	0.77	+4.66dB/+0.03
Pepper Removal	19.22dB	0.51	23.01dB	0.58	25.20dB	0.75	+2.19dB/+0.17
Mask Removal	20.54dB	0.83	20.00dB	0.73	25.83dB	0.80	+5.83dB/+0.07
Rotation	inf[7]	1.00	29.52dB	0.89	27.18dB	0.83	-2.34dB/-0.06
Flipping	inf	1.00	29.52dB	0.89	27.28dB	0.84	-2.24dB/-0.05

Experiment

degraded

MAE-r

LM4LV



Experiment

- Auto Regression matters.
- ViT-LLM generation: directly output curated image tokens in a single forward process

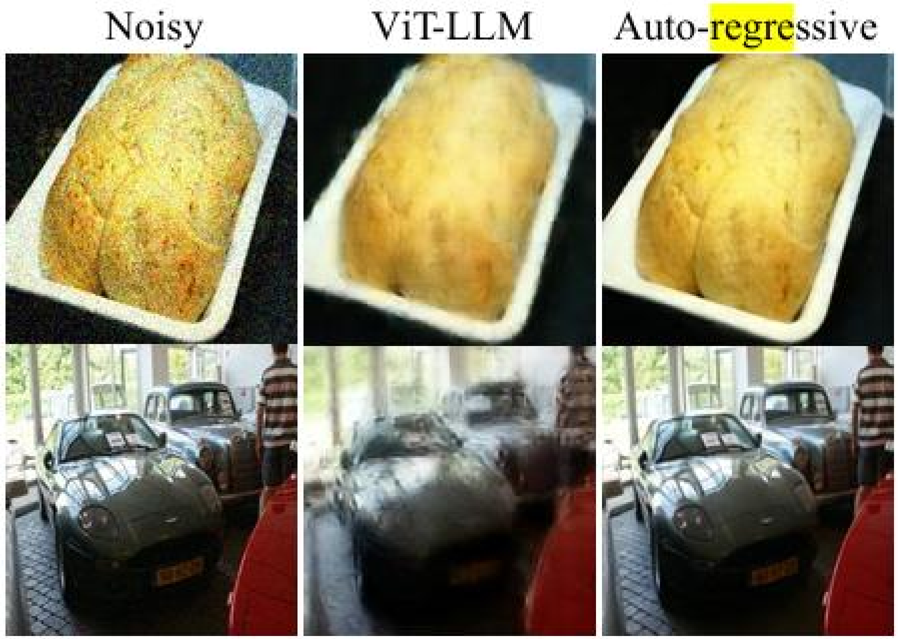


Figure 5: ViT-LLM generation fails for image denoising even when the noise level is low (2nd row), producing low-quality and blurred images.

Experiment

- Is the Linear Layer Doing the Task?
- Leaving only the linear adaptation module.

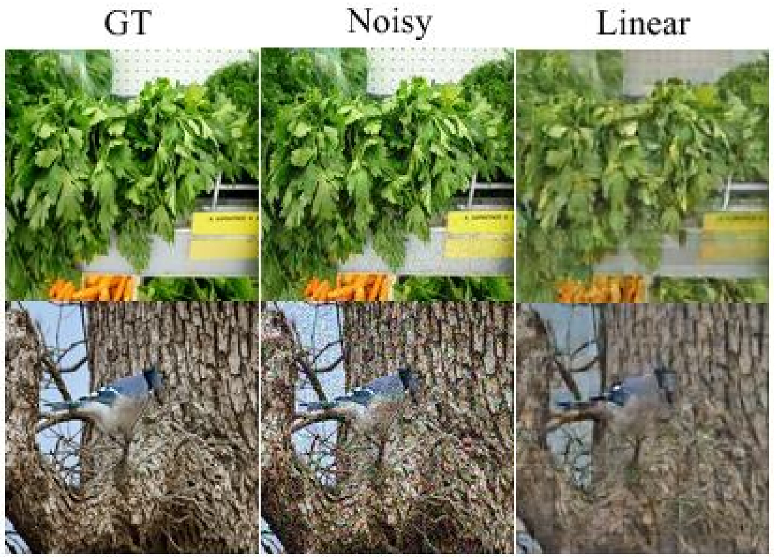


Figure 6: Using a single linear layer for de-noising yields bad results.

