LM4LV: A Frozen Large Language Model for Low-level Vision Tasks

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

arXiv 2024

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Introduction

- Current major direction for many MLLM related works is towards a better semantic fusion of the text and image modality.

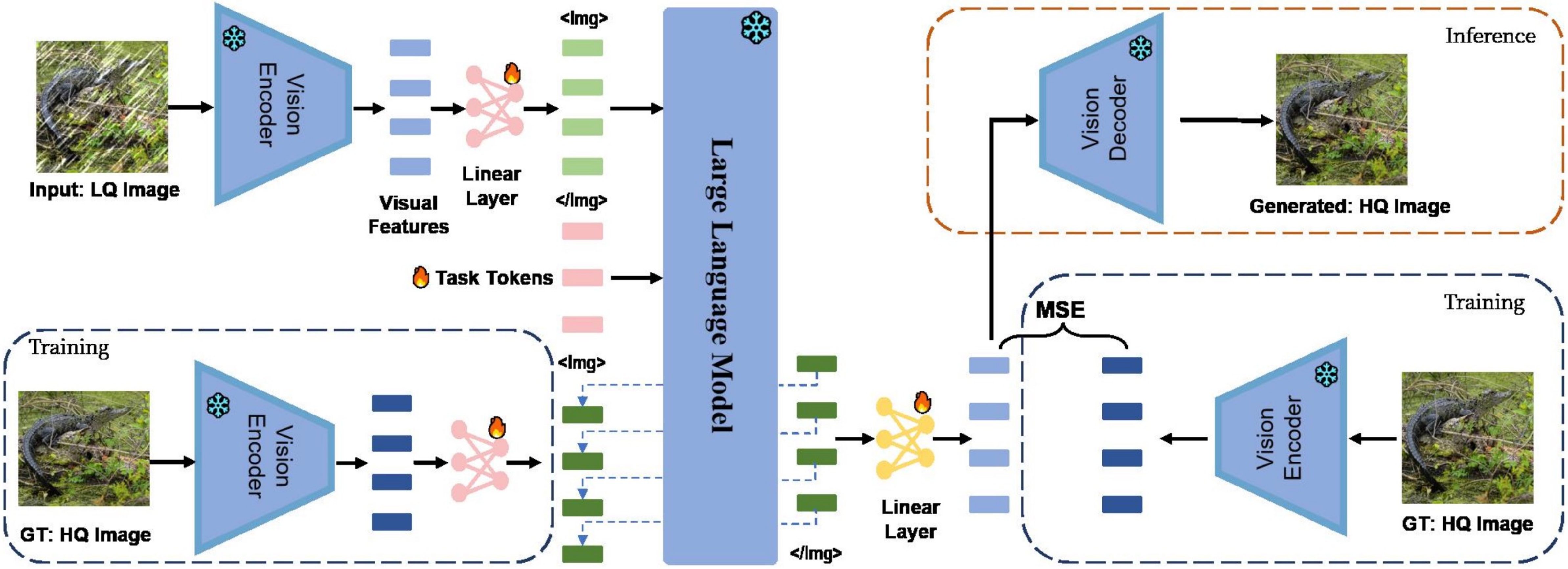
- First attempt to investigate a frozen LLM's capability in processing low-level feature, highlighting the importance of investigating LLMs' capability to process visual features with no multi-modal data or prior, leading to a deeper understanding of LLMs' inner mechanisms.

- By simply training two linear layers with vision-only data, a frozen LLM showcases non-trivial capability on a wide range of low-level vision tasks.

