MoA-VR: A Mixture-of-Agents System Towards All-in-One Video Restoration

Lu Liu*, Chunlei Cai*, Shaocheng Shen, Jianfeng Liang, Weimin Ouyang, Tianxiao Ye, Jian Mao, Huiyu Duan, Jiangchao Yao, Xiaoyun Zhang, Qiang Hu†, *Member, IEEE,* Guangtao Zhai, *Fellow, IEEE*

***Abstract*—Real-world videos often suffer from complex degradations, such as noise, compression artifacts, and low-light distortions, due to diverse acquisition and transmission conditions. Existing restoration methods typically require professional manual selection of specialized models or rely on monolithic architectures that fail to generalize across varying degradations. Inspired by expert experience, we propose MoA-VR, the first Mixture-of-Agents Video Restoration system that mimics the reasoning and processing procedures of human professionals through three coordinated agents: Degradation Identification, Routing and Restoration, and Restoration Quality Assessment. Specifically, we construct a large-scale and high-resolution video degradation recognition benchmark and build a vision-language model (VLM) driven degradation identifier. We further introduce a self-adaptive router powered by large language models (LLMs), which autonomously learns effective restoration strategies by observing tool usage patterns. To assess intermediate and final processed video quality, we construct the Restored Video Quality (Res-VQ) dataset and design a dedicated VLM-based video quality assessment (VQA) model tailored for restoration tasks. Extensive experiments demonstrate that MoA-VR effectively handles diverse and compound degradations, consistently outperforming existing baselines in terms of both objective metrics and perceptual quality. These results highlight the potential of integrating multimodal intelligence and modular reasoning in general-purpose video restoration systems.**

***Index Terms*—Video Restoration, Agentic System, Video Quality Assessment**

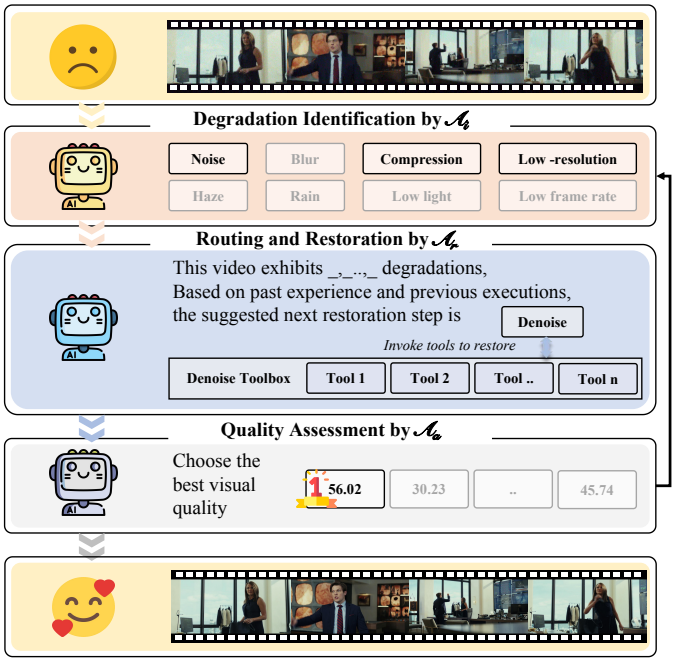


Fig. 1. Overview of the agents in MoA-VR. MoA-VR restores low-quality video clips with complex degradations through the collaboration of three agents: the degradation identification agent, the routing and restoration agent, and the quality assessment agent.

I. Introduction

The growing ubiquity of high-quality video data has fueled advances in diverse applications such as entertainment, surveillance, and autonomous driving, *etc*. However, videos captured or transmitted in real-world environments often suffer from complex and heterogeneous degradations, including noise, compression artifacts, motion blur, and low-light conditions, *etc*. These degradations, which may co-occur or vary spatially and temporally, significantly impair visual quality and downstream processing. Although video restoration (VR) techniques have seen notable progress, achieving robust and high-quality restoration in an all-in-one manner remains a formidable challenge due to the diversity, overlap and unpredictability of degradation types.

Traditional video restoration methods [1]–[8] typically employ a multi-model combination strategy, where each specialized model is trained to handle a specific type of degradation. Representative works such as BasicVSR++ [9] and VRT [10] have demonstrated promising results by effectively targeting particular degradation scenarios. However, it is very difficult for these specialized methods to achieve good performance in practical applications, as they heavily rely on accurately identifying the type of degradation, leveraging expert knowledge to carefully select appropriate models, and performing extensive visual assessment to choose the best result. Moreover, when multiple models need to be combined to handle complex real-world degradations, improper execution order can easily lead to error accumulation, resulting in poor robustness and unsatisfactory performance.

Manuscript received May 31, 2025, revised September 30, 2025, accepted October 7, 2025. This work is supported by National Natural Science Foundation of China (62571322, 62271308, 62401365, 62225112, 62132006, U24A20220), STCSM (24ZR1432000, 24511106902, 24511106900, 22DZ2229005), 111 plan (BP0719010), China Postdoctoral Science Foundation (BX20250411, 2025M773473) and State Key Laboratory of UHD Video and Audio Production and Presentation. (* Equal contribution authors: Lu Liu and Chunlei Cai, † Corresponding author: Qiang Hu)

Lu Liu, Shaocheng Shen, Huiyu Duan, Jianfeng Liang, Jiangchao Yao, Xiaoyun Zhang, Qiang Hu and Guangtao Zhai are with the Shanghai Jiao Tong University, Shanghai, 200240, China. (email:lettieliu, shenshaocheng, huiyuduan, blur_damon, sunarker, xiaoyun.zhang, qiang.hu, zhaiguangtao@sjtu.edu.cn)

Weimin Ouyang is with University of Electronic Science and Technology of China, Hainan, 572400, China. (email:2022360903022@std.uestc.edu.cn)

Chunlei Cai, Tianxiao Ye and Jian Mao are with Bilibili Inc., Shanghai, 200433, China. (email:caichunlei, yetianxiao, maojian@bilibili.com)

In contrast, all-in-one models [9]–[12] aim to streamline the restoration process by using a single unified network to handle diverse degradation types. While recent advances such as AverNet [12] demonstrate promising generalization across tasks, these models often rely on increased capacity and are limited by their static architecture. In practical deployments, especially in large-scale or resource-constrained video applications, model size constraints, domain variability, and unknown degradation compositions make it difficult to achieve consistent performance. As a result, even all-in-one solutions are typically deployed as multiple specialized smaller distilled models. Proper use of them still requires manual tuning and massive evaluation. This reveals a key limitation: static restoration pipelines struggle with real-world complexity, highlighting the necessity for dynamic and adaptive solutions.

