2025 summer seminar

# VLM/MLLM-based Agentic Systems: Methods and Applications


***Pusan National University***
*Computer Vision & Signal Processing Lab, Dept. of Electronic Engineering*

***Presented By***
*정다운*


CVSP Lab
Computer Vision & Signal Processing

# Index

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).
2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.


- Introduction
  - 논문 선정 이유
  - background
- 논문 소개
  - AN INTELLIGENT AGENTIC SYSTEM FOR COMPLEXIMAGE RESTORATION PROBLEMS [ICLR 2025] [1)]
  - RestoreAgent: Autonomous Image Restoration Agent via Multimodal Large Language Models [NeurIPS 2024] [2)]




CVSP Lab
Computer Vision & Signal Processing

# Introduction

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).
2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- 논문 선정 이유
  - 요즘 핫한 "Agent"
    - LLM/VLM의 추론 능력이 향상됨에 따라 자연스럽게 성능 향상
    - 최적화가 어려운 부분에서 스스로 문제를 인지하고 계획하고 수정하는 시스템
  - Agentic System이 어떻게 설계되고 돌아가는 지 설명하기 위함
    - 논문의 Task가 모두 "Image Restoration"이지만 LLM/VLM을 활용한 Agentic system이 어떤 식으로 돌아가는지에 대해 집중해서 들어주시면 됩니다!



CVSP Lab
Computer Vision & Signal Processing

# Background

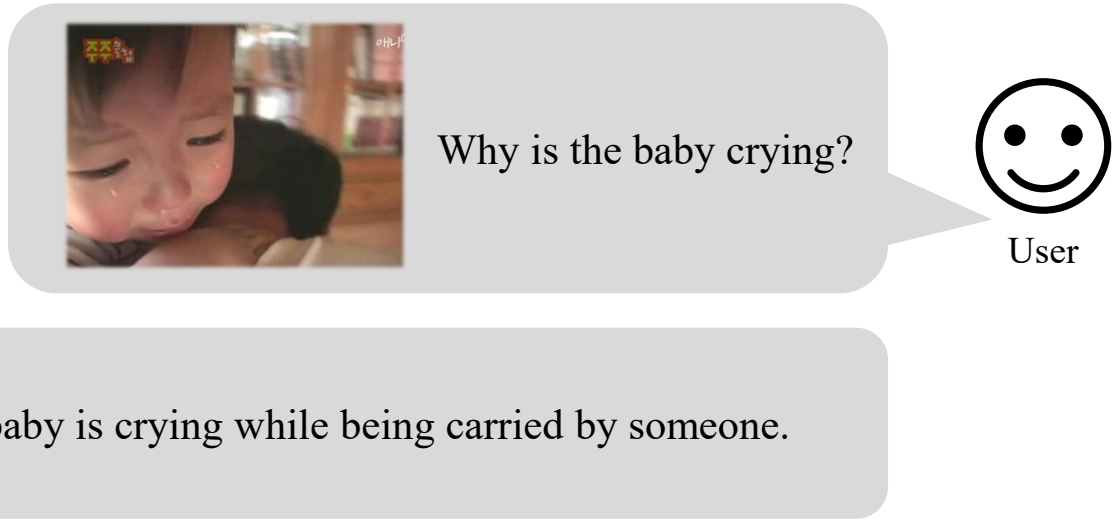
- Vision Language Model(VLM)
  - 이미지와 텍스트를 동시에 이해하고 연결하는 모델
    - 대표적으로 CLIP, ALIGN, BLIP, CoCa 등이 있음
  - Core Idea
    - 시각적 정보(image feature)와 언어적 정보(text feature)를 embedding space에서 정렬
    - 이를 통해 zero-shot image classification, retrieval, captioning 등 다양한 멀티모달 task 수행 가능
  - 구조
    - Vision Encoder: CNN, Vision Transformer(ViT) 등으로 image feature 추출
    - Text Encoder: Transformer 기반 언어모델로 텍스트 embedding 추출
    - Contrastive learning 등을 활용해 시각-언어 feature alignment


CVSP Lab
Computer Vision & Signal Processing

# Background

0) Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.

- Large Language and Vision Assistant (LLaVA) [0)]
  - LLM과 Vision Encoder를 결합한 Visual Instruction Tuning 모델
    - "이미지를 참고하여 질문에 대답하는 모델"을 만들기 위해 이미지, 질문과 대답을 학습시킴
    - ChatGPT 같은 LLM에 시각적 이해 능력을 붙인 대표적인 사례
  - Instruction-following dataset으로 학습하여 질문-응답(QA), 이미지 설명, reasoning 등 가능
    - 말그대로 '이미지', '이미지에 대한 질문'과 그에 대한 '대답'을 dataset으로 만들어서 학습시킴




<Instruction Tuning 이전의 모델 대답 – captioning 등 주어진 일 밖에 못함>

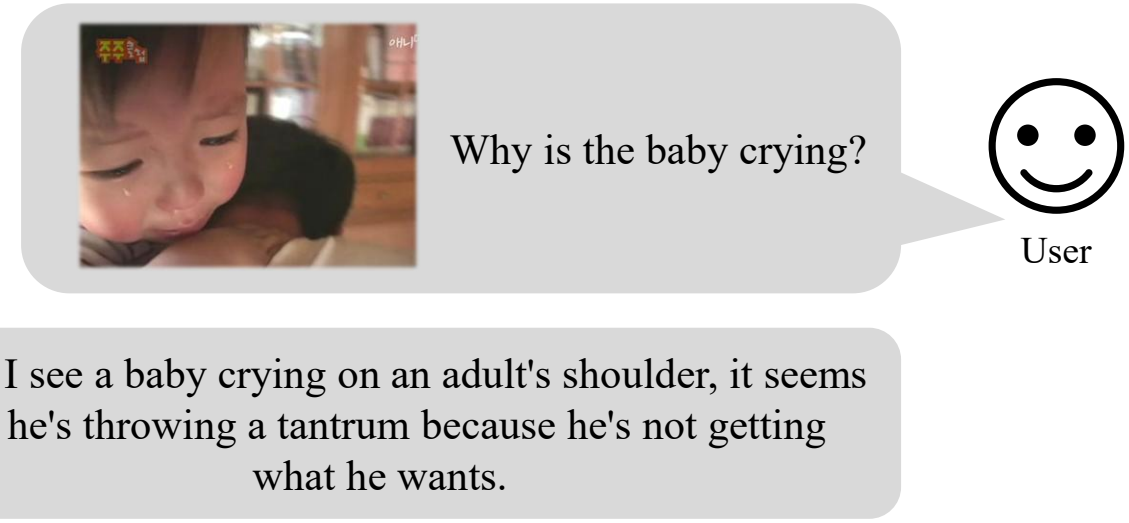


CVSP Lab
Computer Vision & Signal Processing

# Background

0) Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.

- Large Language and Vision Assistant (LLaVA) [0)]
  - LLM과 Vision Encoder를 결합한 Visual Instruction Tuning 모델
    - "이미지를 참고하여 질문에 대답하는 모델"을 만들기 위해 이미지, 질문과 대답을 학습시킴
    - ChatGPT 같은 LLM에 시각적 이해 능력을 붙인 대표적인 사례
  - Instruction-following dataset으로 학습하여 질문-응답(QA), 이미지 설명, reasoning 등 가능
    - 말그대로 '이미지', '이미지에 대한 질문'과 그에 대한 '대답'을 dataset으로 만들어서 학습시킴




<Instruction Tuning 이후의 모델 대답 – 이미지와 질문을 참고하여 알맞은 대답 생성>

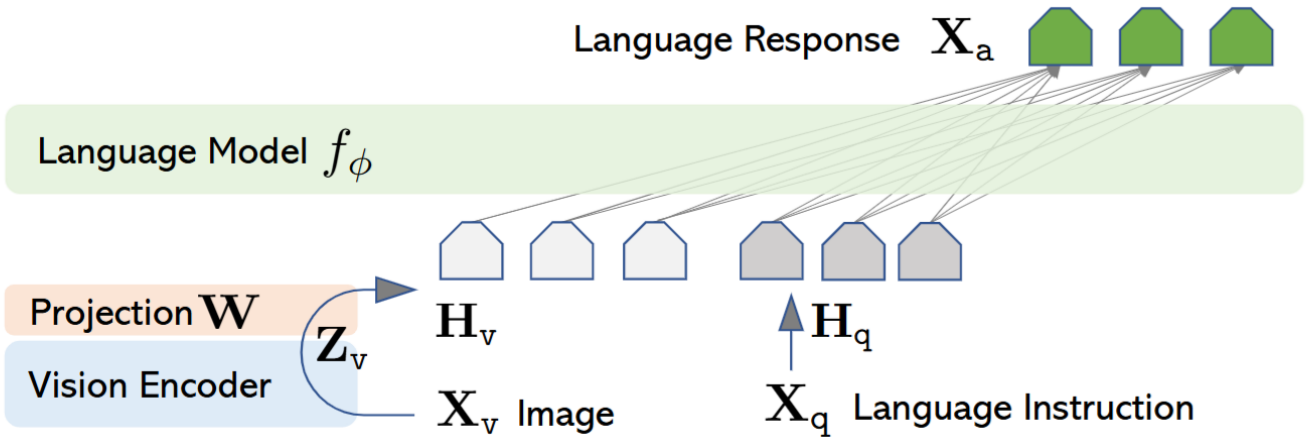


CVSP Lab
Computer Vision & Signal Processing

# Background

0) Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.

- Large Language and Vision Assistant (LLaVA) [0)]
  - 구조
    - Vision Encoder: CLIP ViT 가 image feature 추출
    - Projection Layer: image embedding을 LLM의 토큰 공간으로 매핑
    - LLM Backbone: Vicuna-13B 등의 모델
      - Instruction Tuning으로 이미지 기반 질의 응답가능
  - 단순한 captioning을 넘어 multi-modal reasoning이 가능해짐




<VLM Instruction Tuning model - 구조>



CVSP Lab
Computer Vision & Signal Processing

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

# AN INTELLIGENT AGENTIC SYSTEM FOR COMPLEX IMAGE RESTORATION PROBLEMS [1)]

ICLR 2025

# Introduction

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

- Introduction to paper
  - AN INTELLIGENT AGENTIC SYSTEM FOR COMPLEX IMAGE RESTORATION PROBLEMS [1)]
  - Image Restoration문제를 LLM/VLM모델을 활용한 Agentic System으로 해결함

Published as a conference paper at ICLR 2025

## AN INTELLIGENT AGENTIC SYSTEM FOR COMPLEX IMAGE RESTORATION PROBLEMS

**Kaiwen Zhu**[1,2]* **Jinjin Gu**[3]* **Zhiyuan You**[4,5] **Yu Qiao**[2] **Chao Dong**[5,2]†
[1]Shanghai Jiao Tong University [2]Shanghai Artificial Intelligence Laboratory
[3]The University of Sydney [4]The Chinese University of Hong Kong
[5]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
zhukaiwen@pjlab.org.cn, jinjin.gu@sydney.edu.au, zhiyuanyou@foxmail.com,
qiaoyu@pjlab.org.cn, chao.dong@siat.ac.cn

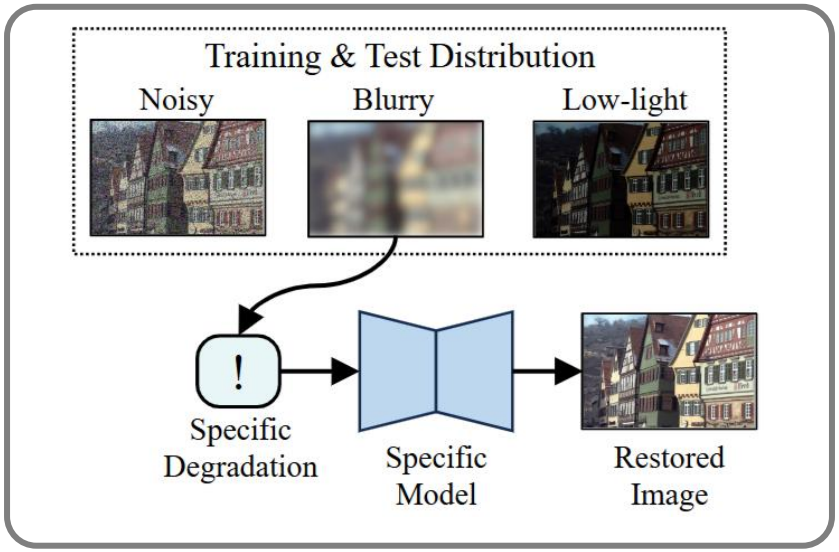


CVSP Lab Computer Vision & Signal Processing

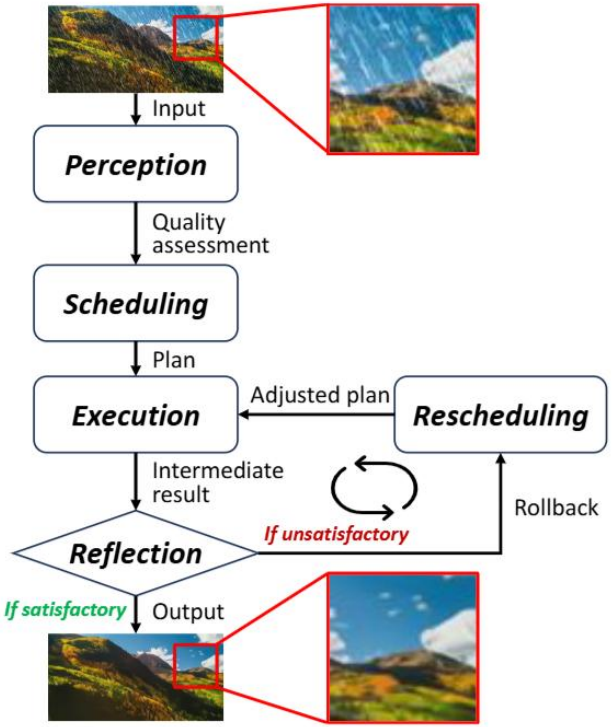
# Introduction

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

- Motivation
  - 실제 Image Restoration문제는 다양한 degradation이 동시에 발생
  - 기존 모델은 특정 degradation만 처리할 수 있어 한계가 있었음
  - 현실에서는 사람이 여러 모델을 조합하여 단계적으로 활용
    - 이 단계들을 Agent system으로 만들어 해결




<**Classic** Image Restoration Model>



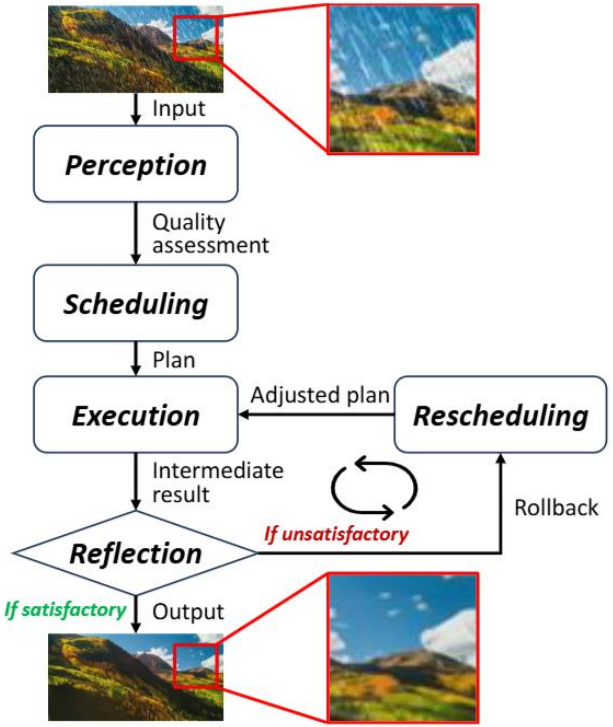



CVSP Lab
Computer Vision & Signal Processing

# Introduction

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

- Motivation
  - 인간의 IR 문제 해결 과정
    - Perception: degradation 종류 파악
    - Scheduling(계획): 어떤 모델을 어떤 순서로 쓸 지 결정
    - Execution(실행): 계획대로 모델을 순차적으로 적용
    - Reflection(반성): 결과 품질 평가
    - Rescheduling(재계획): 성능이 불만족스러우면 계획 수정




  - 필요한 것
    - 이미지 품질 분석 능력
    - 어떤 모델을 써야할 지, 다음 모델을 뭘 써야할 지 결정하는 능력
    - IR 모델 활용 경험 및 지식



CVSP Lab
Computer Vision & Signal Processing

# Introduction

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

- Motivation
  - Image Restoration에서 operation순서의 중요성

Denoising → dehazing



Input image with haze and noise | Result of denoising | Result of first denoising and then dehazing



High-quality image | Result of dehazing | Result of first dehazing and then denoising

Dehazing → Denoising

Deraining → SR



Input image with rain and low-resolution | Result of deraining | Result of first deraining and then SR