Experiment

- Is the Linear Layer Doing the Task?
- Leaving only the linear adaptation module.
- Two linear layers tend to perform a scaled identity mapping even though they are not forced to do so.

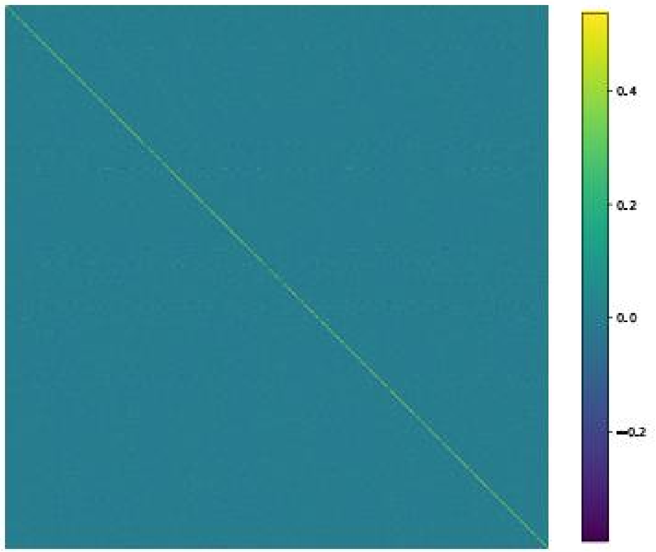


Figure 11: The multiplication matrix tends to center it's weight on the diagonal. Yellow represents a large value, and blue represents a small value.

Experiment

- Does Text Pre-training Play an Important Role?

Figure 7: Using randomly initialized LLM gives messy outputs.

Experiment

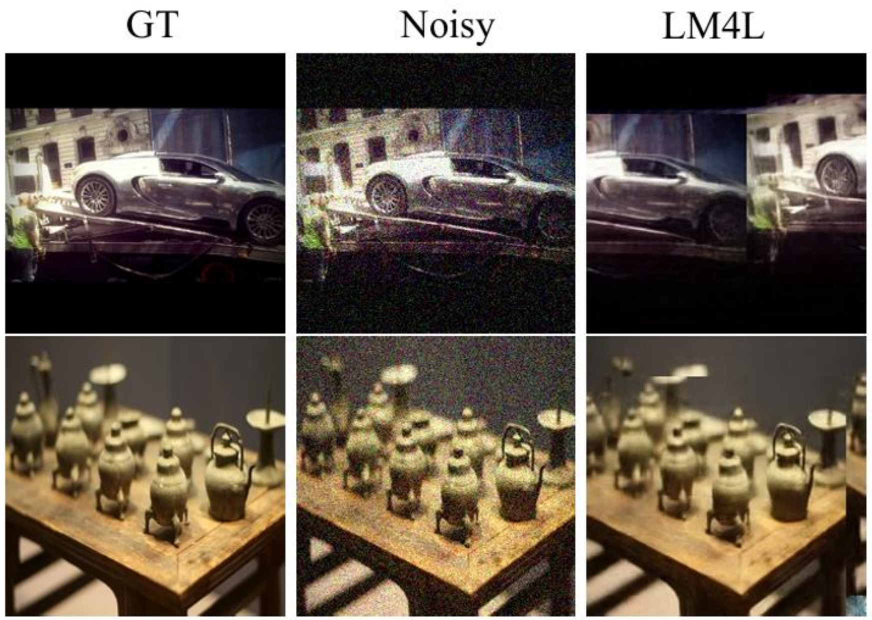
- LLM vs Expert Models

Table 3: Comparisons of different expert models and our methods. Using LLM gain superior performance in image rotation, and surpass MLP in image denoising. Best results are in bold.