Recently, large language models (LLMs), such as DeepSeek [13], and vision-language models (VLMs), such as BLIP [14], GPT-4o [15], and LLaVA [16], have shown remarkable success in multimodal understanding and reasoning tasks. These models inherently possess rich prior knowledge and strong contextual reasoning capabilities, making them well-suited for interpreting complex visual information alongside natural language descriptions. Such abilities suggest a promising direction for video restoration, where diverse and compound degradations need to be accurately identified and adaptively addressed. Since traditional methods struggle with manual degradation classification and fixed restoration pipelines, and all-in-one models lack flexibility and fine-grained control, by leveraging the adaptive reasoning power of VLMs and LLMs, it is possible to build more intelligent and autonomous video restoration systems that dynamically understand and respond to the unique degradation characteristics of each video. However, the application of VLMs and LLMs in video restoration remains largely underexplored, particularly in tasks requiring precise degradation identification and adaptive restoration strategy selection.

To bridge this gap, we propose MoA-VR, a novel intelligent video restoration framework that integrates multimodal perception and modular reasoning within a unified framework, referring the cognitive processes of human experts. MoA-VR decomposes the restoration process into three fundamental capabilities, Identification, Routing, and Assessment, mimicking how human professionals analyze visual degradations, plan restoration workflows, and judge output quality as show in Fig.1. By introducing a modular architecture empowered by VLMs and LLMs, MoA-VR brings flexibility and adaptability to the restoration of complex, mixed, and temporally non-uniform degradations across video sequences. This design enables our system to make task-aware decisions and perform context-sensitive tool selection with minimal human supervision.

To accurately identify degradation types, we establish a vision-language model-based benchmark to systematically evaluate various VLMs on multi-type degradation perception, allowing the model to capture fine-grained and temporally variant artifacts such as compression noise, motion blur, and resolution drop. Based on the recognized degradation profiles, we introduce a self-adaptive routing module powered by LLMs, which learns to synthesize optimal restoration pipelines by interpreting tool execution traces and planning context-aware toolchains. To close the loop, a dedicated video quality assessment (VQA) module is developed and trained on the newly curated Restoration Video Quality (Res-VQ) dataset, which includes high-resolution video pairs and human-annotated perceptual quality scores. This enables MoA-VR to judge the restoration results and guide iterative optimization when needed. Extensive experiments demonstrate that MoA-VR significantly outperforms state-of-the-art video restoration methods, achieving a **3.02 dB** improvement in PSNR and notable gains in both perceptual and pixel metrics, while requiring minimal human intervention.

Our main contributions are summarized as follows:

- We propose MoA-VR, the first multi-agent modular framework for intelligent all-in-one video restoration, integrating identification, routing, and assessment.
- We establish the first VLM-based benchmark for video degradation recognition, enabling accurate analysis of eight common real-world distortions.
- We design a self-adaptive routing module guided by LLMs that learns optimal restoration pipelines from historical tool usage traces.
- We curate the Restored Video Quality (Res-VQ) dataset with human-rated quality scores and build a dedicated VQA model tailored for restoration quality evaluation.

II. Related Works

A. Video Restoration

Video restoration is a fundamental research field focused on reconstructing high-quality video content from degraded observations. Beyond the application of single-frame image methodologies [17]–[19], current VR methods can be categorized into two paradigms based on their application scenarios: task-specific degradation processing models [1]–[5], [20] and all-in-one models [9]–[12]. Task-specific degradation processing models are designed to address individual degradation types through dedicated learning frameworks. These models typically require prior human expert assessment to identify the video degradation characteristics before processing.

However, real-world video degradation typically involves complex mixed artifacts, making single-task restoration methods fundamentally limited. To address this, VRT [10] and BasicVSR++ [9] proposed a unified framework trained simultaneously across multiple tasks, including denoising, deblurring, super-resolution and so on. Moreover, AverNet [12] effectively handle previously unseen degradation patterns by demonstrating a single set of weights. Although all-in-one video restoration models have demonstrated preliminary success, their robustness in handling diverse unknown degradations remains insufficient, often requiring manual post-processing by human experts. This critical limitation motivates the urgent need for an intelligent framework capable of automatically identifying degradation patterns and performing adaptive restoration.

B. LLMs and VLMs

Large Language Models [13], [21], [22] and Vision-Language Models [23], [23], [24] have emerged as pioneering

Fig. 2. MoA-VR incorporates three specialized agents within a closed-loop architecture. For a low-quality input video, $\mathcal{A}_i$ identifies the degradation type and level; $\mathcal{A}_r$ generates a degradation removal plan and then invokes the corresponding restoration toolbox; $\mathcal{A}_a$ assesses all the intermediate results and chooses the best quality one. Then $\mathcal{A}_i$ identifies whether the previous restoration was successful. If it fails, $\mathcal{A}_r$ rolls back and reroutes; if successful, $\mathcal{A}_r$ follows the previous plan. This loop continues until all degradations are removed.

forces in advancing general-purpose artificial intelligence. Trained on massive-scale multimodal datasets encompassing both textual and visual data, these models demonstrate remarkable problem-solving capabilities that extend far beyond conventional language processing tasks. Some studies extend LLMs to specialized domains, such as mathematical problem-solving and legal query processing [25], further blurring the line between human and machine intelligence. Meanwhile, models such as GPT-4o [15] and Qwen-Serials [26] demonstrate unprecedented performance in real-world multimodal interactions. The demonstrated cognitive and perceptual competencies of LLMs and VLMs have established them as foundational architectures for developing sophisticated multi-agent decision systems.