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Framework



Input: LQ Image
Vision Encoder
Visual Features
Linear Layer
<Img>
</Img>
Task Tokens
Large Language Model
Training
GT: HQ Image
Vision Encoder
<Img>
</Img>
Linear Layer
MSE
Inference
Vision Decoder
Generated: HQ Image
Training
Vision Encoder
GT: HQ Image

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Vision Modules

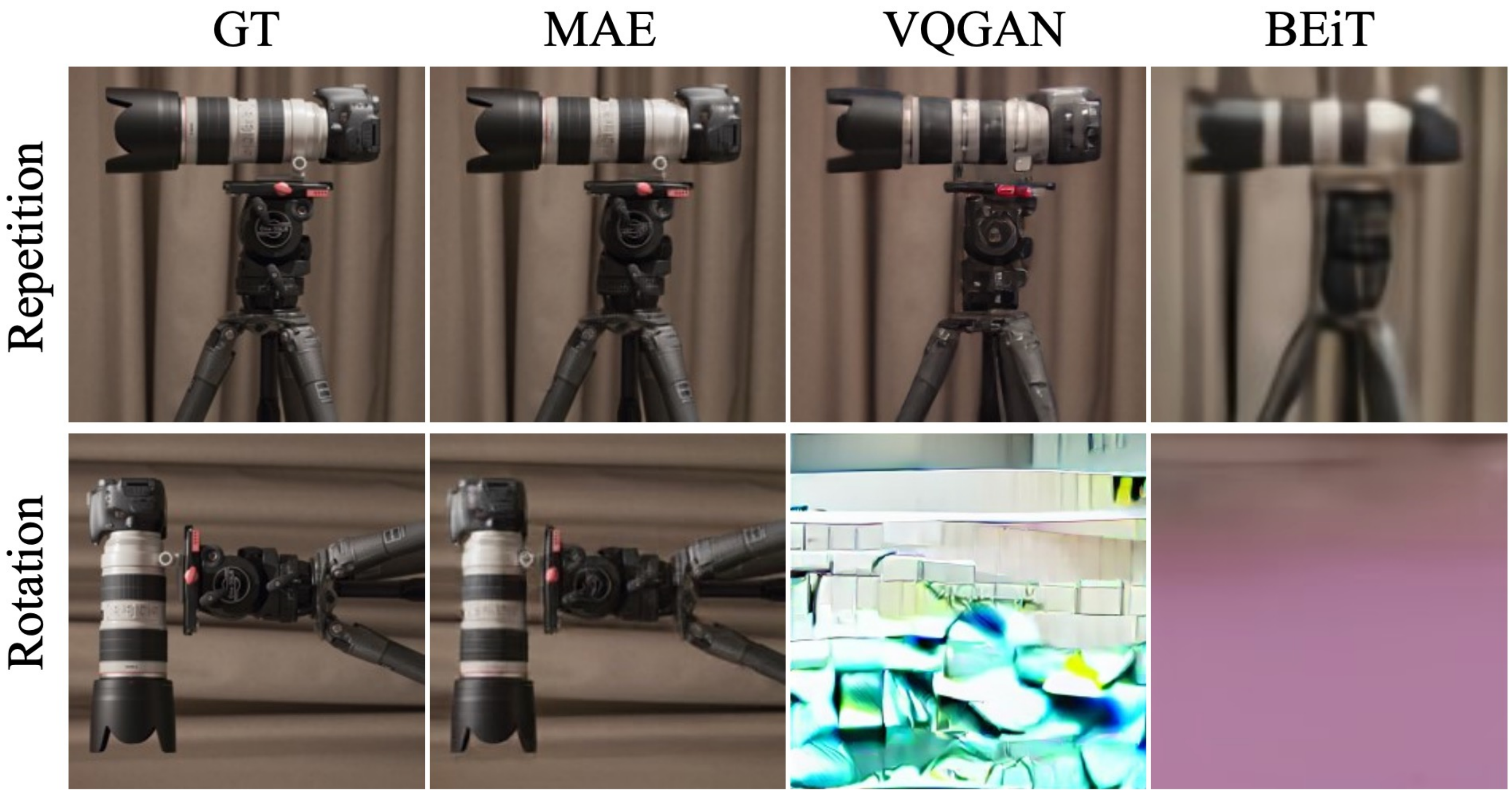
- Training objective of the vision module should be reconstruction
- Vision modules must be trained in an unsupervised manner to avoid any multi-modal training
  - if the encoder has already transformed the image into text-like features, it becomes unclear
  - whether the LLM is leveraging its powerful text processing abilities to handle text features
  - or it inherently has the capability to process other modalities.