High-quality image | Result of SR | Result of first SR and then deraining

SR → Deraining

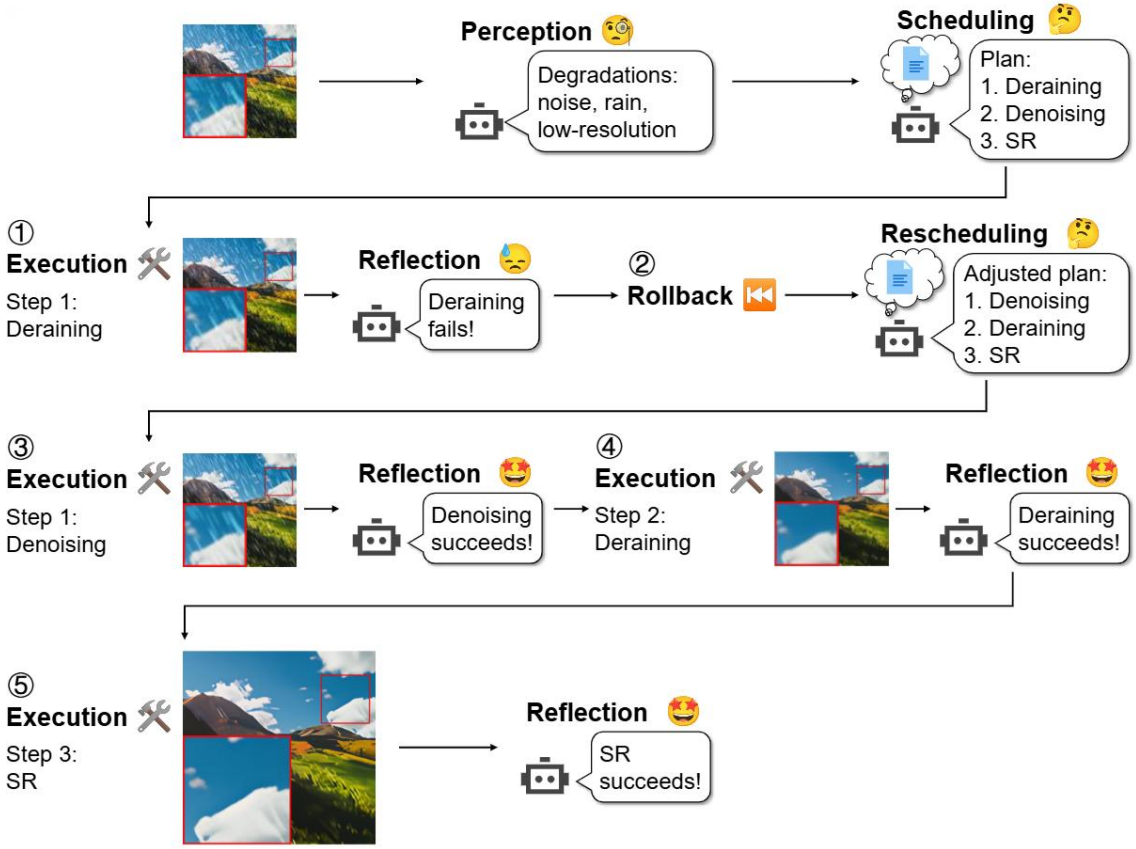


CVSP Lab
Computer Vision & Signal Processing

# Method

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

- Entire Workflow
  - 목표: 최적의 restoration model과 그 순서를 최적화 하여 이미지 복원
  - Perception(인지) → Scheduling(계획) → Execution(수행) → Reflection(재평가, 반성)



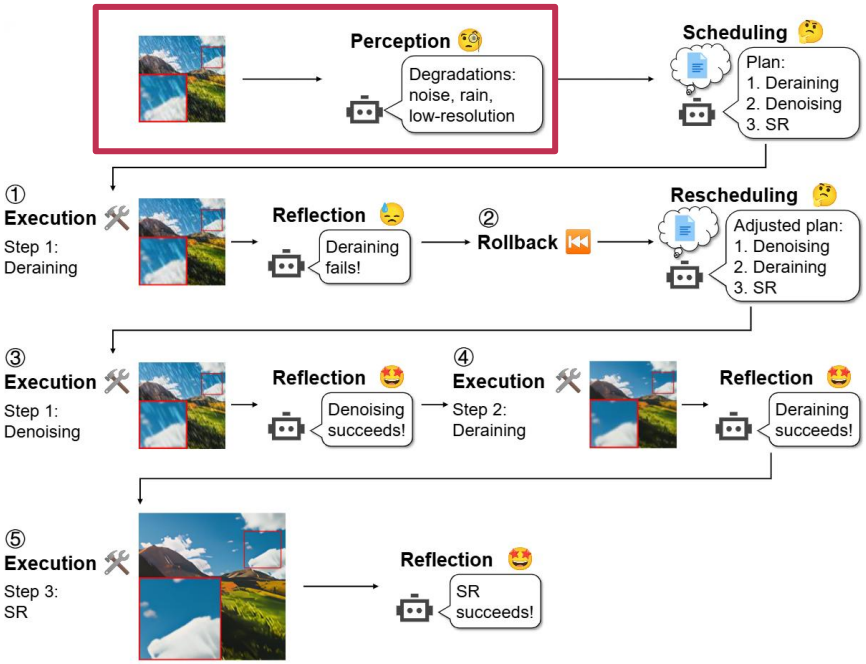



CVSP Lab
Computer Vision & Signal Processing

# Method

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

- Entire Workflow
  - Perception
    - Vision-Language Model(VLM)을 사용하여 이미지의 내용과 Degradation을 분석
      - 예시: noise, rain, low-resolution …



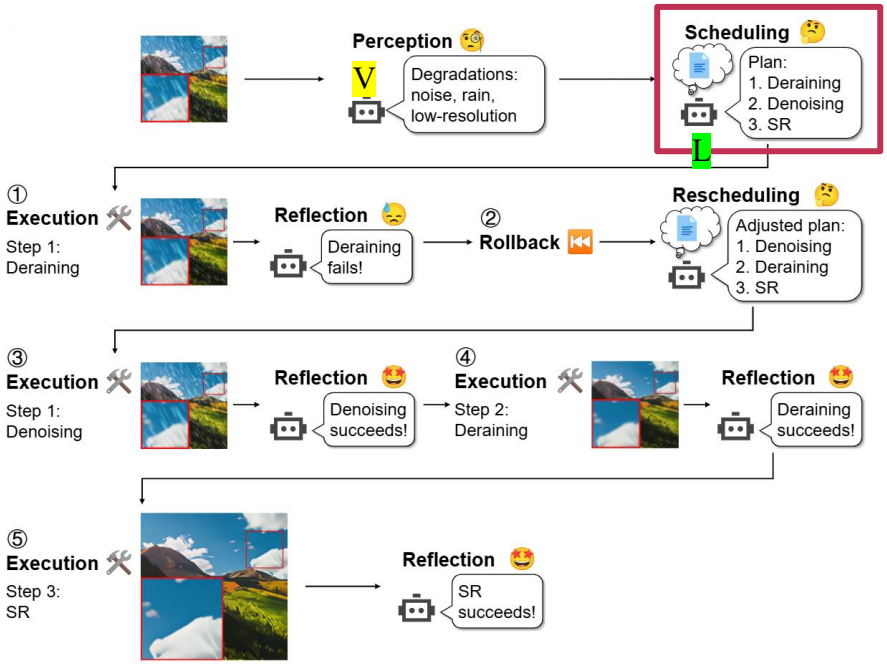



CVSP Lab
Computer Vision & Signal Processing

# Method

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

- Entire Workflow
  - Scheduling
    - 아래 정보를 활용해 LLM이 Image Restoration을 위한 전체적인 Plan을 세움
      - VLM이 Perception한 정보
      - Degradation Image Restoration에 대한 기본 지식
      - Domain 지식



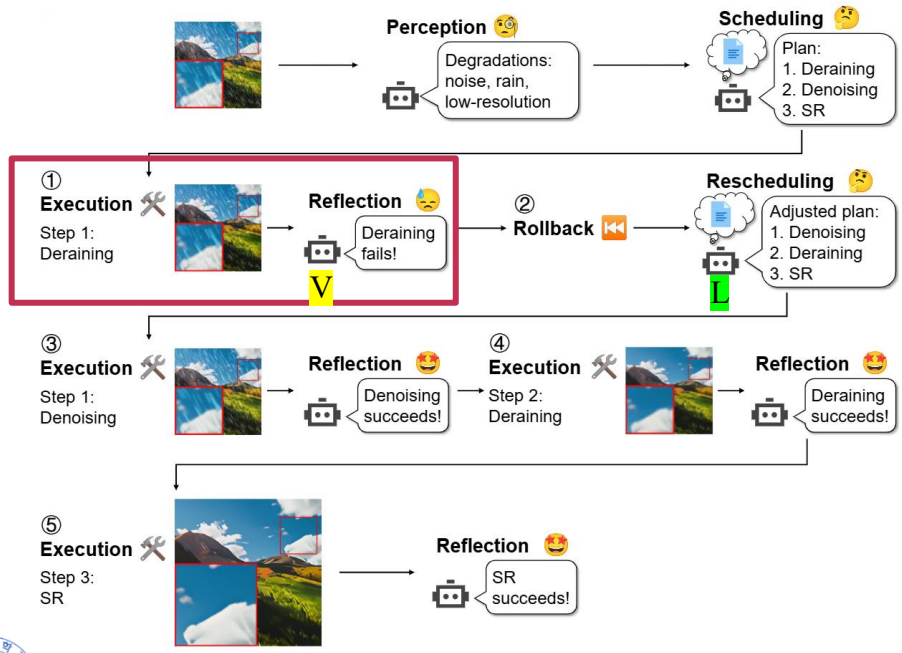


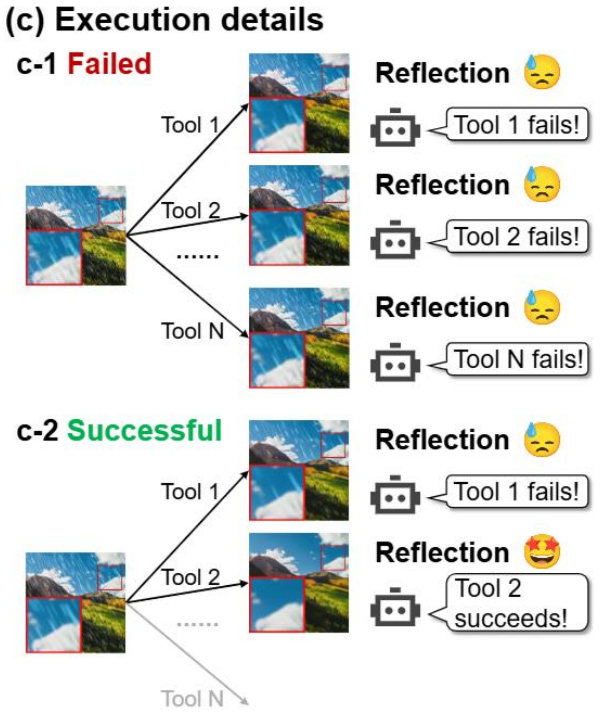

CVSP Lab
Computer Vision & Signal Processing

# Method



- Entire Workflow
  - Execution
    - Agent가 Plan의 첫번째 연산을 수행
    - 목표(Image Restoration)을 위해 순차적으로 tool을 적용



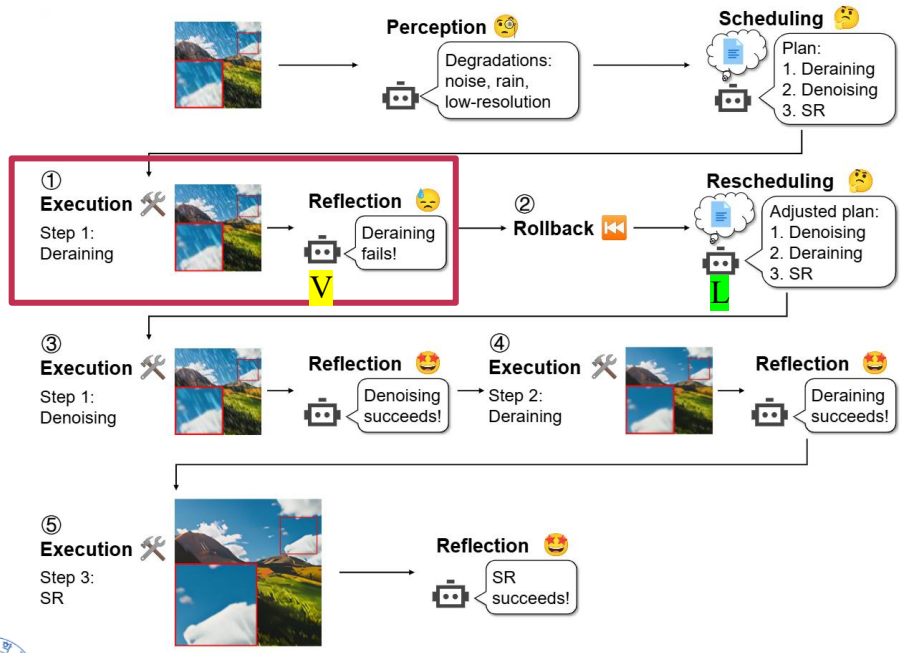


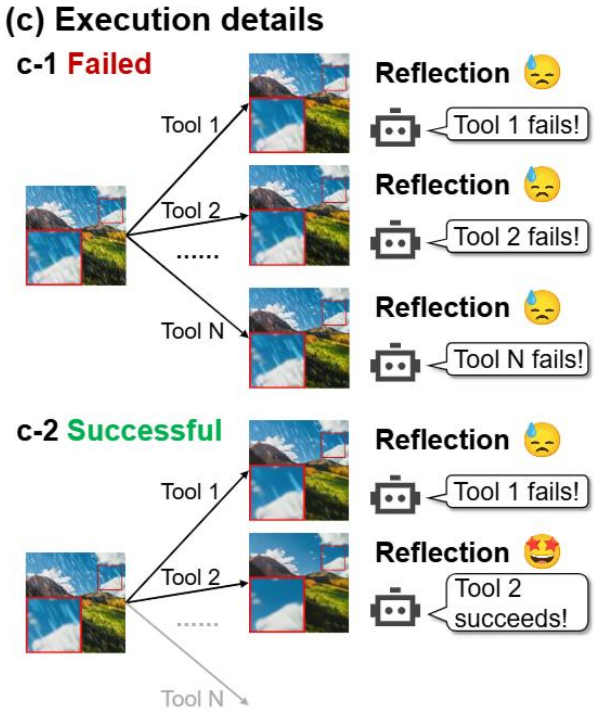

CVSP Lab
Computer Vision & Signal Processing

# Method



- Entire Workflow
  - Reflection
    - Execution에서 수행한 tool이 잘 적용되었는지 이미지 품질 평가
    - 목표가 충족됐으면 Agent는 다음 연산으로 넘어감
    - 모든 연산이 완료될 때 까지 이 과정 반복



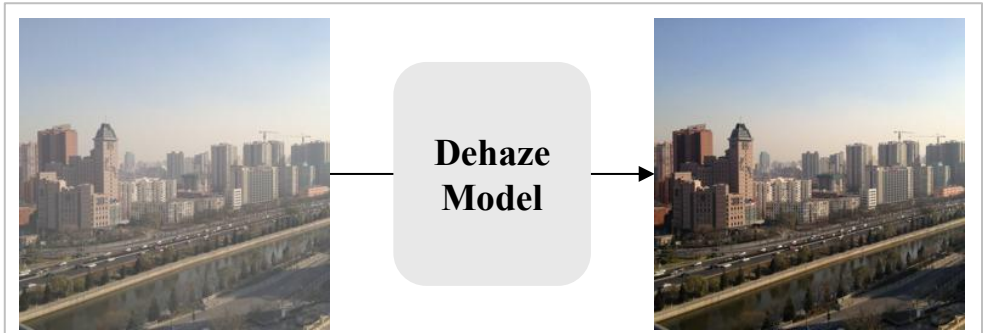



CVSP Lab
Computer Vision & Signal Processing

# Method

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

- Tools setting
  - 실제 Image Restoration 하는 시나리오를 '여러 degradation이 혼합된 집합'으로 가정
  - 실제 환경에서 일어나는 Degradation은 모두 포함하지 못하더라도 일단 아래와 같이 정의
    - Low Resolution(Model 1, 2, 3…)
    - Noise(Model 1, 2, …, 5)
    - Motion Blur
    - Defocus Blur
    - Rain
    - Haze
    - JPEG compression artifacts
    - Low Light




<Image Restoration (Tool) 예시>

  - Agent가 활용가능한 Toolbox 준비
    - 각 Degradation유형마다 '특화 모델'존재 → **각 유형마다 3-6개 모델(Tools) 준비**

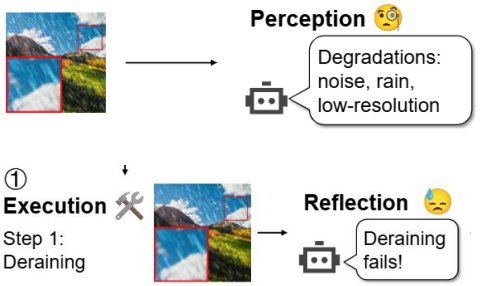


CVSP Lab
Computer Vision & Signal Processing

# Method

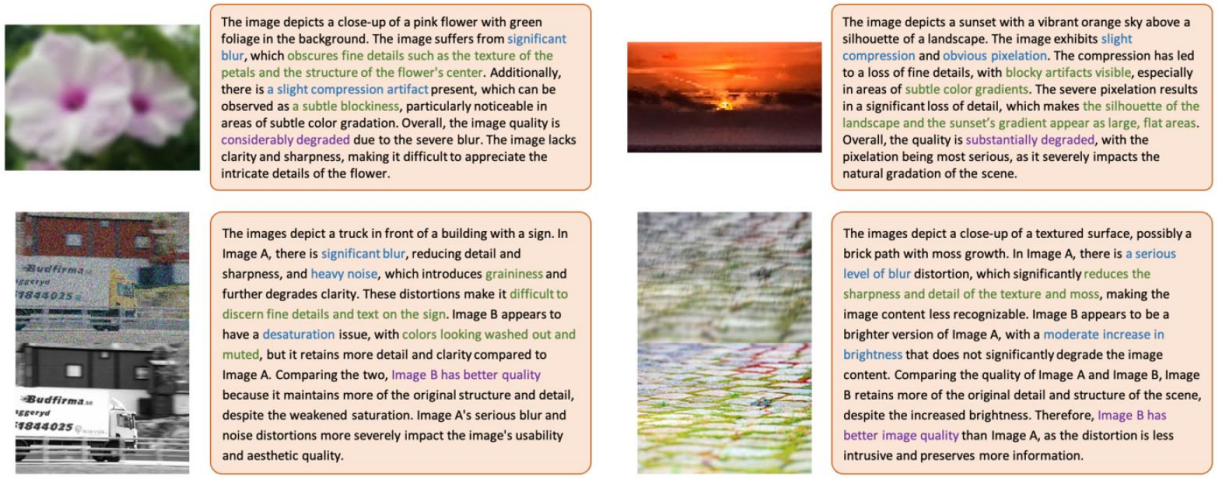
1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).
3)You, Zhiyuan, et al. "Depicting beyond scores: Advancing image quality assessment through multi-modal language models." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
4) Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.