C. Multi-Agent System

Multi-agent systems (MAS) consist of multiple interacting intelligent agents that collaboratively or competitively solve complex problems [27]–[29]. Early MAS research primarily relied on symbolic reasoning [30] and game-theoretic approaches [31], which provided strong theoretical guarantees but struggled with scalability in open-world environments. Recent advances in foundation models have fundamentally transformed this paradigm. Specifically, the emergence of LLMs and VLMs has introduced new paradigms for multi-agent coordination. LLMs enable agents to engage in natural language communication, facilitating role allocation, negotiation, and strategy formulation. For instance, Generative agents [32] demonstrated that LLM-driven agents can simulate human-like social behaviors in virtual environments, while HuggingGPT [33] orchestrates multiple AI models through LLM-based task decomposition.

Moreover, VLMs further extend multi-agent capabilities into visually grounded environments. In virtual environments, Ghost in the Minecraft [34] demonstrated hierarchical planning for long-horizon tasks through visual-language grounding. For distributed perception, AVLEN [35] developed attention mechanisms for VLM-equipped drone swarms, enabling emergent coordination during exploration. In the domain of image restoration research, AgenticIR [36] and Restoreagent [37] independently developed novel multi-agent frameworks incorporating VLMs, marking a significant break-

TABLE I
COMPARISON OF POPULAR VIDEO RESTORATION DATASETS. THIS TABLE INCLUDES DATASETS FOR VARIOUS TASKS SUCH AS SUPER-RESOLUTION, DEBLURRING, AND COMPRESSION ARTIFACT REMOVAL.

Dataset	Task Type	Resolution	Clips	Total Frames	Frames per Clip
GoPro [38]	Deblur	1280×720	33	3214	97
DVD [39]	Deblur	1280×720	71	5708	95
BSD [40]	Deblur	640×480	300	30000	100
REDS [41]	SR, Deblur	1280×720	300	30000	100
Vimeo-90K [42]	SR, Denoising, Decomp.	448×256	91,701	641,970	7
MoA-VD	8 tasks: SR, Denoising, Decompression, Deblur, Low-light Enhancement, Dehazing, Deraining, Interpolation	1920×1080	3,300	330,000	100

through by extending multi-agent system applications to image inpainting tasks.

D. Visual Quality Assessment

For Visual Quality Assessment, conventional approaches [43]–[46] typically evaluate video quality through handcrafted feature extraction and regression analysis. With the advancement of deep learning techniques, numerous deep neural network (DNN)-based VQA models have been proposed and demonstrated superior performance. Several representative approaches [47]–[53] employ pre-trained deep neural networks to extract semantic features and predict quality scores through training regression evaluator. However, these methods exhibit limited capability in restoration-type understanding and can only provide holistic quality assessments for videos.

III. METHODOLOGY

A. Problem Formulation and System Architecture

We consider a comprehensive degradation space $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, where each d_i denotes a specific degradation type, such as noise, blur, compression artifacts, rain, haze, low resolution, low frame rate, or low-light conditions. For each degradation d_i, we maintain a dedicated toolset $\mathcal{T}_{d_i} = \{T^1_{d_i}, T^2_{d_i}, \ldots\}$, where each tool $T^j_{d_i}$ is tailored to address d_i under varying contexts and severity levels. Given a degraded video V potentially affected by an unknown combination of degradations from $\mathcal{D}$, our objective is to design an agentic system that can iteratively identify active degradations, determine an effective removal sequence, select appropriate tools, and refine its strategy based on restoration quality feedback. Formally, the restoration process is defined as:

$$\tilde{V} = \mathcal{F}(V),$$

where $\tilde{V}$ denotes the restored video output.

To this end, we propose MoA-VR, a modular multi-agent video restoration system inspired by the collaborative workflow of human video repair experts. As illustrated in Fig. 2, MoA-VR comprises three VLM-empowered agents, each specializing in a critical aspect of video restoration: the *Degradation Identification Agent* $\mathcal{A}_i$, the *Routing and Restoration Agent* $\mathcal{A}_r$, and the *Quality Assessment Agent* $\mathcal{A}_a$.

The Degradation Identification Agent $\mathcal{A}_i$ performs comprehensive analysis of the input video, detecting degradation types and their severity levels (e.g., low, medium, high). It generates a structured diagnostic output that serves as a precise foundation for restoration planning. This agent is implemented via a fine-tuned vision-language model, enabling robust and context-aware degradation diagnosis.

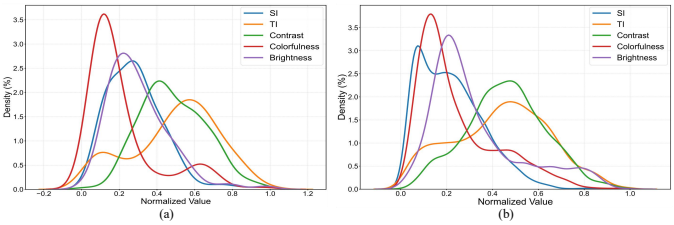


Fig. 3. Feature distribution of (a) MoA-VD-GT and (b) MoA-VD-LQ. SI and TI indicate spatial and temporal information, respectively.

Based on the diagnostic output, the Routing and Restoration Agent $\mathcal{A}_r$ formulates an explicit restoration sequence, decomposing the task into subtasks targeting specific degradations. It then selects and applies restoration tools from the respective toolsets $\mathcal{T}_{d_i}$. Importantly, $\mathcal{A}_r$ iteratively refines the restoration plan by incorporating feedback from intermediate restoration results, enabling a flexible, adaptive, and content-aware enhancement process.

To ensure high restoration quality, the Quality Assessment Agent $\mathcal{A}_a$ acts as an automated evaluator, estimating the visual quality of intermediate outputs. By assigning quantitative quality scores, it assists $\mathcal{A}_r$ in selecting the most effective tools for each subtask. This agent emulates the human-in-the-loop quality inspection process, thereby enhancing robustness and decision reliability.