Fine-tuning MAE for Image Reconstruction

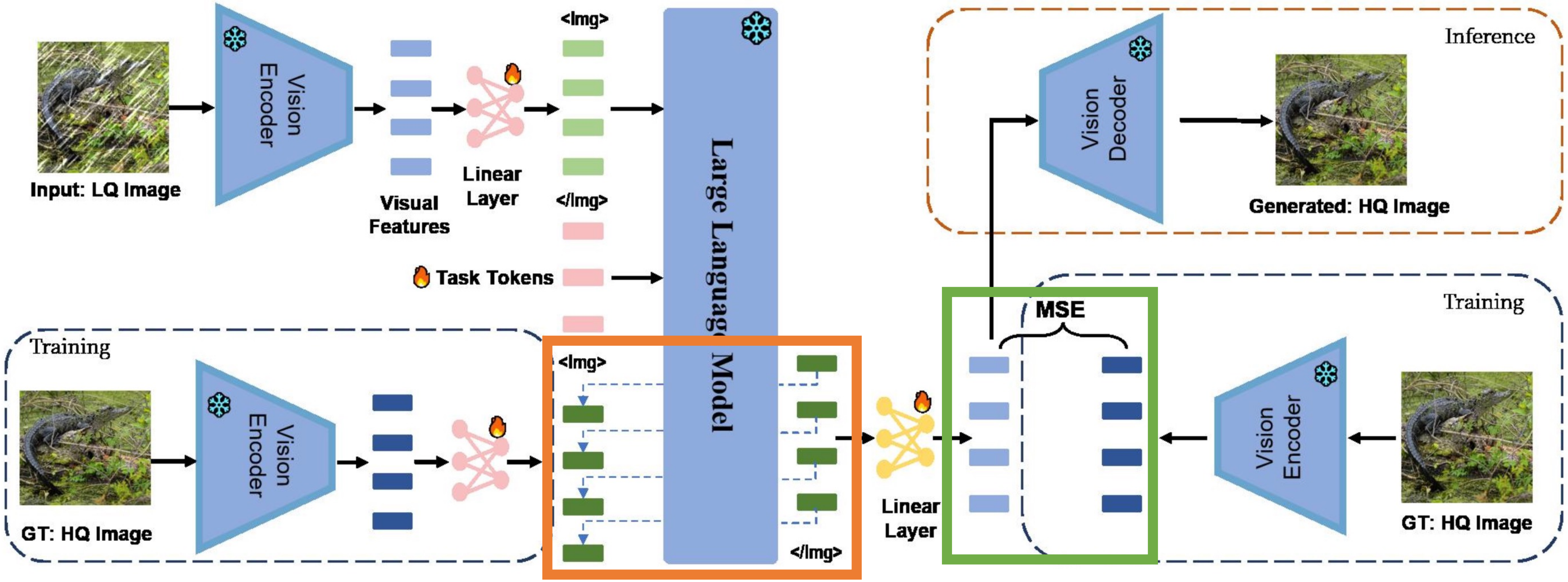
Model	rFID↓	prec(%)↑	recall(%)↑	PSNR↑
MAE	84.22	13.35	45.78	19.15
MAE-L1	9.96	88.46	97.57	**29.21**
VQGAN	1.49	94.90	99.67	22.61
MAE*	**1.24**	**99.94**	**99.97**	28.96

- 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.
- Released version of MAE calculates the reconstruction loss solely on masked tokens, leads to inconsistency between training and inference.