- Ability to Analyze Image Quality
  - VLM 기반의 Image Quality Assesment(IQA) 모델— DepictQA[3] 사용
    - Image Quality와 내용을 자연어로 설명(Scheduling에 사용)
    - Reasoning이 가능하여 Execution의 성공/실패를 판단할 수 있음




<VLM의 품질평가(Perception, Reflection 단계>

  - DepictQA[3]
    - Image 내용을 분석하고 Image Quality를 분석할 수 있는 VLM기반 단일 이미지 품질평가 모델
    - 이 모델을 LoRA[4] finetuning, Degradation 정도를 판단할 수 있도록 finetuning




<DepictQA의 이미지 내용 파악 및 품질 분석 예시>

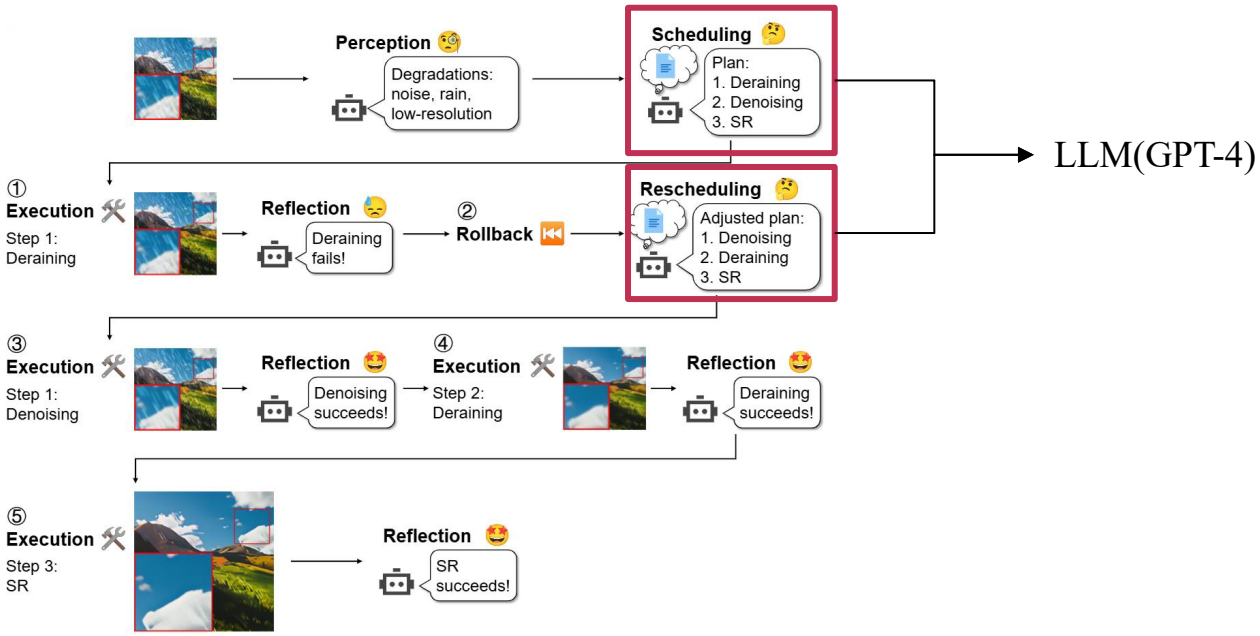


CVSP Lab
Computer Vision & Signal Processing

# Method

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).
3)You, Zhiyuan, et al. "Depicting beyond scores: Advancing image quality assessment through multi-modal language models." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
4) Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.

- Ability to Reason Based on Context
  - Image 품질 평가 내용을 토대로 Scheduling과 Rescheduling을 하려면 깊은 이해가 필요
    - Tool이 많아 경우의 수가 아주 많기 때문에 규칙 기반 접근은 불가능함
    - Agent의 지능적인 사고(Thinking)와 추론(Reasoning)을 위해 LLM을 사용
      - GPT-4(고성능 LLM) 사용(Finetuning 없이)




<AgenticIR의 Entire Workflow – LLM의 추론이 필요한 Scheduling>

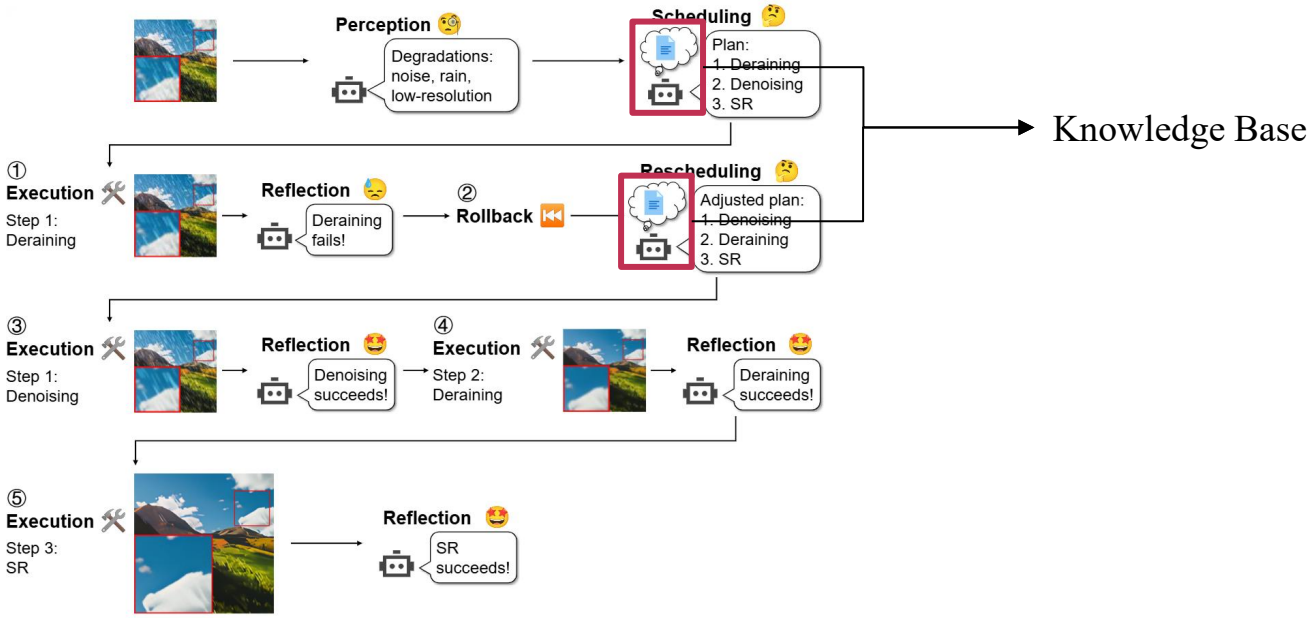


CVSP Lab
Computer Vision & Signal Processing

# Method

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).
3)You, Zhiyuan, et al. "Depicting beyond scores: Advancing image quality assessment through multi-modal language models." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
4) Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.

- Knowledge and Experience in IR
  - LLM(GPT-4)이 참고할만한 Knowledge 및 Experience가 필요함(기존 LLM은 일반적 지식만 있음)
  - 이미지들에 대해 모든 경우의 수의 조합(Sequence)을 적용 후 이미지 품질 평가 모델(VLM-DepictQA)로 성공률 계산
  - 수집된 조합마다의 이미지 품질 통계를 LLM(GPT-4)에게 요약시키고 Knowledge base에 저장
  - 이 경험적 Knowledge base를 참조하여 추론




<AgenticIR의 Entire Workflow – LLM의 추론에 필요한 Knowledge Base>

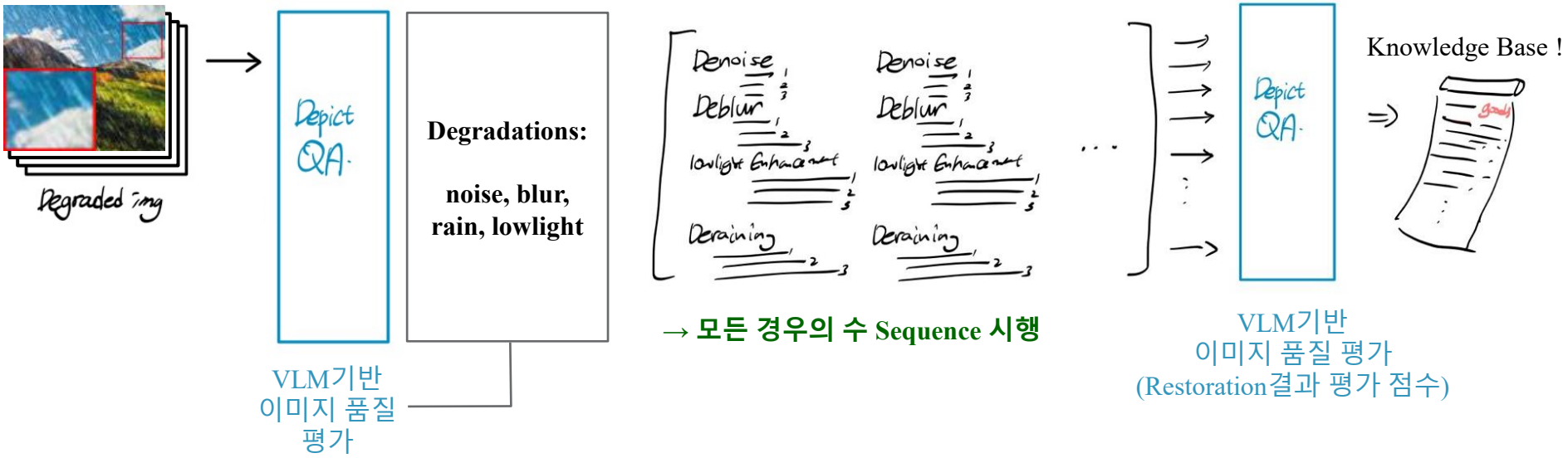


CVSP Lab
Computer Vision & Signal Processing

# Method

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).
3)You, Zhiyuan, et al. "Depicting beyond scores: Advancing image quality assessment through multi-modal language models." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
4) Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.

- Knowledge and Experience in IR
  - Knowledge Base 구축




<AgenticIR의 Entire Workflow – LLM의 추론에 필요한 Knowledge Base>

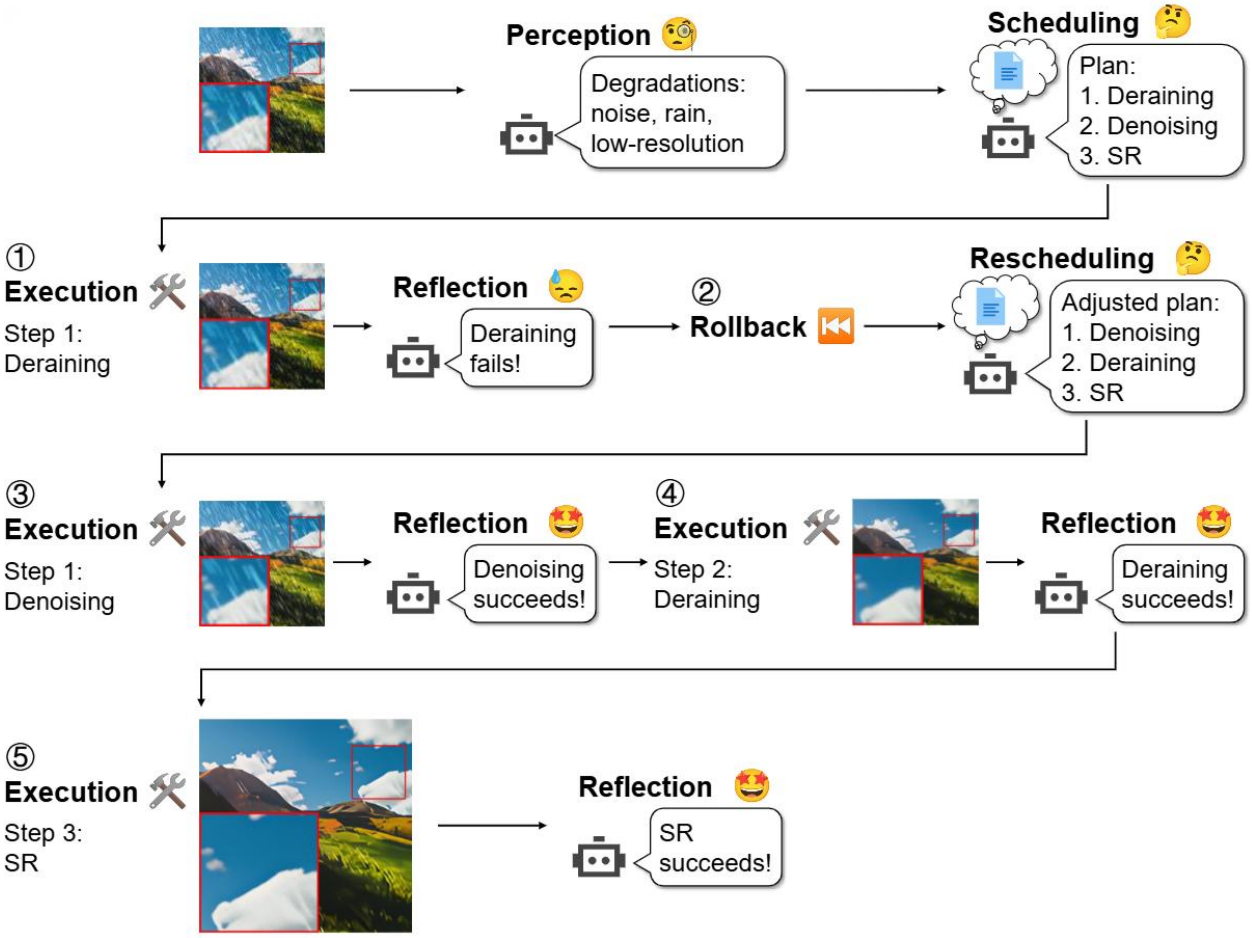


CVSP Lab
Computer Vision & Signal Processing

# Method

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).
3)You, Zhiyuan, et al. "Depicting beyond scores: Advancing image quality assessment through multi-modal language models." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
4) Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.

- Entire Workflow






CVSP Lab
Computer Vision & Signal Processing

# Experiments

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

- Quantitative Results
  - Knowledge base기반 계획과 랜덤 셔플
    - 전체 탐색을 통해 얻은 경험 기반의 Planning 효과
  - Group A, B, C
    - A: Agent가 Experience를 하여 Knowledge base를 쌓는 데 쓰는 조합
    - B: 테스트 전용. Group A에서와 같은 복잡도이지만 새로운 조합
    - C: 테스트 전용. A, B보다 더 복잡한 상황을 시뮬레이션

| Degradations | As planned | PSNR | SSIM | LPIPS↓ | MANIQA | CLIP-IQA | MUSIQ |
|---|---|---|---|---|---|---|---|
| Group A | ✓ | **21.14** | **0.6836** | **0.2753** | **0.3469** | **0.5091** | **60.77** |
| | ✗ | 20.79 | 0.6652 | 0.3060 | 0.3385 | 0.4819 | 59.85 |
| Group B | ✓ | **21.14** | **0.7088** | **0.2683** | 0.3588 | **0.5275** | **61.92** |
| | ✗ | 20.32 | 0.6811 | 0.2976 | **0.3623** | 0.5257 | 60.15 |
| Group C | ✓ | **18.78** | **0.5352** | **0.4239** | **0.3118** | **0.4876** | 51.08 |
| | ✗ | 18.49 | 0.5277 | 0.4345 | 0.3058 | 0.4719 | **51.32** |

# Experiments

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

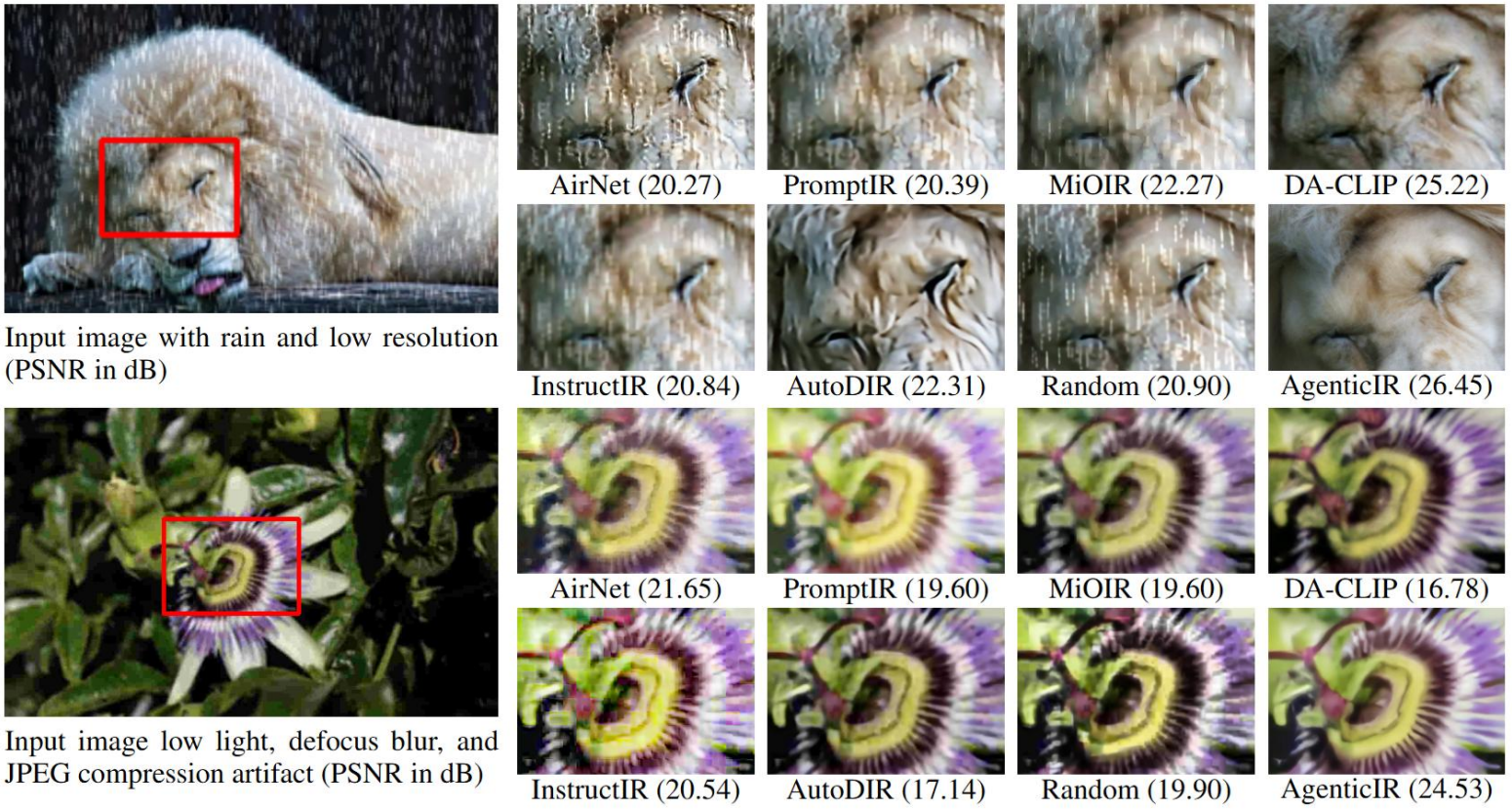
- Quantitative Results
  - Plan과 다르게 무작위로 planning한 후 수행한 것과 성능 비교
    - PSNR, SSIM, LPIPS, MANIQA 외 2개의 BIQA 모델(GT 없이 이미지 품질 평가 모델)
  - Group A, B, C
    - A: Agent가 Experience를 하여 Knowledge base를 쌓는 데 쓰는 조합
    - B: 테스트 전용. Group A에서와 같은 복잡도이지만 새로운 조합
    - C: 테스트 전용. A, B보다 더 복잡한 상황을 시뮬레이션

| Degradations | Method | PSNR | SSIM | LPIPS↓ | MANIQA | CLIP-IQA | MUSIQ |
|---|---|---|---|---|---|---|---|
| Group A | AgenticIR | **21.04** | **0.6818** | **0.3148** | **0.3071** | **0.4474** | **56.88** |
| | Random | 20.90 | 0.6642 | 0.3368 | 0.2963 | 0.4394 | 55.30 |
| Group B | AgenticIR | **20.55** | **0.7009** | **0.3072** | **0.3204** | **0.4648** | **57.57** |
| | Random | 20.06 | 0.6766 | 0.3351 | 0.3120 | 0.4514 | 56.15 |
| Group C | AgenticIR | 18.82 | **0.5474** | **0.4493** | **0.2698** | **0.3948** | **48.68** |
| | Random | **18.87** | 0.5456 | 0.4796 | 0.2354 | 0.3543 | 44.61 |

# Experiments



- Qualitative Results
  - Comparisons with other methods






CVSP Lab
Computer Vision & Signal Processing

# Experiments

1)Zhu, Kaiwen, et al. "An intelligent agentic system for complex image restoration problems." arXiv preprint arXiv:2410.17809 (2024).