Together, these three agents form a tightly coupled closed-loop system that iteratively processes the input video: $\mathcal{A}_i$ diagnoses degradations, $\mathcal{A}_r$ executes and adapts restoration operations, and $\mathcal{A}_a$ provides continuous quality feedback. This collaborative architecture offers strong flexibility, scalability, and generalization capacity, effectively supporting robust restoration across diverse and complex degradation scenarios.

B. Degradation Identification Agent $\mathcal{A}_i$

To effectively support downstream restoration agents, the Degradation Identification Agent $\mathcal{A}_i$ is introduced to recognize both the type and severity of degradations present in input videos. This subsection elaborates on the construction of a comprehensive training dataset and the multi-modal model design of $\mathcal{A}_i$, highlighting its capabilities in diverse and compound degradation recognition.

Fig. 4. Visual Examples of Different Video Degradations

Dataset Construction. Training a robust degradation identifier demands a large-scale, high-fidelity dataset with fine-grained annotations. Existing video restoration datasets, summarized in Table I, often suffer from limited diversity in degradation types and insufficient resolution. To overcome these limitations, we construct the MoA-VD dataset, featuring 300 high-quality 1080P-resolution videos paired with 3000 degraded versions. The source content consists of 300 diverse 4K videos, including humans, objects, scenes, and animations, collected from online platforms and downsampled to 1080P using bicubic interpolation to preserve quality while avoiding upsampling artifacts. The feature diversity of MoA-VD is shown in Fig. 3, covering a wide range of color distributions, contrast variations, and temporal dynamics.

To generate degraded counterparts, we design a comprehensive degradation pipeline inspired by Real-ESRGAN [54] and AgenticIR [36], simulating eight common real-world degradation types: rain, haze, blur, low resolution, low frame rate, low light, noise, and compression artifacts. Each degraded video is labeled with its degradation type(s) and corresponding severity levels (low, medium, high). Notably, both single and mixed degradations are included, enhancing model generalizability.

Specifically, haze is generated using an atmospheric scattering model with depth-based transmission estimation; rain is simulated via directional filtering and Gaussian noise perturbations; blur includes both defocus (circular kernel) and gaussian blur; noise covers Gaussian and Poisson variants with adjustable intensity; low-light degradation is achieved by modifying the luminance channel in HSV color space; low resolution is obtained through bicubic downsampling; low frame rate via frame dropping; and compression artifacts are introduced by H.264 encoding. For mixed degradations, up to three types are randomly selected and applied sequentially in a realistic order. The resulting dataset consists of 300 ground-truth videos and 3000 degraded clips, totaling 300,000 frames, as illustrated in Fig. 4.

Degradation Identification. The Degradation Identification Agent $\mathcal{A}_i$ is designed as a VLM capable of multi-modal reasoning to identify degradation types and assess their severity in video content. Inspired by recent advances in instruction-tuned multimodal learning [55]–[57], we reformulate the degradation assessment task as a *text-based classification problem*, enabling the model to better exploit the semantic priors embedded in large language models and improve interpretability.

The degradation identification agent is capable of identifying eight degradation types: noise, blur, compression, low resolution, rain, haze, low frame rate, and low light. These types were selected based on their frequency in real-world scenarios, where degradations often appear as combinations of these factors. In comparison to other all-in-one restoration methods that typically address only a limited set of degradations, our approach considers a broader range, as supported by prior works [17], [58] and real-world applications. These degradations are among the most common and urgent to address in all-in-one restoration tasks. Some other less typical degradations may be considered for future work.

As depicted in Fig. 5, the input video is first processed by the vision encoder Qwen-VL-2.5-ViT [26], which has been fine-tuned using Low-Rank Adaptation (LoRA) to capture rich spatial-temporal representations efficiently. The extracted visual features are then projected into the language embedding space via two stacked multilayer perceptrons (MLPs). These projected visual embeddings are concatenated with tokenized textual prompts to form joint multi-modal tokens. Subsequently, these fused representations are passed into the pre-trained language backbone Qwen-VL2.5-7B [26] for contextual understanding and multi-modal inference.

Rather than regressing scalar values for degradation levels, $\mathcal{A}_i$ generates human-readable textual statements, such as *"The noise level is medium."* This formulation aligns more naturally with the strengths of LLMs, which are more adept at modeling discrete semantic categories than continuous numerical values. Moreover, this approach enhances interpretability and generalizability to unseen degradation types or ambiguous cases.

To adapt the pre-trained VLM to the degradation classification task, we apply LoRA [59] to both the vision encoder and the LLM. Specifically, for a frozen linear transformation $h = Wx$, LoRA introduces a low-rank update:

$$h = Wx + \Delta Wx = Wx + \frac{\alpha}{r} BAx, \tag{1}$$

where $A \in \mathbb{R}^{r \times d_i}$, $B \in \mathbb{R}^{d_o \times r}$, and $r \ll \min(d_o, d_i)$ is the rank of the adaptation. The scalar α is a learnable scaling factor.

We fine-tuned the LLM and vision encoder by minimizing a token-level cross-entropy loss. Formally, the language loss is defined as:

$$\mathcal{L}_{\text{language}} = -\frac{1}{N} \sum_{i=1}^{N} \log P(y_{\text{label}} \mid y_{\text{pred}}), \tag{2}$$

where y_{pred} is the predicted token, y_{label} is the ground truth token, $P(y_{\text{label}} \mid y_{\text{pred}})$ is the predicted probability of the correct token, and N is the total number of tokens. This fine-tuning strategy ensures lightweight yet effective adaptation, allowing $\mathcal{A}_i$ to deliver accurate, fine-grained classification of degradation types and severities. The identified degradation cues provide essential semantic guidance for downstream restoration agents.