Vision Modules



Next Element Prediction



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

- Two simple linear layers as the adapter modules between the LLM and the vision encoder/decoder to align the feature dimension
- **Next-token cross-entropy loss** for text tokens, and for continuous visual tokens, we apply a **next token l2-regression loss**

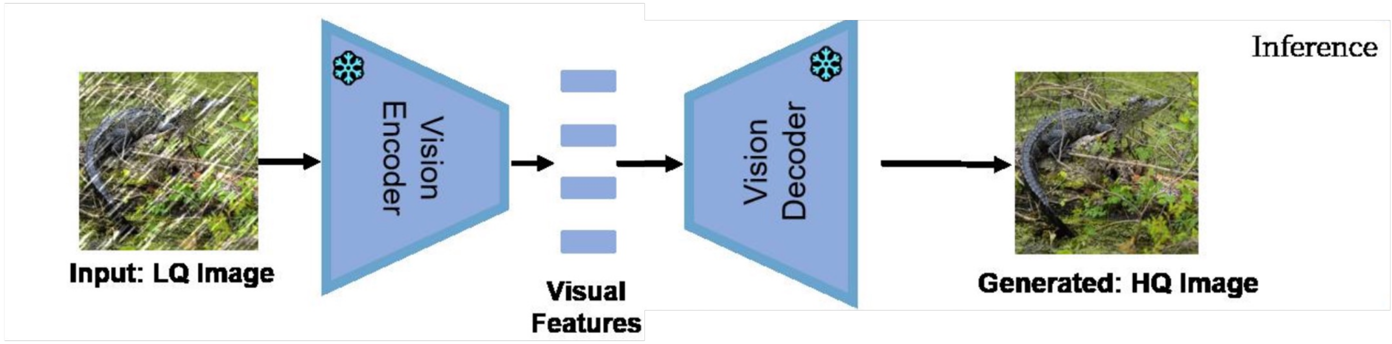
Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Experiment setup

- Evaluation tasks
  - Restoration tasks
    - Denoising
    - Deblurring
    - Pepper noise removal
    - Deraining
    - Mask removal
  - Spatial operation tasks
    - image rotation
    - image flipping

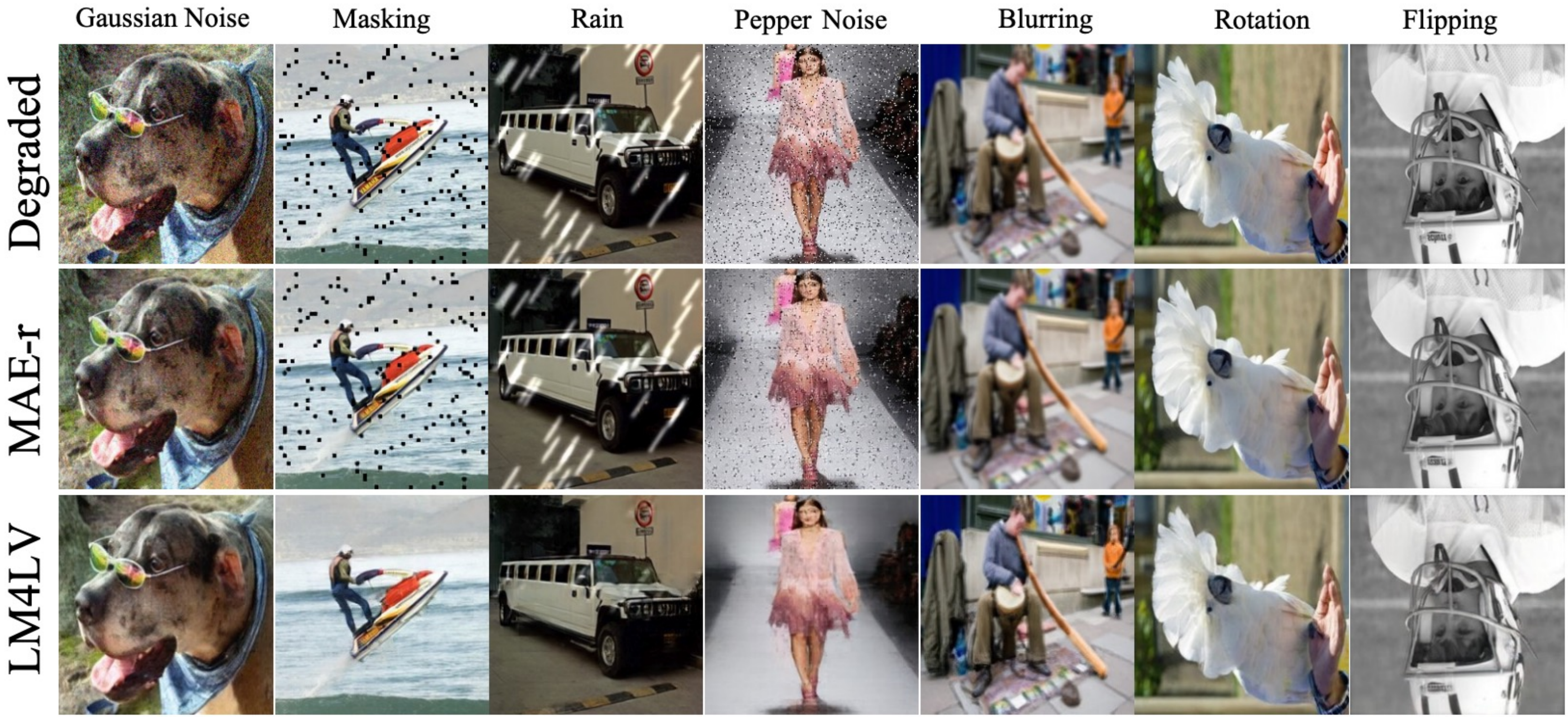
- Baseline
  - MAE to reconstruct degraded images without further modification (denote as MAE-r)