- Qualitative Results
  - Comparisons with other methods
    - AgenticIR과 deraining model, dehazing model, All-in-one model 비교



Input | Restormer (Deraining) | DehazeFormer (Dehazing) | AutoDIR | AgenticIR



CVSP Lab
Computer Vision & Signal Processing

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

# RestoreAgent: Autonomous Image Restoration Agent via Multimodal Large Language Models [2)]

NeurIPS 2024

# Introduction

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Introduction to paper
  - RestoreAgent: Autonomous image restoration agent via multimodal large language models
    - VLM을 finetuning하여 Agentic system을 설계하고 Restoration을 수행한 논문
    - 탄탄한 Introduction이 특징

## RestoreAgent: Autonomous Image Restoration Agent via Multimodal Large Language Models

**Haoyu Chen**[1], **Wenbo Li**[2], **Jinjin Gu**[3], **Jingjing Ren**[1], **Sixiang Chen**[1], **Tian Ye**[1], **Renjing Pei**[2], **Kaiwen Zhou**[2], **Fenglong Song**[2], **Lei Zhu**[1,4*]

[1]The Hong Kong University of Science and Technology (Guangzhou) [2]Huawei Noah's Ark Lab
[3]The University of Sydney [4]The Hong Kong University of Science and Technology

Project page: https://haoyuchen.com/RestoreAgent



CVSP Lab Computer Vision & Signal Processing

# Introduction

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Introduction to paper
  - RestoreAgent 제안
    - VLM에 기반한 지능적인 Image Restoration system
    - Framework
      - Degradation type Identification
      - Adaptive Restoration Sequence
      - Optimal Model Selection
      - Automated Execution



# Introduction

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Motivation
  - 다양한 Degradation 유형을 가진 Image를 복원하기 위해 다양한 model들이 필요함
  - 이 model들을 실행하는 순서와 조합은 최종 복원 결과에 큰 영향을 미침
  - 최근에는 이 모델들을 통합하는 All-in-one model을 개발하거나 Image restoration 전문가 지식을 활용함

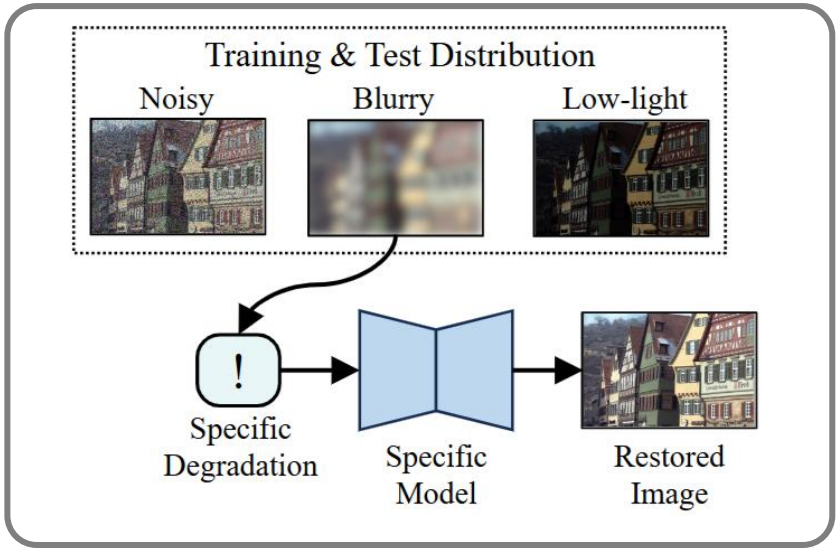


CVSP Lab
Computer Vision & Signal Processing

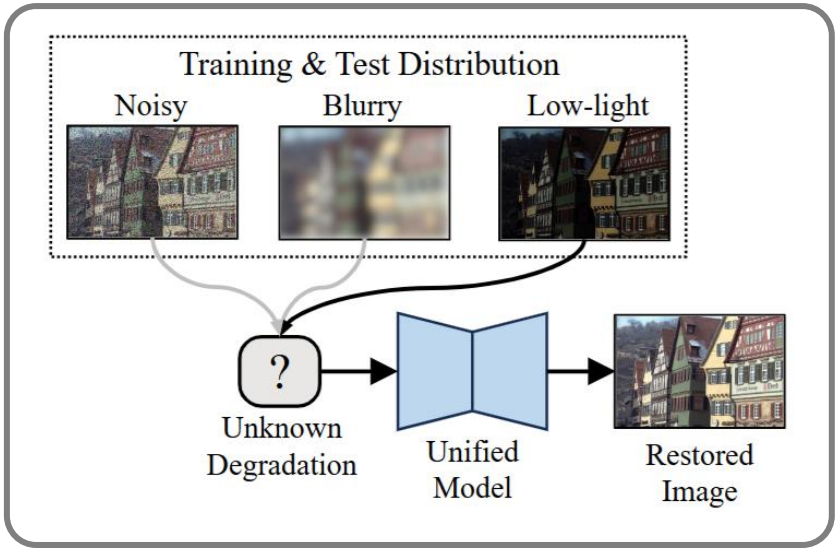
# Introduction



- Motivation
  - All-in-one Image Restoration model과 비교
    - 단일 프레임워크를 통해 여러 degradation을 동시에 처리하고자 함




<**Classic** Image Restoration Model>

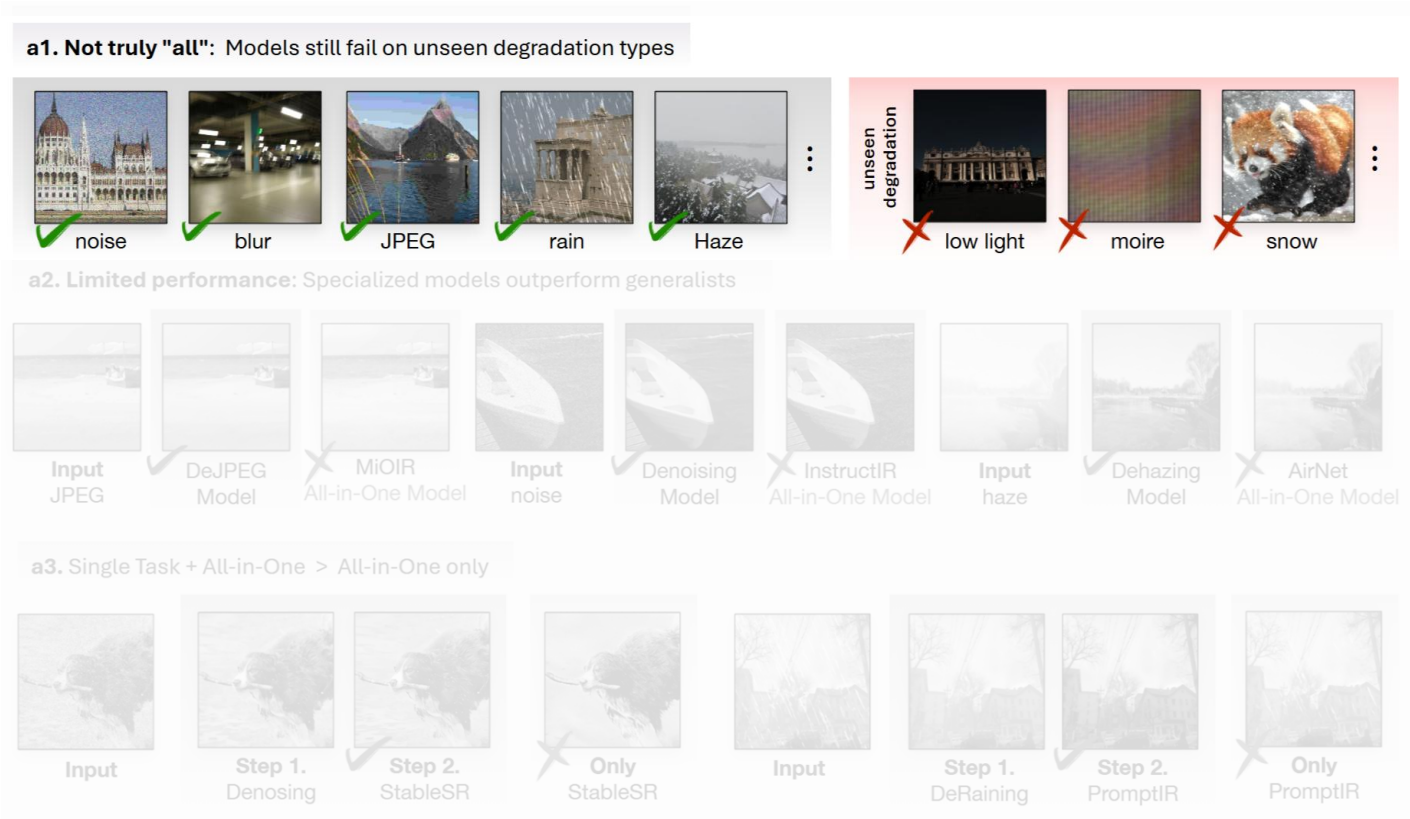



<**All-in-one** Image Restoration Model>

# Introduction

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Motivation
  - All-in-one Image Restoration model과 비교
    - 단일 프레임워크를 통해 여러 degradation을 동시에 처리하고자 함
    - 하지만 unseen degradation에 약하고 Single-task restoration model에 비해 덜 효과적임



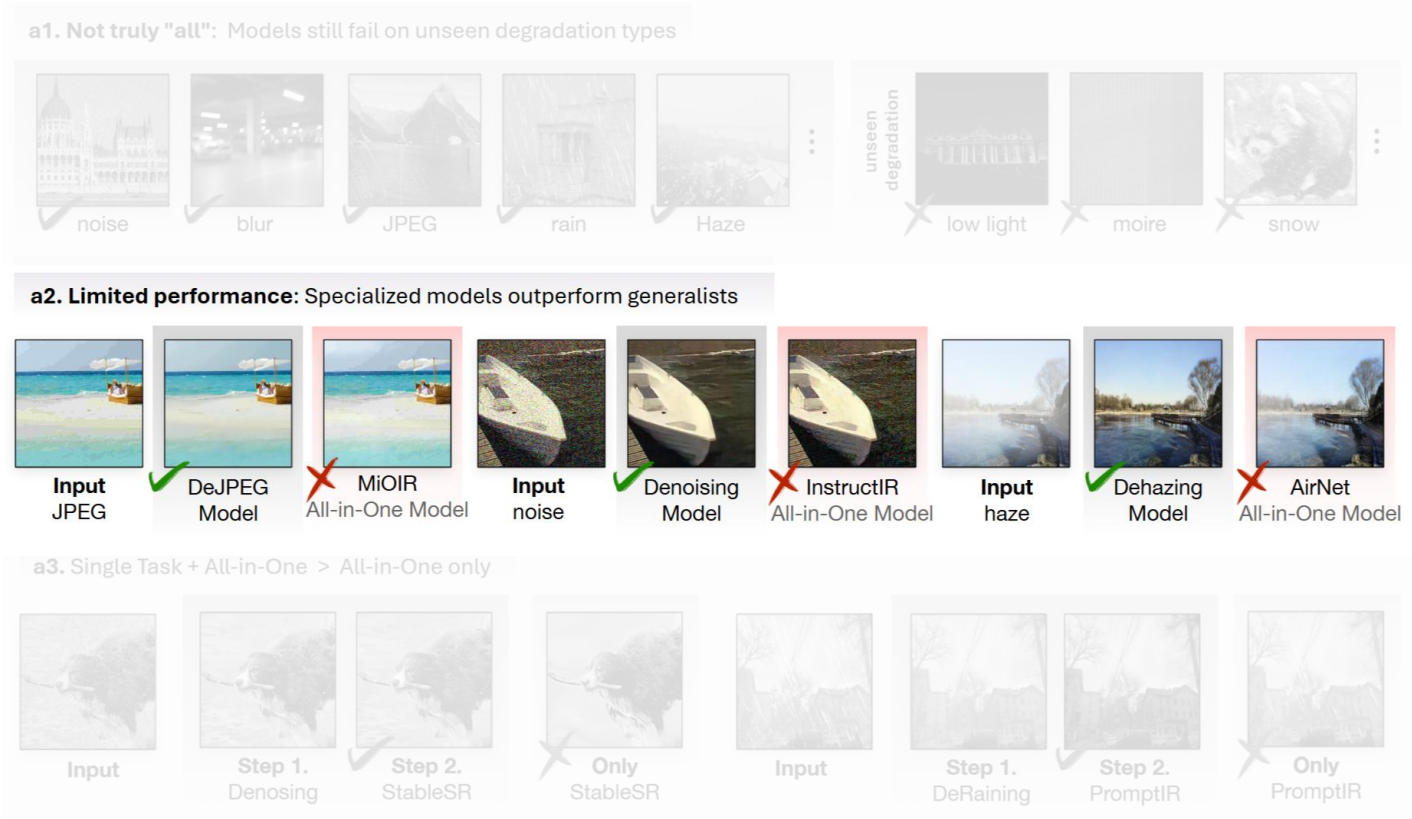



CVSP Lab Computer Vision & Signal Processing

# Introduction

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Motivation
  - All-in-one Image Restoration model과 비교
    - 단일 프레임워크를 통해 여러 degradation을 동시에 처리하고자 함
    - 하지만 unseen degradation에 약하고 Single-task restoration model에 비해 덜 효과적임



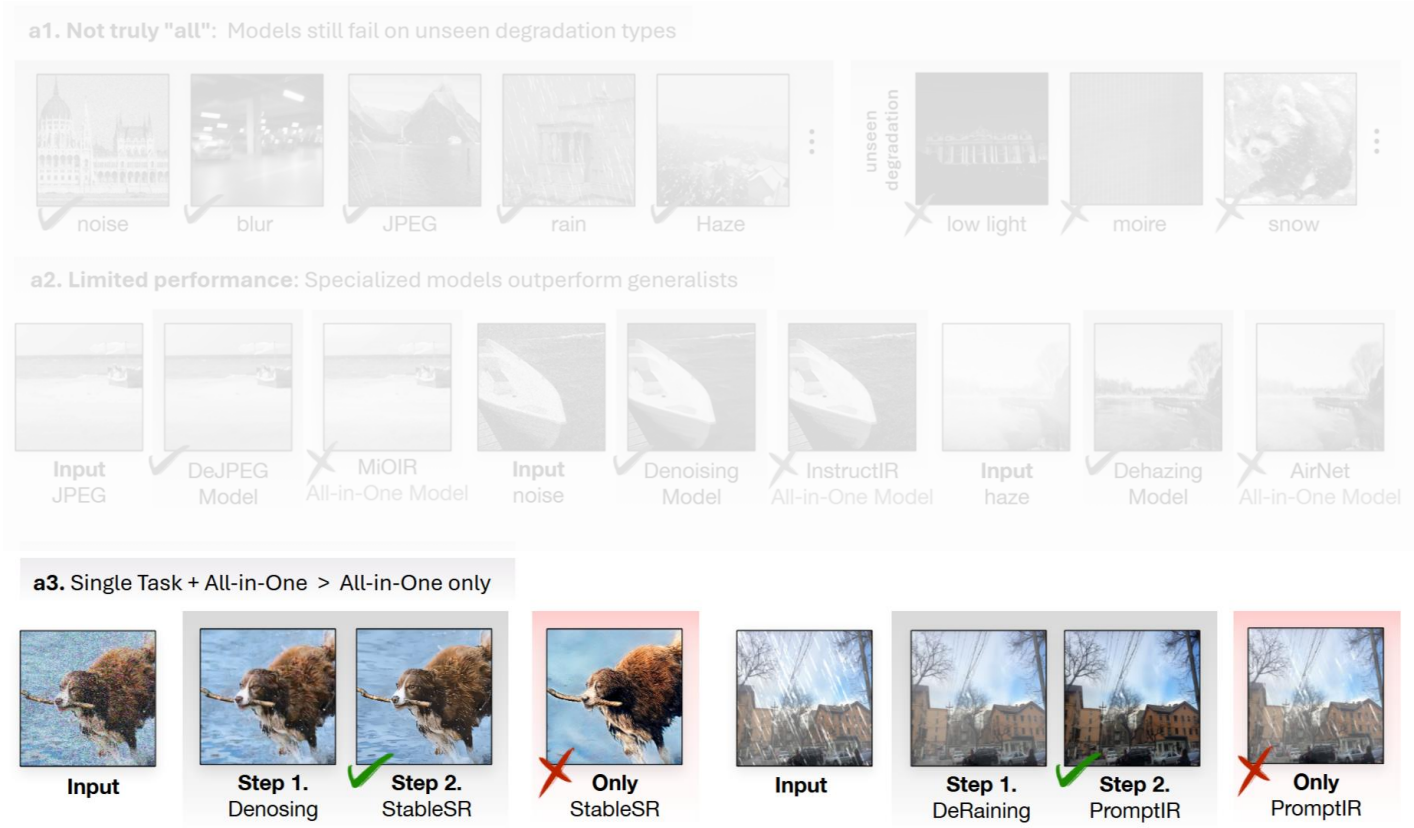



CVSP Lab
Computer Vision & Signal Processing

# Introduction

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Motivation
  - All-in-one Image Restoration model과 비교
    - 단일 프레임워크를 통해 여러 degradation을 동시에 처리하고자 함
    - 하지만 unseen degradation에 약하고 Single-task restoration model에 비해 덜 효과적임