	Denoising		Rotation	
	PSNR↑	SSIM ↑	PSNR↑	SSIM ↑
MLP	25.87dB	0.76	13.29dB	0.32
Transformer	**27.42dB**	**0.81**	10.52dB	0.23
Ours*	26.77dB	0.80	**27.18dB**	**0.83**

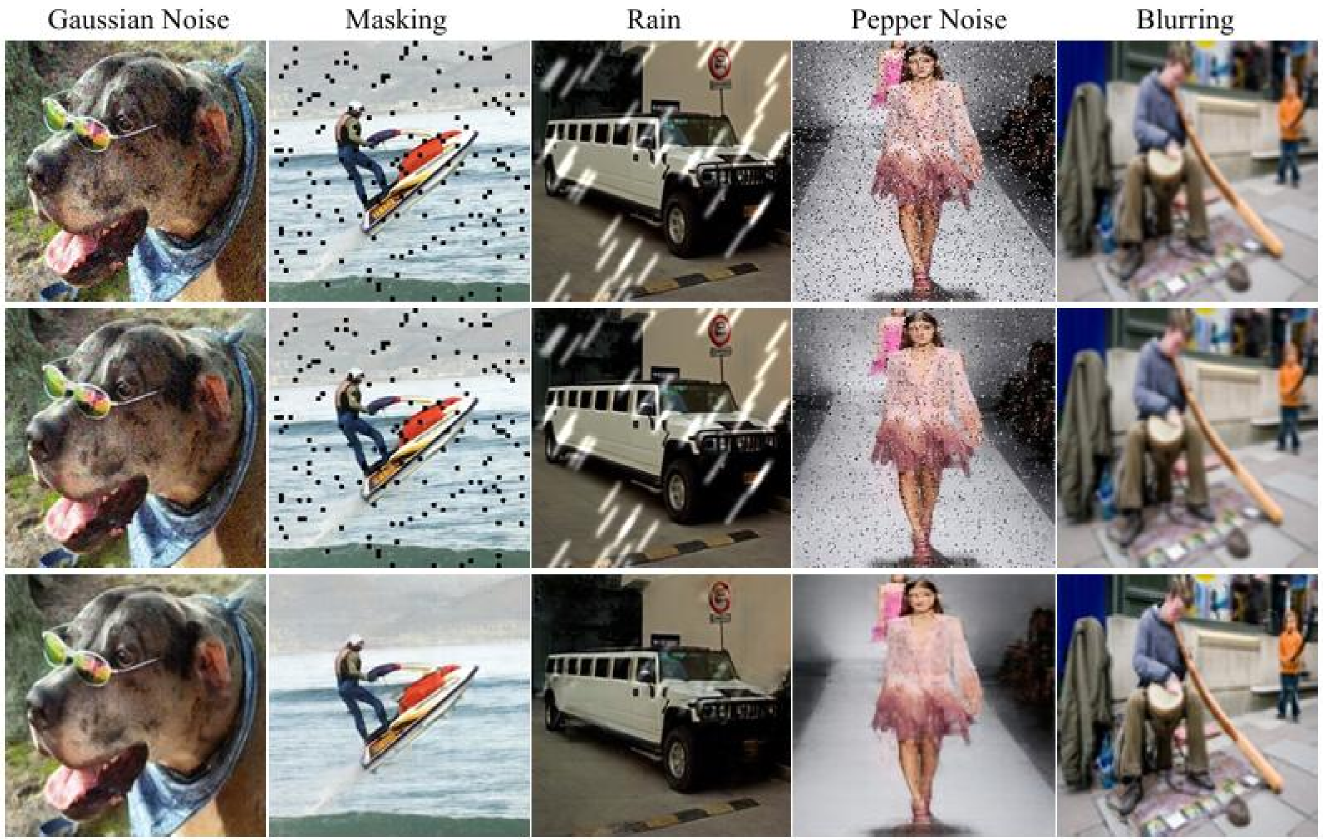
Experiment

- Failure case
- fails to align the visual tokens correctly



Limitation

- Lack high-frequency details
- Could be improved by adding skip-connection or multi-modal data



Conclusion

- Does a frozen LLM has the ability to accept, process, and output low-level features?
- By designing a framework from bottom to top, give a positive answer, showing LLMs’ non-trivial performance on various low-level tasks.

Thanks for your listening!