Result

Tasks	Degraded		MAE-r		LM4LV		
	PSNR ↑	SSIM ↑	PSNR↑	SSIM ↑	PSNR ↑	SSIM ↑	$\Delta_{PSNR/SSIM}$
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

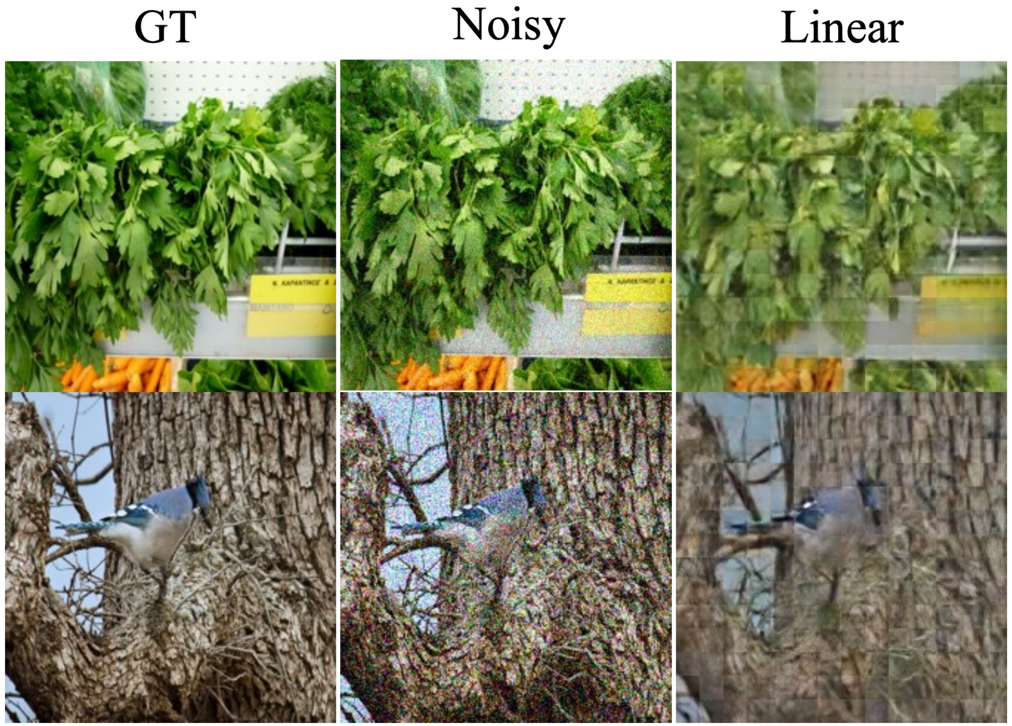
Result


Gaussian Noise
Masking
Rain
Pepper Noise
Blurring
Rotation
Flipping
Degraded
MAE-r
LM4LV