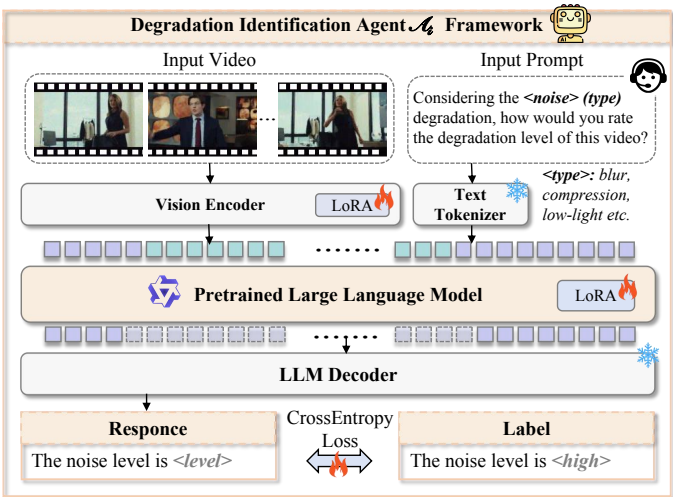


Fig. 5. The overall framework of $\mathcal{A}_i$. $\mathcal{A}_i$ can evaluate all types of degradation levels in an all-in-one framework. It can process videos, along with prompts, to identify the degradations. It consists of a vision encoder to extract both spatial and temporal features and a text tokenizer to tokenize the input prompts. These features are projected into the same space by trained projectors. A pre-trained LLM is utilized to fuse the features while fine-tuned with LoRA.

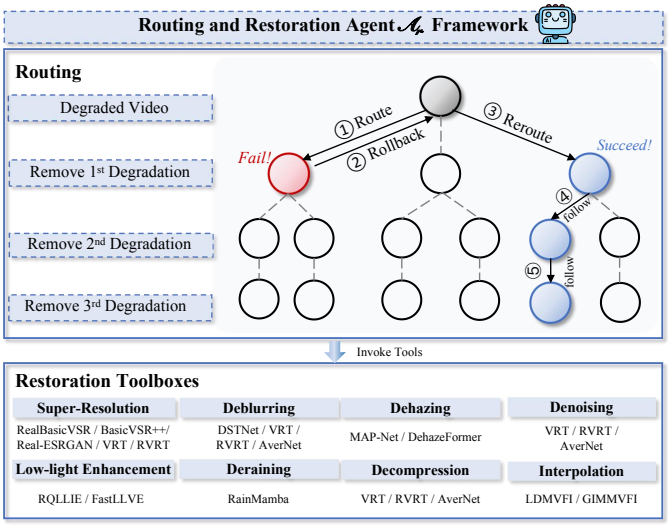


Fig. 6. Illumination of degradation removal process by Routing and Restoration Agent $\mathcal{A}_r$. $\mathcal{A}_r$ is able to route the degradation removal orders, rollback when restoration fails, reroute to another degradation removal orders.

C. Routing and Restoration Agent $\mathcal{A}_r$

To tackle the diversity and context-dependency inherent in video degradations, we propose the Routing and Restoration Agent $\mathcal{A}_r$ that adaptively plans multi-step degradation removal sequences by leveraging LLM-guided reasoning combined with a self-adapted mechanism. This agent orchestrates the restoration process by integrating high-level decision-making with modular, degradation-specific restoration tools.

Inspired by recent advances in agent-based image restoration [36], [60], we extend adaptive degradation routing into the more complex video domain, where the space of possible degradations and their combinations grows combinatorially. Static or rule-based routing methods quickly become insufficient to address this challenge. To enable scalable and intelligent routing, $\mathcal{A}_r$ employs GPT-4 [21] as a high-level reasoning engine. For each degraded input, $\mathcal{A}_r$ prompts GPT-4 to generate candidate sequences of degradation removal steps. Although GPT-4 demonstrates strong abstract reasoning capabilities [61], it initially lacks specific expertise in video restoration, which can lead to suboptimal routing proposals.

To bridge this gap, we introduce a self-exploration framework that allows $\mathcal{A}_r$ to iteratively improve its routing strategy. The agent experiments with multiple degradation removal sequences for each input, evaluating restoration quality with support from the Degradation Identification Agent $\mathcal{A}_i$. Each trial, whether successful or not, is recorded as experience. Periodically, GPT-4 consolidates these accumulated experiences into a task-specific routing knowledge base, refining its future planning to better handle composite and previously unseen degradation patterns.

Given the inherent complexity of multi-step restoration, failures in routing can still occur. To enhance robustness, $\mathcal{A}_r$ incorporates a rollback and rerouting mechanism. Upon encountering ineffective degradation removal, the agent backtracks to a prior state, excludes the failed route, and generates an alternative plan. Failed sequences are cached to avoid repetition and inform subsequent explorations (see Fig. 6).