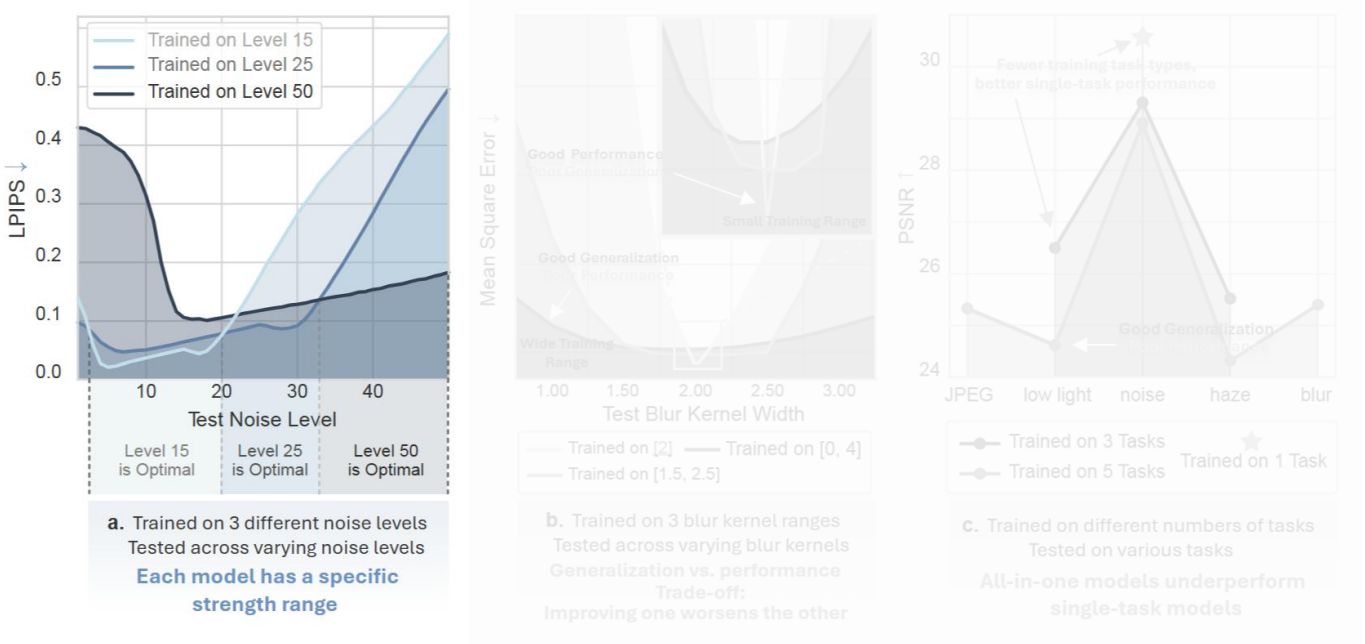



CVSP Lab Computer Vision & Signal Processing

# Introduction

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Motivation
  - All-in-one model 대신 Task-specific model을 써야 하는 이유 설득
    - (a) 세 가지 noise level에서 학습된 모델들이 학습된 근처 수준에서는 성능이 좋지만 다른 noise level에서는 성능이 급격히 떨어짐
      - 학습한 dataset의 분포와 test set의 분포가 다르면 성능이 급격히 떨어짐



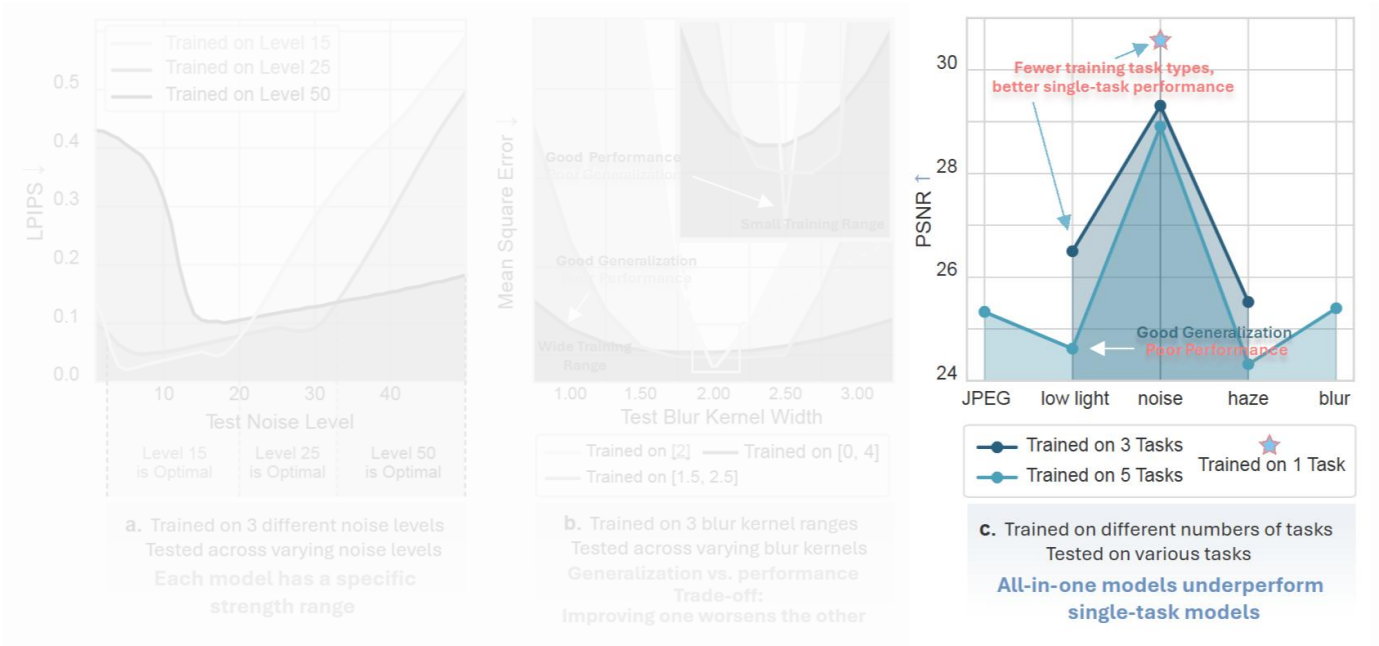



CVSP Lab
Computer Vision & Signal Processing

# Introduction

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Motivation
  - All-in-one model 대신 Task-specific model을 써야 하는 이유 설득
    - (c) 1개 Task, 3개 Task, 5개 Task로 학습된 모델들을 다양한 degradation에 대해 test
      - 적은 Task로 학습된 모델이 개별 Task 성능이 좋음
      - 많은 Task를 학습한 모델은 여러 Task에서 두루 쓸 수 있지만, 각 task에 대해서는 specific-Task model이 나음



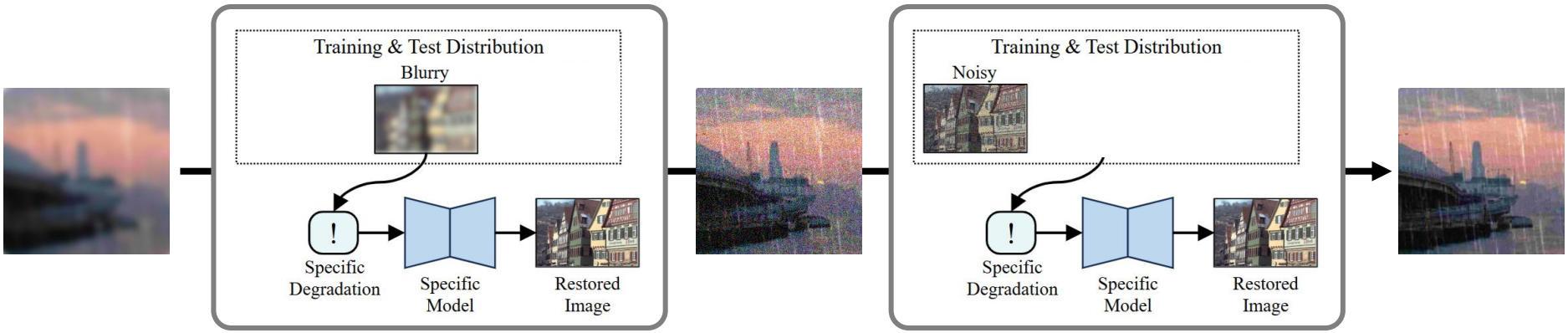



CVSP Lab
Computer Vision & Signal Processing

# Introduction

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Motivation
  - Task-specific models의 조합으로 풀자
    - 여러 Degradation 유형을 하나의 model로 Restoration하는 것에는 한계가 있음
    - 특정 Degradation을 restoration하는 것에 집중한 model들을 결합하여 문제 해결
    - 각 특화 model은 담당하는 특정 Degradation에서 강점을 보이기 때문에 더 좋은 결과를 얻을 수 있음




<Deblur Model> <Denoise Model>

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

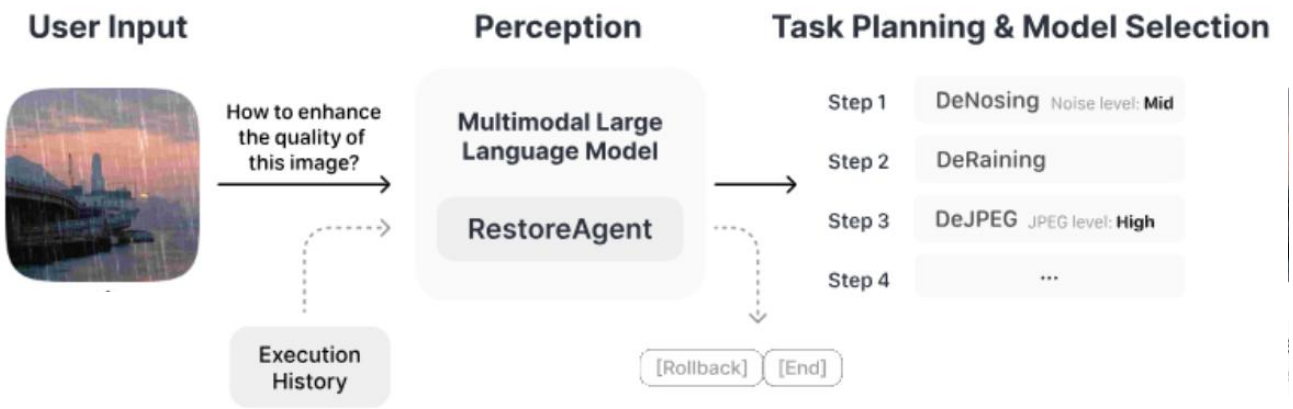
- Overall Framework
  - RestoreAgent의 목적
    - 복잡한 Degradation 상황에서 전문가 수준의 **Adaptive한 Restoration pipeline을 자동으로 생성**함
  - 전체 과정
    1) Degraded Image 입력
    2) MLLM 기반 Degradation 분석(노이즈, 블러, JPEG 등)
    3) 최적의 작업 순서와 모델을 선택(Task Planning)
    4) 단계별 실행 및 피드백 반영 (Rollback/Replanning) 지원
    5) 최종 복원 이미지 출력



# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- RestoreAgent[2] 구현
  - Backbone Model: LLaVA-Llama-8B 사용
  - Vision module과 Language module을 finetuning
    - Input: (instruction) How to enhance the quality of this image?
    - Output: 1.Denoising(Mid), 2.Deraining, 3.DeJPEG(High)






CVSP Lab
Computer Vision & Signal Processing

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

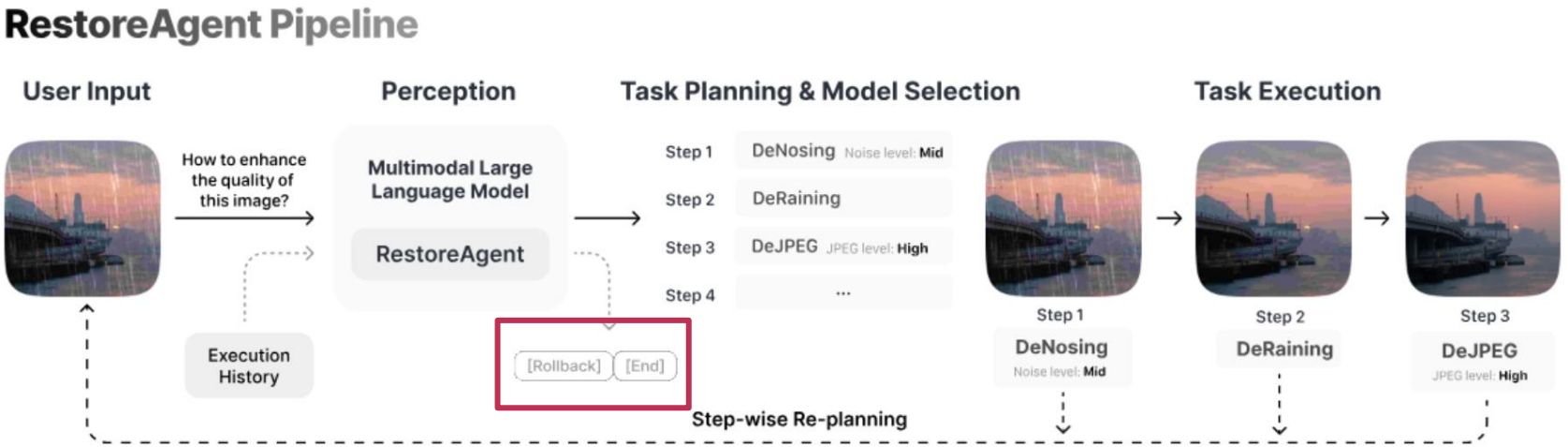
- RestoreAgent[2] 구현
  - 각 복원단계 후 이미지의 상태를 다시 평가하고 Rescheduling을 할 수 있음
  - 재평가 과정에서 실행했던 Restoration 이력이 들어가서 다음 결정에 영향을 줌
  - Rollback 기능 포함
    - 계획대로 수행했는데 중간 결과에서 restoration 품질이 좋지 않으면 이전 상태로 되돌릴 수 있음
  - 반복적 평가와 실행했던 history제공, Rollback기능을 결합하여 진행 중간중간 수정이 가능하도록 함
    - End기능으로 최적의 결과를 찾았을 때 pipeline을 종료함



# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

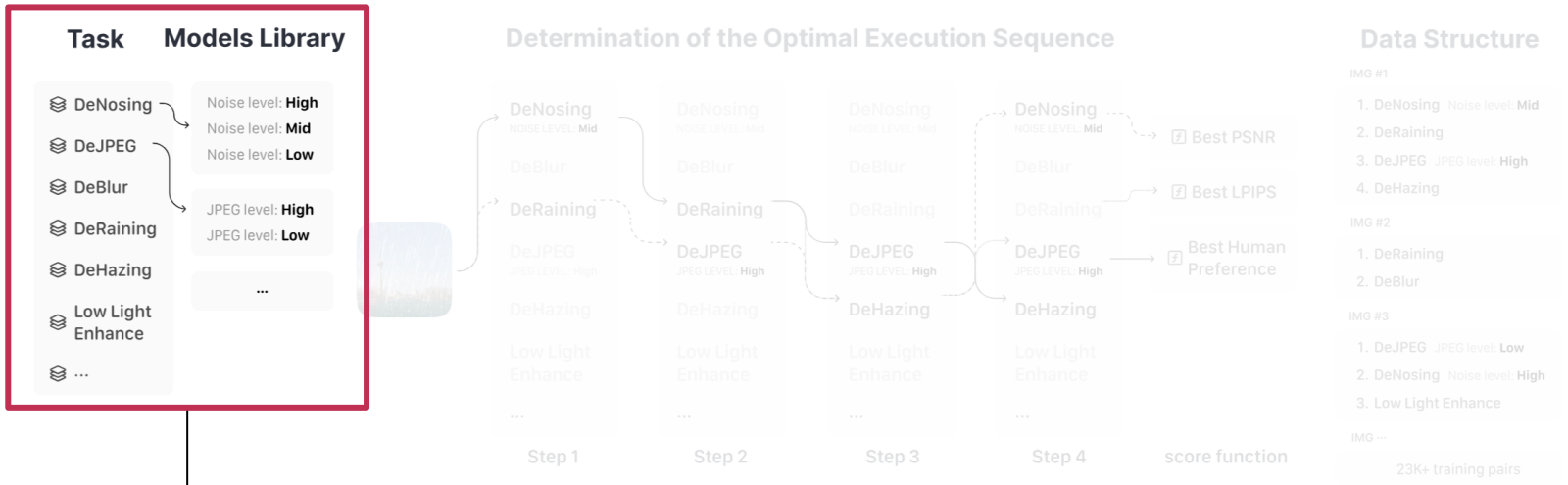
- RestoreAgent[2] 구현
  - 각 복원단계 후 이미지의 상태를 다시 평가하고 Rescheduling을 할 수 있음
  - 재평가 과정에서 실행했던 Restoration 이력이 들어가서 다음 결정에 영향을 줌
  - Rollback 기능 포함
    - 계획대로 수행했는데 중간 결과에서 restoration 품질이 좋지 않으면 이전 상태로 되돌릴 수 있음
  - 반복적 평가와 실행했던 history제공, Rollback기능을 결합하여 진행 중간중간 수정이 가능하도록 함
    - End기능으로 최적의 결과를 찾았을 때 pipeline을 종료함



# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - 포괄적인 Degradation 유형 집합 $D = \{d_1, d_2, \cdots, d_n\}$ 정의
    - 여기서 $d_i$는 Noise, Blur, Rain, Low Light 와 같은 Degradation 유형들을 나타냄
  - 각 Degradation 유형 $d_i$마다 모델 라이브러리(Tools) $M_{d_i}$를 마련
    - $M_{d_i} = \{M_{d_i}^1, M_{d_i}^2, \cdots\}$