Ablation study

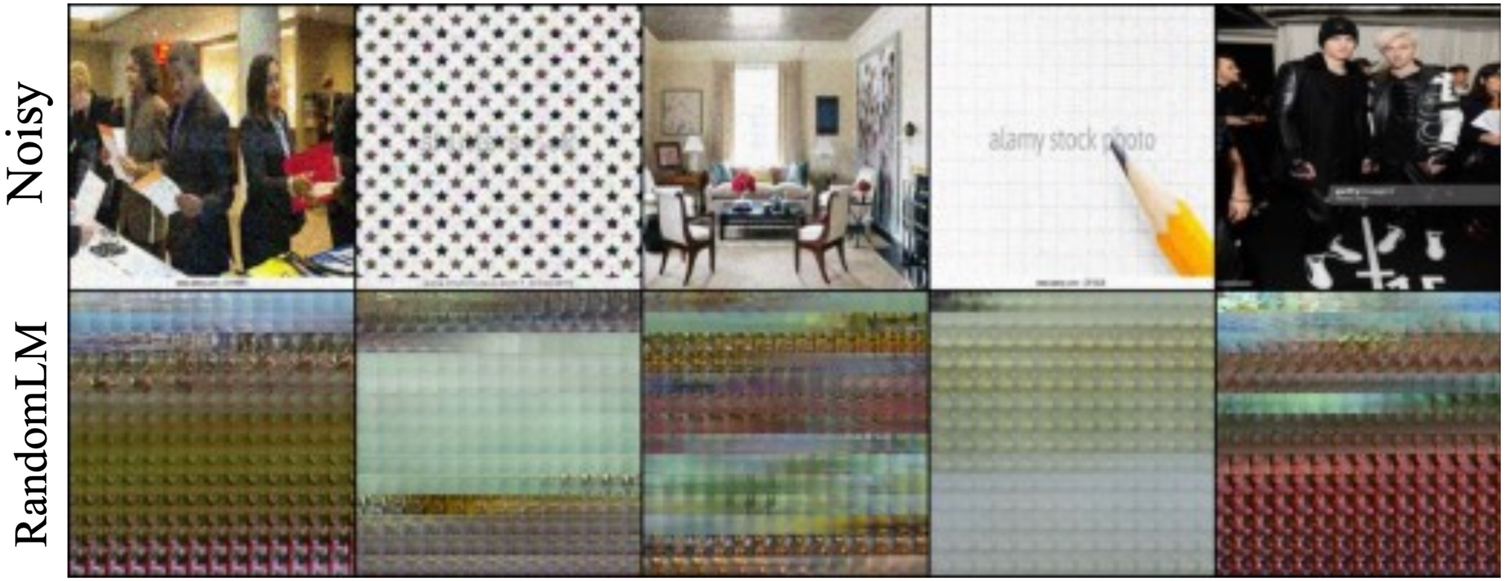
- Auto-regressive Generation Matters
  - ViT-LLM generation produces low-quality and blurred images
  - auto-regressive feature generation naturally aligns with LLM's behavior

- Linear Layer
  - single linear layer is insufficient to handle low-level vision
  - two linear layers tend to perform a scaled identity mapping



Ablation study

- Text Pre-training



- LLM vs Expert Models

	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**

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- The goal for this work is not to achieve the best performance in image restoration, but to demonstrate the potential of LLMs in processing low-level features.

- LM4LV could not restore high-frequency details in degraded images. This is natural because the LLM does not have image prior, which could be improved by adding skip-connection or multi-modal data.

- LLMs' non-trivial performance on various low-level tasks, hope inspiring new perspectives on the capabilities of LLMs and deeper understanding of their mechanisms.