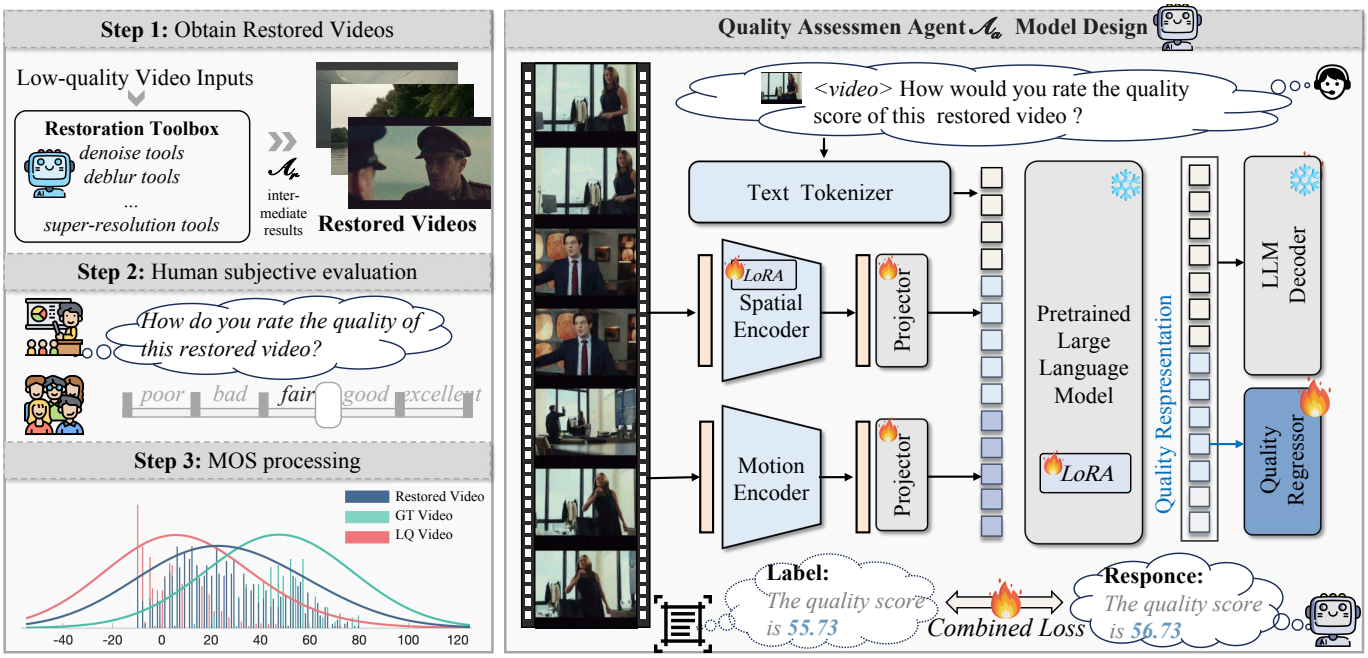
The actual restoration steps are executed by invoking a suite of modular restoration toolboxes, each specialized for a particular degradation type such as blur, noise, or compression artifacts. These toolboxes integrate state-of-the-art open-source models and can be updated flexibly as new methods emerge. Once $\mathcal{A}_r$ selects a degradation class, the corresponding toolbox is applied sequentially to the degraded input. We select 1 to 4 tools for each task. For example, BasicVSR++ [9] for video super-resolution and VRT [10] for deblurring.

By synergizing LLM-guided strategic planning, iterative experience-driven refinement, failure-aware rollback, and modular restoration toolbox, $\mathcal{A}_r$ achieves highly adaptive, scalable, and robust restoration performance across a broad spectrum of video degradations.

D. Quality Assessment Agent $\mathcal{A}_a$

In our agent-based restoration system, reliable video quality assessment (VQA) is crucial for guiding the self-evolution of routing policies and evaluating restoration effectiveness in alignment with human preferences. Unlike traditional User-generated Content (UGC) video quality assessments, our VQA evaluates restored videos generated by our system, which exhibit distinct characteristics. However, directly applying VQA models trained on UGC videos often fails to accurately reflect human preferences, and there is currently no dedicated database for restored videos quality assessment. To address this gap, we develop the Res-VQ dataset, specifically annotated for restored videos and capturing the features introduced during the restoration process. Based on this dataset, we propose a tailored VQA agent that better evaluates restored content and aligns more closely with human preferences.

Restored Video Dataset Construction and Subjective Scoring. To train a quality assessment model capable of



(a) Res-VQ Database Curation (b) Quality assessment agent framework

Fig. 7. (a) Res-VQ quality assessment dataset curation process. (b) An overview of quality assessment agent. It consists of three feature encoders, including an image feature extractor for extracting spatial features from sparse video frames, a motion feature extractor for extracting motion features from the entire video, and a text encoder for extracting aligned text features from prompts. The extracted features are then aligned through projectors and fed into a pre-trained LLM to generate the output results. LoRA weights are introduced to the pre-trained image encoder and the large language model to adapt the models to the quality assessment task.

robustly evaluating diverse restoration outcomes, we first construct Res-VQ, a subjective quality dataset comprising 2,000 restored video outputs generated by our agentic system. Res-VQ contains both the restored clips and their corresponding human mean opinion scores (MOSs), and is used exclusively for training and evaluating the quality assessment agent. Compared with the previously proposed MoA-VD, these two datasets are disjoint and serve orthogonal purposes: MoA-VD focuses on the restoration process and consists of LQ–HQ video counterparts with degradation labels, whereas Res-VQ is dedicated to perceptual quality prediction of restored outputs.

For source video collection, we randomly sample 1,500 intermediate outputs from our system across different stages of degradation removal, meaning that these clips have been restored by the restoration tools one or more times. To ensure quality diversity, we additionally include 250 original low-quality inputs and 250 corresponding high-quality video references in the source video set. These source videos span a wide range of degradation types, restoration methods, and restoration plan sequences, ensuring diversity and realism.

For ground-truth annotation, we conduct a controlled subjective study with 15 human raters having normal or corrected-to-normal vision. Prior to rating, all participants undergo expert-led training to calibrate their perception of restoration quality. During the experiment, videos are displayed in randomized order on 27-inch 4K monitors under standardized lighting conditions. Each clip is rated on a 5-point scale based on perceived restoration effectiveness. The protocol strictly follows ITU-R BT.500-14 [62] guidelines to ensure consistency and validity.

To process subjective scores, we apply kurtosis-based outlier rejection with a 3% threshold to reduce inter-rater variance. The remaining scores are normalized into a [0,100] range using z-score standardization and linear rescaling:

$$z_{ij} = \frac{r_{ij} - \mu_j}{\sigma_i}, \qquad z'_{ij} = \frac{100(z_{ij} + 3)}{6}, \tag{3}$$

$$\mu_j = \frac{1}{M_i} \sum_{i=1}^{M_i} r_{ij}, \ \sigma_i = \sqrt{\frac{1}{M_i - 1} \sum_{j=1}^{M_j} (r_{ij} - \mu_j)^2}, \tag{4}$$

where r_{ij} denotes the score from the i-th participant on video j, and M_i is the number of rated samples for participant i. Final mean opinion scores (MOS) are calculated by averaging the normalized scores:

$$MOS_j = \frac{1}{N} \sum_{i=1}^{N} z'_{ij} \tag{5}$$

where N is the number of valid participants, and z'_{ij} denotes rescaled z-scores.