< 다양한 Restoration model 적용을 위해 Model Library 마련 >

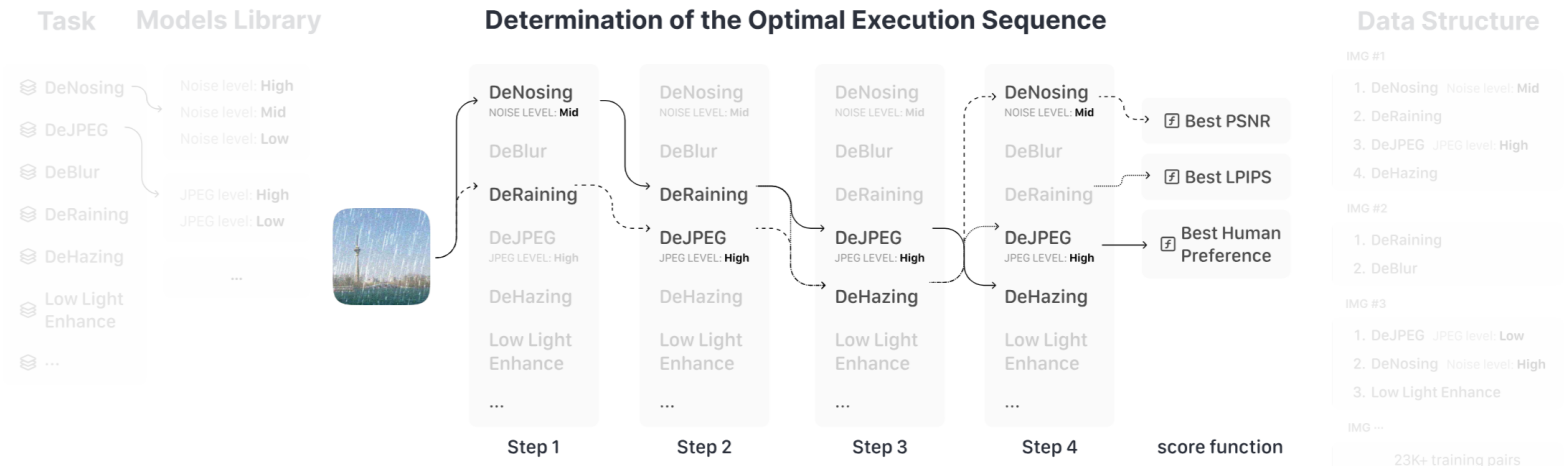


CVSP Lab
Computer Vision & Signal Processing

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - 전체 가능한 Sequence를 모두 수행하여 포괄적인 탐색(Comprehensive search)
    - 모든 가능한 Restoration 실행 순서와 모델 조합의 permutation을 생성
      - 최고 성능을 가지는 복원 sequence 선별하여 학습데이터로 선정
      - 23K 이상의 이미지 데이터셋으로 수행




<실행 가능한 모든 조합을 모두 수행하여 품질 평가>

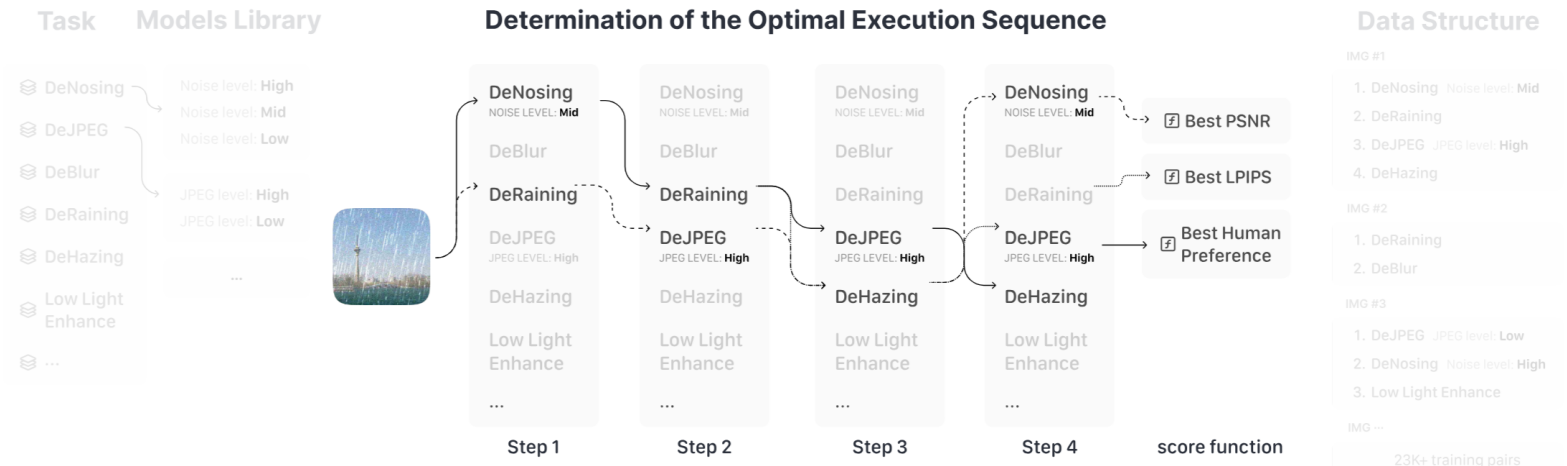


CVSP Lab
Computer Vision & Signal Processing

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - 모든 경우의 pipeline을 이미지에 적용 및 restoration 결과 품질을 $S(I,\sigma)$로 평가
    - PSNR, LPIPS, Human Preference를 종합하거나 선택하여 평가
    - 모든 경우의 수들을 수행했을 때 점수 S를 비교하여 가장 높은 점수를 얻은 파이프라인을 해당 이미지 $I$의 최적의 restoration pipeline $\sigma$로 선택




<실행 가능한 모든 조합을 모두 수행하여 품질 평가>



CVSP Lab
Computer Vision & Signal Processing

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - 총 다섯 개의 시나리오 구성
    - RestoreAgent가 상황에 따라 결정을 할 수 있도록 데이터셋에 다섯 개의 시나리오를 포함하여 MLLM finetuning

| | | Input | Answer | Function |
|---|---|---|---|---|
| 1 (Primary) | Input | How to enhance the quality of this image? Execution history: None. | 1. Denoising, low noise level, 2. Dehazing, 3. DeJPEG, high quality factor. | Initiate full enhancement sequences for degraded images. |
| 2 | 1. Denoising | How to enhance the quality of this image? Execution history: 1. Denoising, low noise level. | 1. Denoising, low noise level, 2. Dehazing, 3. DeJPEG, high quality factor | Dynamically adjust strategies based on intermediate results. |
| 3 | ~~1. Denoising 2. Dehazing~~ | How to enhance the quality of this image? Execution history: 1. Denoising, low noise level, 2. Dehazing | Rollback. | Identify and correct suboptimal steps through rollback mechanisms. |
| 4 | Input | How to enhance the quality of this image? Execution history: 1. Denoising, low noise level. Rollback from Dehazing. | 1. Denoising, low noise level, 2. DeJPEG, high quality factor, 3. Dehazing. | Avoid repetition of ineffective procedures post-rollback. |
| 5 | 1. Denoising 2. DeJPEG 3. Dehazing | How to enhance the quality of this image? Execution history: 1. Denoising, low noise level, 2. DeJPEG, high quality factor, 3. Dehazing. | Stop. | Recognize when image quality has reached its optimal state. |

<MLLM tuning을 위한 데이터셋 - 다섯가지의 시나리오>

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - Scenario 1(초기 Restoration 시나리오)
    - Image와 Instruction이 주어졌을 때 최적의 pipeline을 계획하는 능력 학습
      - Input: **복원할 Degraded Image** + Instruction*
      - Output: 전체탐색 중 최고 점수를 받은 pipeline들을 정답으로 설정
      - 최소 23k pairs로 구성

| | | Input | Answer | Function |
|---|---|---|---|---|
| 1 (Primary) | Input | How to enhance the quality of this image? Execution history: None. | 1. Denoising, low noise level,<br>2. Dehazing,<br>3. DeJPEG, high quality factor. | Initiate full enhancement sequences for degraded images. |

<MLLM tuning을 위한 데이터셋 – Scenario 1>

Instruction* : How to enhance the quality of this image?



CVSP Lab
Computer Vision & Signal Processing

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - Scenario 2(개선 시나리오)
    - Restoration pipeline을 수행하는 중간에 유연하게 pipeline을 수정하는 능력 학습
      - Input: **pipeline이 부분적으로 처리된 Image** + **실행 이력(history)** + Instruction*
      - Output: 그 단계 이후 최적의 pipeline을 정답으로 설정

| | | Input | Answer | Function |
|---|---|---|---|---|
| 2 | 1. Denoising | How to enhance the quality of this image?<br>Execution history: 1. Denoising, low noise level. | 1. Denoising, low noise level,<br>2. Dehazing,<br>3. DeJPEG, high quality factor | Dynamically adjust strategies based on intermediate results. |

<MLLM tuning을 위한 데이터셋 – Scenario 2>

Instruction* : How to enhance the quality of this image?



CVSP Lab
Computer Vision & Signal Processing

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - Scenario 3(Rollback 시나리오)
    - 특정 단계에서 잘못된 선택을 인식하고 되돌아가는 결정 학습
      - Input: **전체 탐색 중 결과 품질점수가 가장 낮은 pipeline의 결과 Image** + **실행 이력(history)** + Instruction*
      - Output: [Rollback]

| | | Input | Answer | Function |
|---|---|---|---|---|
| 3 |  | How to enhance the quality of this image?<br>Execution history:<br>1. Denoising, low noise level, 2. Dehazing | Rollback. | Identify and correct suboptimal steps through rollback mechanisms. |

<MLLM tuning을 위한 데이터셋 – Scenario 3>

Instruction* : How to enhance the quality of this image?

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - Scenario 4(Rollback 후 개선 시나리오)
    - 동일한 실수를 반복하지 않고 더 나은 전략으로 수정하는 능력 학습
      - Input: **[Rollback]이 실행 이력에 포함된 pipeline을 거친 결과 Image** + **실행 이력(history)** + Instruction*
      - Output: Rollback 이후 수정된 pipeline을 정답으로 설정

| | | Input | Answer | Function |
|---|---|---|---|---|
| 4 | Input | How to enhance the quality of this image? Execution history: 1. Denoising, low noise level. Rollback from Dehazing. | 1. Denoising, low noise level, 2. DeJPEG, high quality factor, 3. Dehazing. | Avoid repetition of ineffective procedures post-rollback. |

<MLLM tuning을 위한 데이터셋 – Scenario 4>

Instruction* : How to enhance the quality of this image?

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - Scenario 5(End 시나리오)
    - 최적의 Pipeline을 거쳐 더 이상 개선이 필요 없을 때 [End]를 뱉는 능력
      - Input: **전체 탐색 중 최고 점수를 받은 최종 결과 Image** + **실행 이력(history)** + Instruction*
      - Output: [End]

| | | Input | Answer | Function |
|---|---|---|---|---|
| 5 | 1. Denoising<br>2. DeJPEG<br>3. Dehazing | How to enhance the quality of this image?<br>Execution history:<br>1. Denoising, low noise level,<br>2. DeJPEG, high quality factor, 3. Dehazing. | Stop. | Recognize when image quality has reached its optimal state. |

<MLLM tuning을 위한 데이터셋 – Scenario 5>

Instruction* : How to enhance the quality of this image?

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - 총 다섯 개의 시나리오 정리
    - RestoreAgent가 상황에 따라 결정을 할 수 있도록 데이터셋에 다섯 개의 시나리오를 포함하여 MLLM finetuning

| | | Input | Answer | Function |
|---|---|---|---|---|
| 1 (Primary) | Input | How to enhance the quality of this image? Execution history: None. | 1. Denoising, low noise level, 2. Dehazing, 3. DeJPEG, high quality factor. | Initiate full enhancement sequences for degraded images. |
| 2 | 1. Denoising | How to enhance the quality of this image? Execution history: 1. Denoising, low noise level. | 1. Denoising, low noise level, 2. Dehazing, 3. DeJPEG, high quality factor | Dynamically adjust strategies based on intermediate results. |
| 3 | ~~1. Denoising 2. Dehazing~~ | How to enhance the quality of this image? Execution history: 1. Denoising, low noise level, 2. Dehazing | Rollback. | Identify and correct suboptimal steps through rollback mechanisms. |
| 4 | Input | How to enhance the quality of this image? Execution history: 1. Denoising, low noise level. Rollback from Dehazing. | 1. Denoising, low noise level, 2. DeJPEG, high quality factor, 3. Dehazing. | Avoid repetition of ineffective procedures post-rollback. |
| 5 | 1. Denoising 2. DeJPEG 3. Dehazing | How to enhance the quality of this image? Execution history: 1. Denoising, low noise level, 2. DeJPEG, high quality factor, 3. Dehazing. | Stop. | Recognize when image quality has reached its optimal state. |

<VLM tuning을 위한 데이터셋 - 다섯가지의 시나리오>



CVSP Lab
Computer Vision & Signal Processing

# Method

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Overall Framework 정리
  1) Degraded Image 입력
  2) MLLM 기반 Degradation 분석(노이즈, 블러, JPEG 등)
  3) 최적의 작업 순서와 모델을 선택(Task Planning)
  4) 단계별 실행 및 피드백 반영 (Rollback/Replanning) 지원
  5) 최종 복원 이미지 출력



# Experiments

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Comparisons with other strategies
  - Compared methods
    - Random Order & Model : 작업 순서와 모델 모두 랜덤 선택
    - Random Order & Predicted Model : Task순서는 랜덤, 모델은 RestoreAgent가 예측
    - Random Model & Predicted Order : 모델은 랜덤, Task 순서는 RestoreAgent가 예측
    - Predefined Order & Model : 모든 이미지에 동일 규칙 적용
    - Human Expert : vision 연구를 5년 이상 한 전문가가 이미지를 직접 보고 맞춤형 restoration pipeline 구성
    - RestoreAgent : MLLM 기반 자동 pipeline결정 및 반복



CVSP Lab
Computer Vision & Signal Processing

# Experiments

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Comparisons with other strategies
  - Results

| | Noise + JPEG | | | | | | Low Light + Noise | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /17 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /10 |
| Random Order & Model | 24.52 | 0.7273 | 0.2889 | 0.2212 | 1.47 | 6.7 | 15.57 | 0.6541 | 0.4351 | 0.2588 | 1.98 | 3.9 |
| Random Oder + Predict Model | 25.24 | 0.7765 | 0.2327 | 0.1960 | 3.07 | 4.2 | 15.62 | 0.6887 | 0.3651 | 0.2283 | 3.03 | 3.0 |
| Random Model + Predict Order | 24.90 | 0.7568 | 0.2597 | 0.2132 | 2.03 | 6.0 | 17.57 | 0.7044 | 0.3685 | 0.2324 | 3.75 | 2.3 |
| Pre-defined Oder and Model | 25.29 | 0.7828 | 0.2366 | 0.2037 | 2.47 | 5.3 | 17.75 | 0.7098 | 0.3385 | 0.2260 | 3.93 | 2.1 |
| Human Expert | 25.06 | 0.7588 | 0.2551 | 0.2121 | 2.25 | 5.5 | 18.05 | 0.7239 | 0.3278 | 0.2220 | 4.29 | 1.9 |
| **RestoreAgent** | 25.32 | 0.7806 | 0.2308 | 0.1958 | 3.17 | 3.9 ↑1.6 | 17.80 | 0.7226 | 0.3259 | 0.2138 | 4.39 | 1.7 ↑0.2 |
| | **Motion Blur + Noise + JPEG** | | | | | | **Rain + Noise + JPEG** | | | | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /64 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /64 |
| Random Order & Model | 24.81 | 0.7816 | 0.2381 | 0.1747 | 2.32 | 19.5 | 25.64 | 0.7970 | 0.2412 | 0.2020 | 2.90 | 16.1 |
| Random Oder + Predict Model | 24.73 | 0.7787 | 0.2261 | 0.1684 | 2.69 | 16.1 | 25.67 | 0.8008 | 0.2368 | 0.1956 | 3.11 | 15.0 |
| Random Model + Predict Order | 24.95 | 0.7912 | 0.2263 | 0.1647 | 3.18 | 13.6 | 26.14 | 0.8074 | 0.2314 | 0.1996 | 3.49 | 13.3 |
| Pre-defined Oder and Model | 24.84 | 0.7895 | 0.2305 | 0.1662 | 2.97 | 15.0 | 25.80 | 0.7981 | 0.2360 | 0.2041 | 2.83 | 16.7 |
| Human Expert | 25.20 | 0.795 | 0.2205 | 0.1646 | 3.82 | 9.0 | 25.99 | 0.8063 | 0.2258 | 0.1992 | 3.58 | 12.6 |
| **RestoreAgent** | 25.16 | 0.7939 | 0.2042 | 0.1546 | 4.35 | 4.6 ↑4.4 | 26.38 | 0.8136 | 0.2200 | 0.1891 | 4.67 | 6.4 ↑6.2 |
| | **Haze + Noise + JPEG** | | | | | | **Haze + Rain + Noise + JPEG** | | | | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /64 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /287 |
| Random Order & Model | 18.98 | 0.7156 | 0.3267 | 0.2212 | 1.52 | 23.4 | 15.13 | 0.6300 | 0.4464 | 0.2800 | 1.28 | 102.5 |
| Random Oder + Predict Model | 19.00 | 0.7235 | 0.3133 | 0.2081 | 2.03 | 20.4 | 17.45 | 0.6897 | 0.3692 | 0.2400 | 2.86 | 72.8 |
| Random Model + Predict Order | 19.67 | 0.7653 | 0.2778 | 0.2010 | 2.95 | 15.9 | 19.79 | 0.7833 | 0.2815 | 0.1991 | 5.66 | 16.9 |
| Pre-defined Oder and Model | 19.47 | 0.7803 | 0.2641 | 0.1912 | 3.51 | 12.4 | 19.29 | 0.7785 | 0.2815 | 0.1974 | 5.502 | 26.1 |
| Human Expert | 19.50 | 0.7753 | 0.2703 | 0.1982 | 3.36 | 12.7 | 19.39 | 0.7802 | 0.2928 | 0.2043 | 5.503 | 21.3 |
| **RestoreAgent** | 19.55 | 0.7794 | 0.25663 | 0.1863 | 3.93 | 8.4 ↑4.3 | 19.72 | 0.7816 | 0.2741 | 0.1903 | 5.86 | 9.7 ↑11.6 |
| | **Motion Blur + Rain + Noise + JPEG** | | | | | | **Average Result Across All Datasets** | | | | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /287 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking (%) |
| Random Order & Model | 21.96 | 0.6672 | 0.3366 | 0.2239 | 2.57 | 85.3 | 21.31 | 0.7139 | 0.3246 | 0.2241 | 1.92 | 34.7 |
| Random Oder + Predict Model | 22.11 | 0.6667 | 0.3038 | 0.2122 | 3.66 | 58.8 | 21.74 | 0.7385 | 0.2848 | 0.2045 | 2.89 | 26.1 |
| Random Model + Predict Order | 22.74 | 0.6996 | 0.2794 | 0.1979 | 5.39 | 24.7 | 22.42 | 0.7574 | 0.2750 | 0.2027 | 3.44 | 22.7 |
| Pre-defined Oder and Model | 22.35 | 0.6862 | 0.2858 | 0.1997 | 4.65 | 35.7 | 22.38 | 0.7639 | 0.2644 | 0.1986 | 3.48 | 22.1 |
| Human Expert | 22.96 | 0.7092 | 0.2861 | 0.2031 | 5.42 | 21.2 | 22.51 | 0.7634 | 0.2670 | 0.2014 | 3.73 | 19.5 |
| **RestoreAgent** | 22.95 | 0.7097 | 0.2615 | 0.1887 | 6.35 | 5.7 ↑15.5 | 22.61 | 0.7700 | 0.2513 | 0.1890 | 4.38 | 12.9 ↑6.6 |