Vision-Language Quality Regression Model. To model quality in a way that aligns with human judgment while scaling across unseen degradations, we build a VLM-based quality assessor. Compared to classification, regression tasks impose stricter requirements on the fidelity of visual representations, especially for subtle perceptual differences in restored outputs. Our model incorporates both spatial and temporal cues: visual frames are processed by a vision encoder, and motion information is captured using the SlowFast network [63].

TABLE II
COMPARISON OF MOA-VR WITH ALL-IN-ONE METHODS FOR MULTI-DEGRADED VIDEO RESTORATION. WE HIGHLIGHT BEST AND SECOND-BEST VALUES FOR EACH METRIC. ♡ AND ♠, AND DENOTE THE ALL-IN-ONE IMAGE RESTORATION AND ALL-IN-ONE VIDEO RESTORATION METHODS, RESPECTIVELY.

	Double Degradation						Triple Degradation					
	PSNR ↑	SSIM ↑	LPIPS ↓	MANIQA↑	CLIP-IQA↑	MUSIQ↑	PSNR ↑	SSIM ↑	LPIPS ↓	MANIQA↑	CLIP-IQA↑	MUSIQ↑
♡AirNet [58]	15.52	0.4709	0.6137	0.1815	0.2921	29.56	12.70	0.3408	0.7958	0.1420	0.2941	22.82
♡PromptIR [64]	15.84	0.4553	0.6363	0.1903	0.2956	29.85	13.04	0.3189	0.8339	0.1474	0.2939	23.01
♡DA-CLIP [65]	16.79	0.4909	0.6177	0.2142	0.3093	35.27	15.66	0.5254	0.5603	0.2196	0.2632	25.83
♡HAIR [66]	15.86	0.4593	0.6308	0.1875	0.2811	29.76	13.07	0.3247	0.8247	0.1456	0.2727	23.29
♠BasicVSR++ [9]	20.45	0.6049	0.5146	0.2108	0.2767	25.65	17.09	0.5026	0.5579	0.2474	0.3322	24.65
♠AverNet [12]	19.03	0.6583	0.3949	0.1894	0.2394	26.02	20.62	0.6481	0.3564	0.1556	0.1907	15.40
MoA-VR-Ours	23.47	0.6852	0.3386	0.2238	0.3367	36.14	22.94	0.6497	0.3074	0.2585	0.3156	30.43

These representations are projected into a shared token space and concatenated. However, as prior studies indicate that VLMs often struggle with numerical precision, we introduce a lightweight two-layer multilayer perceptron as a quality regressor to decode final scores. The hidden state from the token immediately preceding the numeric prediction in the LLM sequence is extracted and passed to a quality decoder. We optimize the training using a combination of language loss and L1 loss. The training objective combines language loss and L1 loss. The L1 loss is defined as:

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} |q_{\text{pred}} - q_{\text{label}}|, \tag{6}$$

where q_{pred} and q_{label} are the predicted and ground truth quality scores, and N is the number of videos in a batch. The overall loss function is the sum of the two:

$$\mathcal{L}_{\text{combined}} = \mathcal{L}_{\text{language}} + \mathcal{L}_1. \tag{7}$$

This architecture benefits from the interpretability and generalization of LLMs while addressing their inherent limitations on quantitative prediction tasks.

E. Agent Collaboration and Closed-Loop Design

Building upon the specialized capabilities of the three agents introduced earlier, we now detail their collaborative interaction within a unified closed-loop framework designed for robust and adaptive video restoration. This framework emulates human-like iterative problem-solving by leveraging inter-agent communication, feedback-driven decision making, and dynamic adjustments to progressively remove complex and mixed degradations. An overview of the MoA-VR workflow is illustrated in Fig. 2, and the corresponding agent collaboration process is detailed in Algorithm 1.

Given an input video suffering from unknown and compound degradations, the restoration process begins with the Degradation Identification Agent $\mathcal{A}_i$, which performs a fine-grained analysis to detect the presence and severity of each degradation type. This is achieved by querying all known degradation categories and applying a predefined *low-level* threshold as the detection criterion. The resulting degradation profile serves as the initial condition for subsequent routing decisions.

Algorithm 1: Agent Collaboration in MoA-VR

Input: Low-quality video V
Output: Restored high-quality video $\tilde{V}$

 # Identify the type of degradation in the input video
$degradation \leftarrow \mathcal{A}_i(V)$;
 # Route the degradation removal sequence
$plan \leftarrow \mathcal{A}_r(degradation)$;
while *plan is not empty* **do**
  $subtask \leftarrow \text{First}(plan)$;
  $tools \leftarrow \text{Toolbox}(subtask)$;
  $candidates \leftarrow \emptyset$;
  foreach $tool \in tools$ **do**
    $V' \leftarrow tool(V)$;
    # Evaluate the quality of the restored video to determine a best tool
    $score \leftarrow \mathcal{A}_a(V', subtask)$;
    Add $(V', score)$ to $candidates$;
  $(\tilde{V}, best\_score) \leftarrow \text{SelectBest}(candidates)$;
  $success \leftarrow \mathcal{A}_i(\tilde{V}, subtask)$;
  if *success* **then**
    $V \leftarrow \tilde{V}$;
    Remove $subtask$ from $plan$;
  else
    # Reroute if previous restoration fails ;
    $plan \leftarrow \mathcal{A}_r(plan, subtask)$;
output V

Next, the Routing and Restoration Agent $\mathcal{A}_r$ consults a knowledge base and leverages prior restoration experiences to formulate a customized restoration path. This path consists of an ordered sequence of subtasks drawn from the restoration toolbox. For each current subtask, $\mathcal{A}_r$ activates all relevant tools, each generating a candidate restored output. These candidates are then evaluated by the Quality Assessment Agent $\mathcal{A}_a$, which employs a vision-language quality regression model to assess perceptual quality and select the best-performing candidate. The selected output becomes the final result of the current iteration and the input for the next iteration.