CVSP Lab
Computer Vision & Signal Processing

# Experiments

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Comparisons with other strategies
  - Results
    - Balanced Score
      - 각 metric을 그대로 합산하면 스케일이 다르니 z-score 표준화(평균, 표준편차 기준으로 정규화)해서 합산
      - 현재 RestoreAgent에서 쓴 학습 목적 함수

| | Noise + JPEG | | | | | | Low Light + Noise | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /17 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /10 |
| Random Order & Model | 24.52 | 0.7273 | 0.2889 | 0.2212 | 1.47 | 6.7 | 15.57 | 0.6541 | 0.4351 | 0.2588 | 1.98 | 3.9 |
| Random Oder + Predict Model | 25.24 | 0.7765 | 0.2327 | 0.1960 | 3.07 | 4.2 | 15.62 | 0.6887 | 0.3651 | 0.2283 | 3.03 | 3.0 |
| Random Model + Predict Order | 24.90 | 0.7568 | 0.2597 | 0.2132 | 2.03 | 6.0 | 17.57 | 0.7044 | 0.3685 | 0.2324 | 3.75 | 2.3 |
| Pre-defined Oder and Model | 25.29 | 0.7828 | 0.2366 | 0.2037 | 2.47 | 5.3 | 17.75 | 0.7098 | 0.3385 | 0.2260 | 3.93 | 2.1 |
| Human Expert | 25.06 | 0.7588 | 0.2551 | 0.2121 | 2.25 | 5.5 | 18.05 | 0.7239 | 0.3278 | 0.2220 | 4.29 | 1.9 |
| **RestoreAgent** | 25.32 | 0.7806 | 0.2308 | 0.1958 | 3.17 | 3.9 ↑1.6 | 17.80 | 0.7226 | 0.3259 | 0.2138 | 4.39 | 1.7 ↑0.2 |
| | **Motion Blur + Noise + JPEG** | | | | | | **Rain + Noise + JPEG** | | | | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /64 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /64 |
| Random Order & Model | 24.81 | 0.7816 | 0.2381 | 0.1747 | 2.32 | 19.5 | 25.64 | 0.7970 | 0.2412 | 0.2020 | 2.90 | 16.1 |
| Random Oder + Predict Model | 24.73 | 0.7787 | 0.2261 | 0.1684 | 2.69 | 16.1 | 25.67 | 0.8008 | 0.2368 | 0.1956 | 3.11 | 15.0 |
| Random Model + Predict Order | 24.95 | 0.7912 | 0.2263 | 0.1647 | 3.18 | 13.6 | 26.14 | 0.8074 | 0.2314 | 0.1996 | 3.49 | 13.3 |
| Pre-defined Oder and Model | 24.84 | 0.7895 | 0.2305 | 0.1662 | 2.97 | 15.0 | 25.80 | 0.7981 | 0.2360 | 0.2041 | 2.83 | 16.7 |
| Human Expert | 25.20 | 0.795 | 0.2205 | 0.1646 | 3.82 | 9.0 | 25.99 | 0.8063 | 0.2258 | 0.1992 | 3.58 | 12.6 |
| **RestoreAgent** | 25.16 | 0.7939 | 0.2042 | 0.1546 | 4.35 | 4.6 ↑4.4 | 26.38 | 0.8136 | 0.2200 | 0.1891 | 4.67 | 6.4 ↑6.2 |
| | **Haze + Noise + JPEG** | | | | | | **Haze + Rain + Noise + JPEG** | | | | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /64 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /287 |
| Random Order & Model | 18.98 | 0.7156 | 0.3267 | 0.2212 | 1.52 | 23.4 | 15.13 | 0.6300 | 0.4464 | 0.2800 | 1.28 | 102.5 |
| Random Oder + Predict Model | 19.00 | 0.7235 | 0.3133 | 0.2081 | 2.03 | 20.4 | 17.45 | 0.6897 | 0.3692 | 0.2400 | 2.86 | 72.8 |
| Random Model + Predict Order | 19.67 | 0.7653 | 0.2778 | 0.2010 | 2.95 | 15.9 | 19.79 | 0.7833 | 0.2815 | 0.1991 | 5.66 | 16.9 |
| Pre-defined Oder and Model | 19.47 | 0.7803 | 0.2641 | 0.1912 | 3.51 | 12.4 | 19.29 | 0.7785 | 0.2815 | 0.1974 | 5.502 | 26.1 |
| Human Expert | 19.50 | 0.7753 | 0.2703 | 0.1982 | 3.36 | 12.7 | 19.39 | 0.7802 | 0.2928 | 0.2043 | 5.503 | 21.3 |
| **RestoreAgent** | 19.55 | 0.7794 | 0.25663 | 0.1863 | 3.93 | 8.4 ↑4.3 | 19.72 | 0.7816 | 0.2741 | 0.1903 | 5.86 | 9.7 ↑11.6 |
| | **Motion Blur + Rain + Noise + JPEG** | | | | | | **Average Result Across All Datasets** | | | | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /287 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking (%) |
| Random Order & Model | 21.96 | 0.6672 | 0.3366 | 0.2239 | 2.57 | 85.3 | 21.31 | 0.7139 | 0.3246 | 0.2241 | 1.92 | 34.7 |
| Random Oder + Predict Model | 22.11 | 0.6667 | 0.3038 | 0.2122 | 3.66 | 58.8 | 21.74 | 0.7385 | 0.2848 | 0.2045 | 2.89 | 26.1 |
| Random Model + Predict Order | 22.74 | 0.6996 | 0.2794 | 0.1979 | 5.39 | 24.7 | 22.42 | 0.7574 | 0.2750 | 0.2027 | 3.44 | 22.7 |
| Pre-defined Oder and Model | 22.35 | 0.6862 | 0.2858 | 0.1997 | 4.65 | 35.7 | 22.38 | 0.7639 | 0.2644 | 0.1986 | 3.48 | 22.1 |
| Human Expert | 22.96 | 0.7092 | 0.2861 | 0.2031 | 5.42 | 21.2 | 22.51 | 0.7634 | 0.2670 | 0.2014 | 3.73 | 19.5 |
| **RestoreAgent** | 22.95 | 0.7097 | 0.2615 | 0.1887 | 6.35 | 5.7 ↑15.5 | 22.61 | 0.7700 | 0.2513 | 0.1890 | 4.38 | 12.9 ↑6.6 |



CVSP Lab
Computer Vision & Signal Processing

# Experiments

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Comparisons with other strategies
  - Results
    - Ranking
      - 가능한 Permuatation 중 metric 상 평균 등수
      - Denoise(model 1, model 2, model 3), JPEG(model 1, model 2) … 로 가능한 모델 선택, 순서의 경우의 수에서 평균 등수

| | Noise + JPEG | | | | | | Low Light + Noise | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /17 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /10 |
| Random Order & Model | 24.52 | 0.7273 | 0.2889 | 0.2212 | 1.47 | 6.7 | 15.57 | 0.6541 | 0.4351 | 0.2588 | 1.98 | 3.9 |
| Random Oder + Predict Model | 25.24 | 0.7765 | 0.2327 | 0.1960 | 3.07 | 4.2 | 15.62 | 0.6887 | 0.3651 | 0.2283 | 3.03 | 3.0 |
| Random Model + Predict Order | 24.90 | 0.7568 | 0.2597 | 0.2132 | 2.03 | 6.0 | 17.57 | 0.7044 | 0.3685 | 0.2324 | 3.75 | 2.3 |
| Pre-defined Oder and Model | 25.29 | 0.7828 | 0.2366 | 0.2037 | 2.47 | 5.3 | 17.75 | 0.7098 | 0.3385 | 0.2260 | 3.93 | 2.1 |
| Human Expert | 25.06 | 0.7588 | 0.2551 | 0.2121 | 2.25 | 5.5 | 18.05 | 0.7239 | 0.3278 | 0.2220 | 4.29 | 1.9 |
| **RestoreAgent** | 25.32 | 0.7806 | 0.2308 | 0.1958 | 3.17 | 3.9 ↑1.6 | 17.80 | 0.7226 | 0.3259 | 0.2138 | 4.39 | 1.7 ↑0.2 |
| | **Motion Blur + Noise + JPEG** | | | | | | **Rain + Noise + JPEG** | | | | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /64 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /64 |
| Random Order & Model | 24.81 | 0.7816 | 0.2381 | 0.1747 | 2.32 | 19.5 | 25.64 | 0.7970 | 0.2412 | 0.2020 | 2.90 | 16.1 |
| Random Oder + Predict Model | 24.73 | 0.7787 | 0.2261 | 0.1684 | 2.69 | 16.1 | 25.67 | 0.8008 | 0.2368 | 0.1956 | 3.11 | 15.0 |
| Random Model + Predict Order | 24.95 | 0.7912 | 0.2263 | 0.1647 | 3.18 | 13.6 | 26.14 | 0.8074 | 0.2314 | 0.1996 | 3.49 | 13.3 |
| Pre-defined Oder and Model | 24.84 | 0.7895 | 0.2305 | 0.1662 | 2.97 | 15.0 | 25.80 | 0.7981 | 0.2360 | 0.2041 | 2.83 | 16.7 |
| Human Expert | 25.20 | 0.795 | 0.2205 | 0.1646 | 3.82 | 9.0 | 25.99 | 0.8063 | 0.2258 | 0.1992 | 3.58 | 12.6 |
| **RestoreAgent** | 25.16 | 0.7939 | 0.2042 | 0.1546 | 4.35 | 4.6 ↑4.4 | 26.38 | 0.8136 | 0.2200 | 0.1891 | 4.67 | 6.4 ↑6.2 |
| | **Haze + Noise + JPEG** | | | | | | **Haze + Rain + Noise + JPEG** | | | | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /64 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /287 |
| Random Order & Model | 18.98 | 0.7156 | 0.3267 | 0.2212 | 1.52 | 23.4 | 15.13 | 0.6300 | 0.4464 | 0.2800 | 1.28 | 102.5 |
| Random Oder + Predict Model | 19.00 | 0.7235 | 0.3133 | 0.2081 | 2.03 | 20.4 | 17.45 | 0.6897 | 0.3692 | 0.2400 | 2.86 | 72.8 |
| Random Model + Predict Order | 19.67 | 0.7653 | 0.2778 | 0.2010 | 2.95 | 15.9 | 19.79 | 0.7833 | 0.2815 | 0.1991 | 5.66 | 16.9 |
| Pre-defined Oder and Model | 19.47 | 0.7803 | 0.2641 | 0.1912 | 3.51 | 12.4 | 19.29 | 0.7785 | 0.2815 | 0.1974 | 5.502 | 26.1 |
| Human Expert | 19.50 | 0.7753 | 0.2703 | 0.1982 | 3.36 | 12.7 | 19.39 | 0.7802 | 0.2928 | 0.2043 | 5.503 | 21.3 |
| **RestoreAgent** | 19.55 | 0.7794 | 0.25663 | 0.1863 | 3.93 | 8.4 ↑4.3 | 19.72 | 0.7816 | 0.2741 | 0.1903 | 5.86 | 9.7 ↑11.6 |
| | **Motion Blur + Rain + Noise + JPEG** | | | | | | **Average Result Across All Datasets** | | | | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking /287 | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ | balanced ↑ | ranking (%) |
| Random Order & Model | 21.96 | 0.6672 | 0.3366 | 0.2239 | 2.57 | 85.3 | 21.31 | 0.7139 | 0.3246 | 0.2241 | 1.92 | 34.7 |
| Random Oder + Predict Model | 22.11 | 0.6667 | 0.3038 | 0.2122 | 3.66 | 58.8 | 21.74 | 0.7385 | 0.2848 | 0.2045 | 2.89 | 26.1 |
| Random Model + Predict Order | 22.74 | 0.6996 | 0.2794 | 0.1979 | 5.39 | 24.7 | 22.42 | 0.7574 | 0.2750 | 0.2027 | 3.44 | 22.7 |
| Pre-defined Oder and Model | 22.35 | 0.6862 | 0.2858 | 0.1997 | 4.65 | 35.7 | 22.38 | 0.7639 | 0.2644 | 0.1986 | 3.48 | 22.1 |
| Human Expert | 22.96 | 0.7092 | 0.2861 | 0.2031 | 5.42 | 21.2 | 22.51 | 0.7634 | 0.2670 | 0.2014 | 3.73 | 19.5 |
| **RestoreAgent** | 22.95 | 0.7097 | 0.2615 | 0.1887 | 6.35 | 5.7 ↑15.5 | 22.61 | 0.7700 | 0.2513 | 0.1890 | 4.38 | 12.9 ↑6.6 |



CVSP Lab
Computer Vision & Signal Processing

# Experiments

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Comparisons with other strategies
  - Results
    - Random Order & Model
      - 성능 최하위, 모든 조합 중 하위 34.7% 차지
    - Predefined Order & Model
      - 상위 22.1%. 실제 응용에서 쓰이는 고정 규칙이 랜덤보다는 훨씬 나음
    - Human Expert : vision 연구를 5년 이상 한 전문가가 이미지를 직접 보고 맞춤형 restoration pipeline 구성
      - 상위 19.5%로 경험적 직관으로 어느정도 최적화 가능
      - 하지만 RestoreAgent보다 뒤처짐(수많은 Permuation 중 최적을 맞히기는 어려움)
    - RestoreAgent : MLLM 기반 자동 pipeline결정 및 반복
      - 상위 12.9%, 전문가보다 6.6%p 개선
      - 복잡한 multi-degradation에서 큰 차이를 보임 (↑ 11.6)
        - ✓데이터 기반 의사결정이 전문가의 판단보다 안정적이고 효율적임



CVSP Lab
Computer Vision & Signal Processing

# Experiments

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Comparisons with other strategies
  - Qualitative Results






CVSP Lab
Computer Vision & Signal Processing

# Conclusion

- Agentic System을 통해 순서와 모델 최적화
  - 스스로 문제를 파악하고 계획하고 결과를 분석하여 rollback하거나 마치는 system으로 인간이 문제해결을 하는 과정과 유사
    - 단일 task에 최적화된 모델 또는 여러 모델이 있는 경우
    - Task 수행 순서에 따라 최적화가 가능한 task의 경우
  - Open-Vocabulary Object Detection, Open-world Object Detection 등 많은 단일 모델과 적용 순서가 영향을 미치는 분야라면 적용 가능



CVSP Lab
Computer Vision & Signal Processing

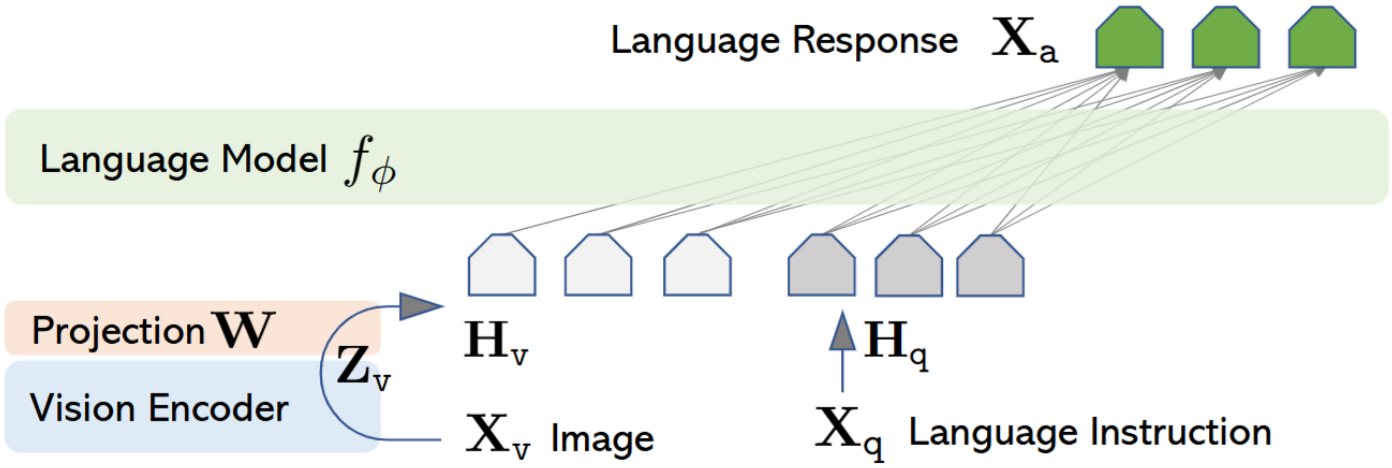
# Appendix



CVSP Lab
Computer Vision & Signal Processing

# Appendix

- Visual Instruction Tuning(LLaVA)
  - Overview Architecture
    - Input: image, instruction
    - Output: response
  - LLaVA Training

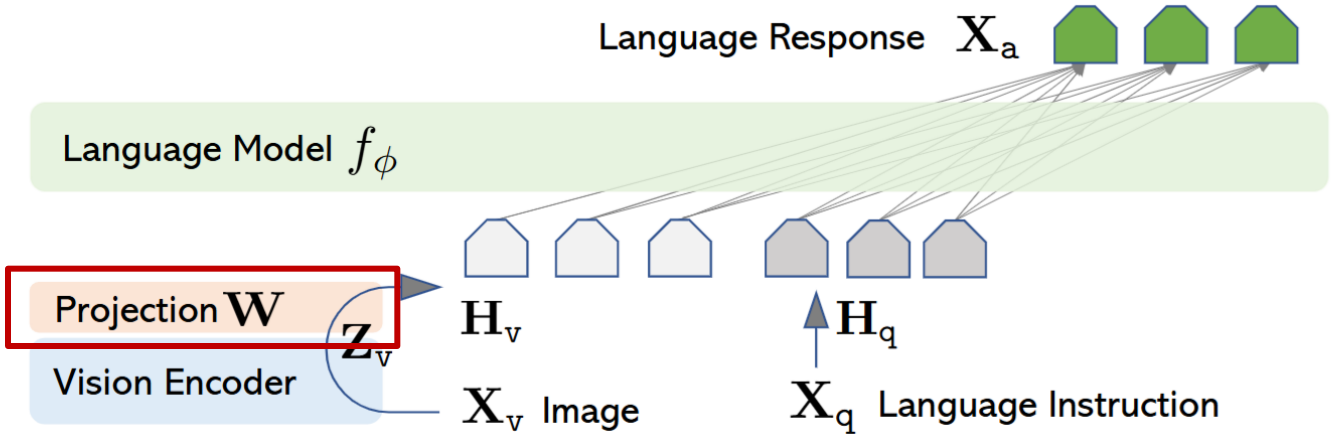



<LLaVA Architecture overview>

# Appendix

1) Peng, Baolin, et al. "Instruction tuning with gpt-4." arXiv preprint arXiv:2304.03277 (2023).

- Visual Instruction Tuning
  - Language Model(Large Language Model)
    - Vicuna[1]를 이용하여 instruction에 따라 응답 생성
      - Vicuna: LLaMA 기반의 모델. 다양한 instruction에 대한 응답 생성 가능
  - Vision Encoder
    - CLIP ViT-L/14 를 사용하여 image를 feature로 변환
  - Projection Layer



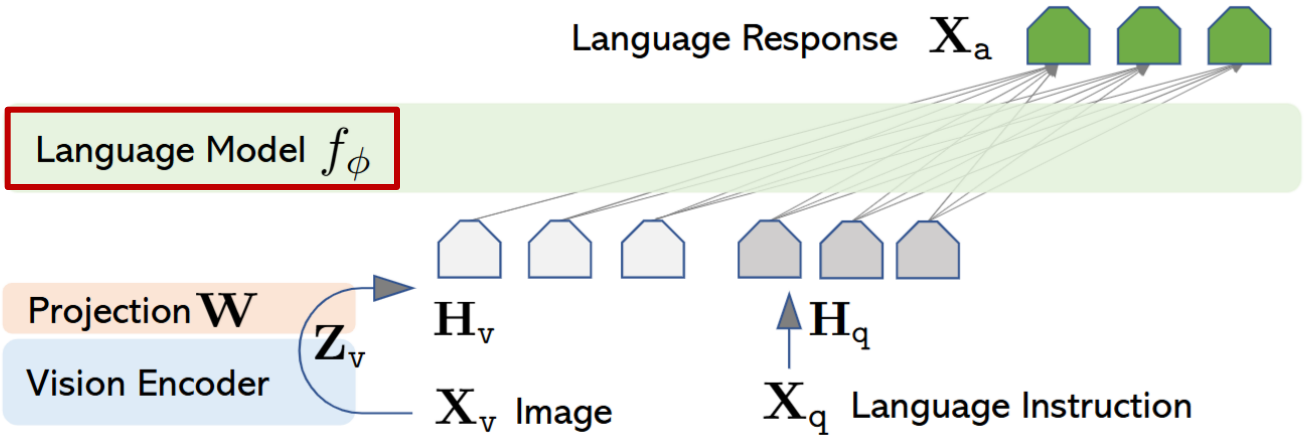



CVSP Lab
Computer Vision & Signal Processing

# Appendix

1) Peng, Baolin, et al. "Instruction tuning with gpt-4." arXiv preprint arXiv:2304.03277 (2023).

- Visual Instruction Tuning (LLaVA)
  - Language Model(Large Language Model)
    - Vicuna[1]를 이용하여 instruction에 따라 응답 생성
      - Vicuna: LLaMA 기반의 모델. 다양한 instruction에 대한 응답 생성 가능
        - ✓LLaMA 모델을 instruction-tuning한 LLM 모델
        - ✓Vicuna $f_\varphi(\cdot)$ 로 표시
          - 기호 $\varphi$ 는 parameter로 LLaVA에서는 7B, 13B짜리를 사용함






CVSP Lab
Computer Vision & Signal Processing

# Appendix

1) Peng, Baolin, et al. "Instruction tuning with gpt-4." arXiv preprint arXiv:2304.03277 (2023).

- Visual Instruction Tuning
  - Vision Encoder

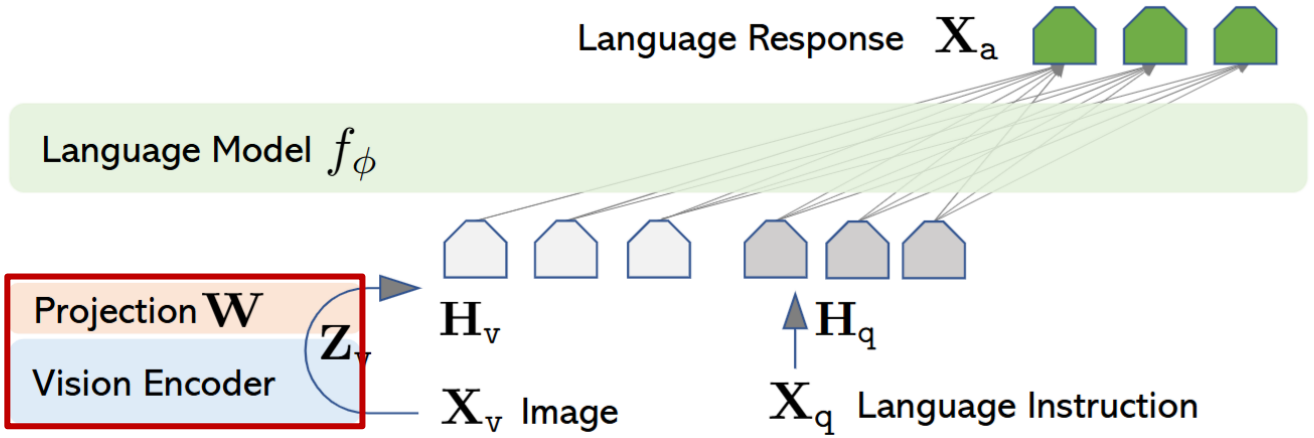
    -CLIP ViT-L/14 를 사용하여 image를 feature로 변환

    $$Z_v = g(X_v)$$

  - Projection layer

    -LLM이 처리할 수 있는 텍스트 임베딩 형태로 변환

    $$H_v = W \cdot Z_v$$

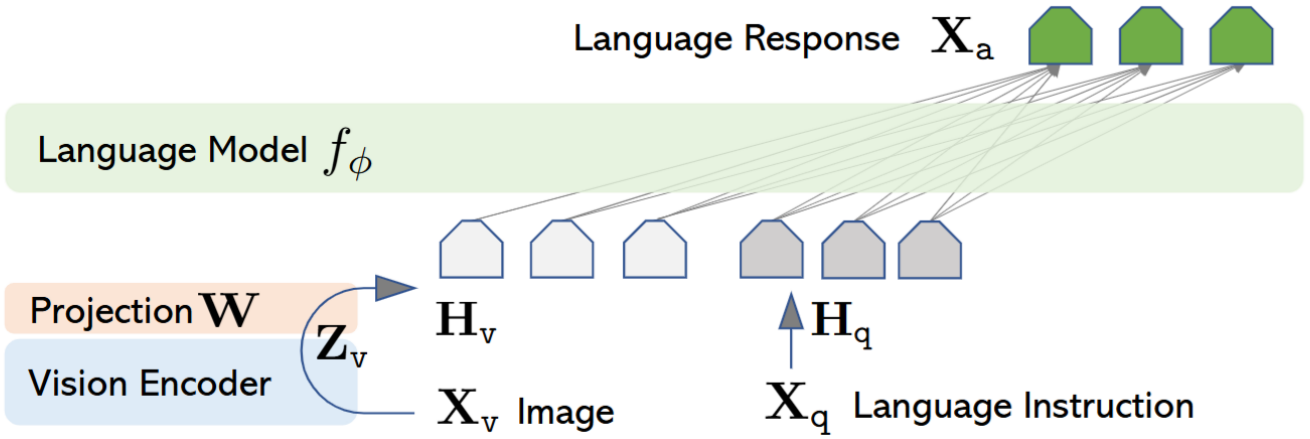





CVSP Lab
Computer Vision & Signal Processing

# Appendix

- LLaVA model Training🔥
  - Training 목적
    - 이미지 input을 받은 상태에서 instruction을 이해하고 답변할 수 있도록 하는 것
    - **LLM(Vicuna)이 instruction 에 맞게 대답을 출력할 수 있도록 tuning**
    - 총 2 stages 로 구성됨






CVSP Lab
Computer Vision & Signal Processing

# Appendix

- LLaVA model Training
  - Training dataset 구성
    - 하나의 이미지 $X_v$에 대해 다수의 instruction-answer pair 생성
      - $(X_q^1,\ X_a^1, \cdots, X_q^T, X_a^T)$
    - multi-turn 대화 시퀀스로 구성
      - $X_{system-message}$ : Vicuna-v0과 똑같이 설정한 system prompt
      - $X_{instruct}$ : 각 t에서의 instruction
      - $X_a$ : 각 t에서의 instruction에 대한 답변

$$\mathbf{X}_{\text{system-message}} \texttt{<STOP>}$$
$$\texttt{Human}: \mathbf{X}_{\text{instruct}}^{1} \texttt{<STOP> Assistant:} \mathbf{X}_{\text{a}}^{1} \texttt{<STOP>}$$
$$\texttt{Human}: \mathbf{X}_{\text{instruct}}^{2} \texttt{<STOP> Assistant:} \mathbf{X}_{\text{a}}^{2} \texttt{<STOP>} \cdots$$



CVSP Lab
Computer Vision & Signal Processing

# Appendix

- LLaVA model Training
  - Training Objective식

-LLaVA가 답변 전체(정답) $X_a$를 생성하는 확률을 어떻게 모델링하는지 보여줌

$$p(X_a|X_v, X_{instruct}) = \prod_{i=1}^{L} p_\theta(x_i|X_v, X_{instruct,<i}, X_{a,<i})$$

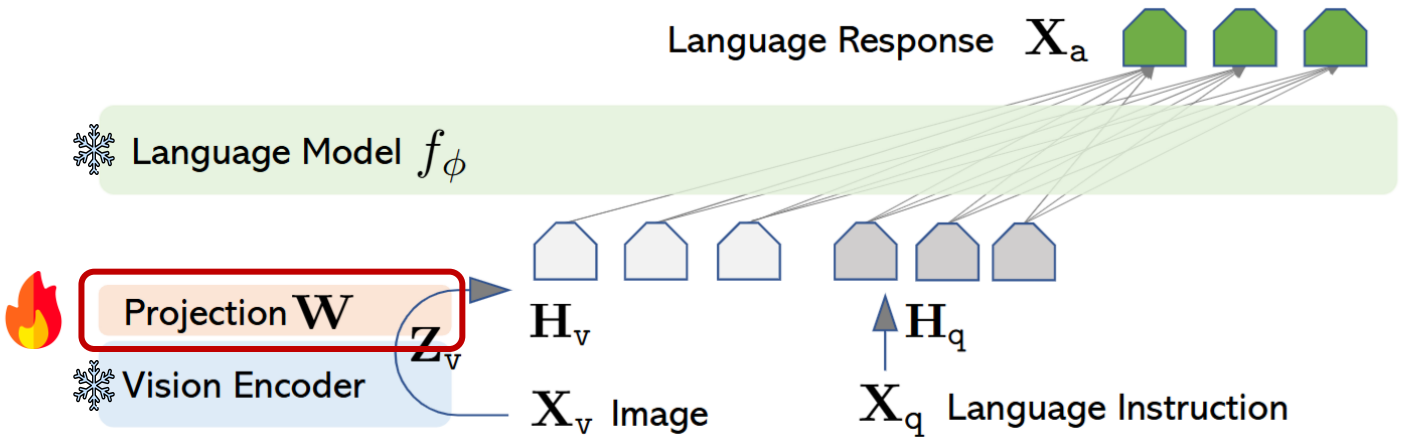
- 목표: X_a 전체의 생성확률을 최대화
- $i$이전 시점까지의 instruction과 답변 토큰들을 조건으로 $x_i$의 확률을 계산
  - ✓ 이미지 $X_v$ 는 모든 시점에서 조건으로 사용됨
  - ✓ $i$시점마다 토큰을 예측하면서 학습됨
  - ✓이미지와 instruction을 조건으로 답변 토큰 sequence를 생성



CVSP Lab
Computer Vision & Signal Processing

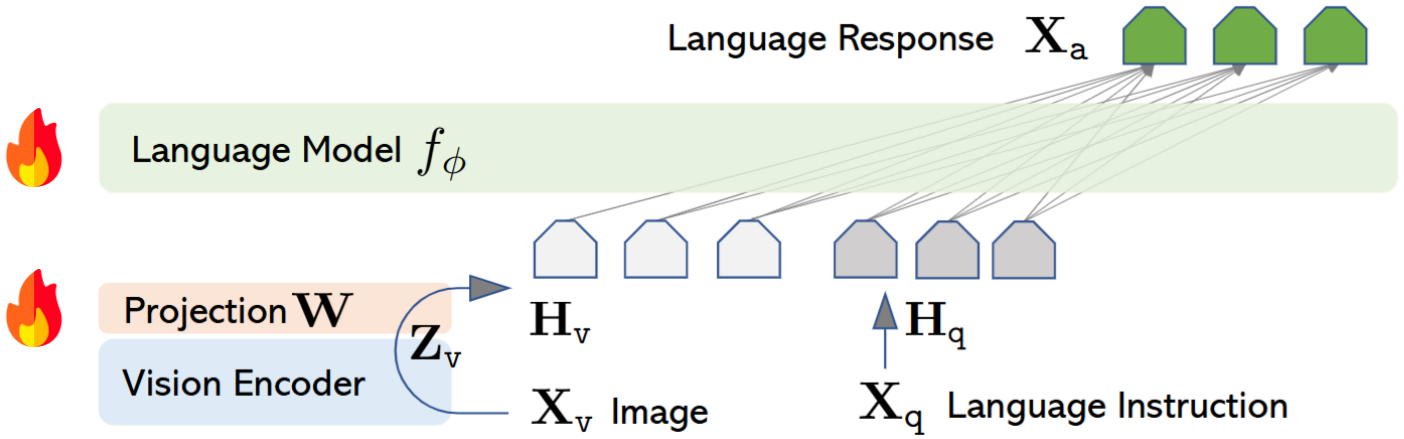
# Appendix

- LLaVA model Training
  - Stage 1: Feature alignment Pre-training
    - -LLM은 고정, 이미지를 LLM의 출력처럼 바꾸는 projection layer를 학습
    - -CLIP Visual Encoder와 Vicuna LLM의 feature space를 맞춰줌
      - 데이터셋: CC3M(595K개의 이미지-텍스트 쌍)
        - ✓이미지 $X_v$에 대해 간단한 설명을 유도하는 질문 $X_q$를 무작위로 샘플링



# Appendix

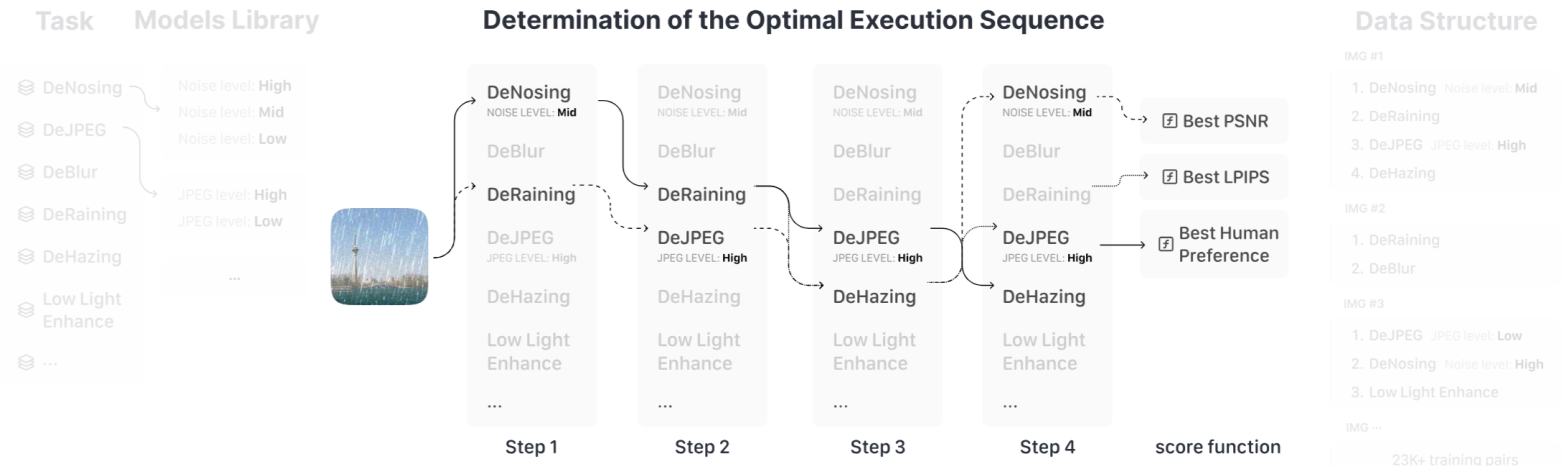
- LLaVA model Training
  - Stage 2: Fine-tuning End-to-End
    - LLaVA-150K 데이터셋으로 Instruction-following 학습
    - Projection Layer와 Language model 모두 업데이트



# Appendix

2)Chen, Haoyu, et al. "Restoreagent: Autonomous image restoration agent via multimodal large language models." Advances in Neural Information Processing Systems 37 (2024): 110643-110666.

- Dataset Construction
  - 모든 경우의 pipeline을 이미지에 적용 및 restoration 결과 품질을 $S(I,\sigma)$로 평가
    - PSNR, LPIPS, Human Preference를 종합하거나 선택하여 평가(Dataset만들기용)
      - Human preference에 대한 정보는 본문에 없음..
    - 모든 경우의 수들을 수행했을 때 점수 S를 비교하여 가장 높은 점수를 얻은 파이프라인을 해당 이미지 $I$의 최적의 restoration pipeline $\sigma$로 선택
      - 평가할 때는 **PSNR, SSIM, LPIPS, DISTS**를 z-score로 표준화해서 사용 (inference후 평가용)




<실행 가능한 모든 조합을 모두 수행하여 품질 평가>



CVSP Lab
Computer Vision & Signal Processing