The closed-loop nature of MoA-VR is embodied in its iterative feedback mechanism. In each subsequent iteration, $\mathcal{A}_i$ re-assesses the degradation state. If a particular degradation type is no longer detected (i.e., judged as *none*), it signifies the success of the preceding restoration attempt. The Routing and Restoration Agent $\mathcal{A}_r$ then advances along the planned route to apply the next subtask. Conversely, if degradation

persists, the system interprets this as a failed restoration step. Consequently, $\mathcal{A}_r$ triggers a rollback and dynamically replans an alternative restoration sequence, as illustrated in Fig. 6, selecting a new toolbox for the subtask. The cycle then continues with updated context.

Throughout this iterative process, the three agents synergistically contribute their strengths: $\mathcal{A}_i$ provides precise degradation perception, $\mathcal{A}_r$ enables adaptive decision-making and planning, and $\mathcal{A}_a$ delivers human-aligned quality evaluation. Together, they allow MoA-VR to flexibly and effectively address diverse and complex mixed degradation scenarios, yielding a scalable and generalizable solution for real-world video restoration challenges.

IV. Experiment

A. Configurations

Dataset. We employ two custom-curated datasets tailored to the multi-agent setting: 1) MoA-VD: A multi-degradation video dataset used to train and evaluate the Degradation Identification Agent ($\mathcal{A}_i$) and the Routing and Restoration Agent ($\mathcal{A}_r$). We construct the dataset with diverse combinations of eight typical distortions (e.g., blur, noise, compression). The training set contains a wide range of degradation permutations, while the test set comprises 400 video clips with degradation combinations not seen during training, ensuring generalization rather than memorization. 2) Res-VQ: A quality-labeled video dataset built for evaluating the Quality Assessment Agent ($\mathcal{A}_a$). It includes both degraded and restored video sequences with MOS as ground truth. An 80:20 split is applied for training and evaluation.

Implementation. All experiments are conducted on NVIDIA GeForce RTX 3090 and RTX A40 GPUs. The degradation identification agent ($\mathcal{A}_i$) is built upon Qwen2.5-VL-7B [26], fine-tuned for 5 epochs using the AdamW optimizer with a learning rate of 1×10^{-3}, a batch size of 4, a LoRA rank of 16, and a LoRA alpha of 32. The routing and restoration agent ($\mathcal{A}_r$) integrates a set of task-specific state-of-the-art models: VRT [10], AverNet [12], and DSTNet [3] for video denoising, decompression, and deblurring; FastLLVE [4] and RQLLIE [67] for low-light video enhancement; MAP-Net [2] and Dehazeformer [68] for video dehazing; Rainmamba [1] for video deraining; GIMMVFI [5] for low-frame-rate video interpolation; and BasicVSR++ [9] and Real-ESRGAN [54] for video super-resolution. For the quality assessment agent ($\mathcal{A}_a$), we adopt InternLM-8B [69] as the base model, which is trained with AdamW for 10 epochs using a batch size of 8 and a LoRA rank of 16.

B. Comparison

To validate the effectiveness of our proposed MoA-VR framework, a flexible and scalable mixture-of-agent system for all-in-one video restoration, we compare it against six representative state-of-the-art methods, including four image-based all-in-one restoration approaches and two video restoration methods designed for mixed distortion scenarios. Notably, while image-based methods are typically trained on a wider variety of degradation types, video-based approaches such as BasicVSR++ are generally limited to handling only two distortion types simultaneously. This limitation reveals a crucial gap in existing video restoration techniques when faced with complex, multi-type distortions, which MoA-VR aims to address comprehensively. All comparative methods are evaluated using the official pre-trained models and code provided by their authors to ensure fairness.

Quantitative comparisons. Table II presents quantitative results where MoA-VR consistently outperforms all existing all-in-one image and video restoration methods across both pixel-level (PSNR, SSIM) and perceptual quality metrics (LPIPS [70], MANIQA [71], CLIP-IQA [72], MUSIQ [73]). Under double degradation, MoA-VR achieves a significant PSNR of 23.47 dB, outperforming the strongest baseline BasicVSR++ by **3.02** dB, and improves LPIPS by 14.2%. When restoring videos containing three types of distortions, the advantages and robustness of our method become even more evident. Compared to BasicVSR++, the PSNR gain increases to **5.85** dB, and it still outperforms all other methods.

Qualitative comparisons. These solid quantitative results are visually corroborated by our qualitative comparisons (Fig. 8), which reveal MoA-VR's exceptional ability to preserve fine details and effectively suppress artifacts under demanding mixed distortion conditions. Unlike existing all-in-one image restoration techniques such as DA-CLIP, which may falter in video restoration due to differing degradation characteristics, or contemporary all-in-one video restoration models like BasicVSR++, which struggle with degradations such as haze and low-light not addressed in their design, MoA-VR effectively restores a wide array of degradations, yielding visually superior results.

These results demonstrate three main advantages of MoA-VR: (1) The collaborative operation of agents $\mathcal{A}_i$, $\mathcal{A}_r$, and $\mathcal{A}_a$ enables effective handling of complex, multi-type degradations, delivering state-of-the-art restoration performance. (2) MoA-VR exhibits remarkable scalability, supported by an extensible restoration toolbox, an adaptive routing strategy, and robust distortion identification coupled with quality assessment. It is the first all-in-one video restoration method capable of addressing eight types of mixed distortions, surpassing existing methods in versatility. (3) MoA-VR sets a new paradigm for multi-agent collaboration in video restoration, facilitating easy extension and optimization to diverse real-world scenarios by leveraging modular agent cooperation.

Comparison on Complex Degradations. We conducted additional experiments to validate the generalization of our model for real-world videos or multi-order degraded videos. Specifically, we generated a synthetic multi-order test set containing 50 videos, and additionally collected 50 real-world videos from existing video restoration datasets [55].

Table III presents quantitative results where MoA-VR consistently outperforms all existing all-in-one video restoration methods across four perceptual quality metrics (MANIQA [71], CLIP-IQA [72], MUSIQ [73], TOPIQ [74]). Under multi-order degradation, MoA-VR improves MUSIQ [71] by **40%**. When restoring real-world videos, the advantages of our method is evident as well, with improvement across all quality metrics compared with

Fig. 8. Performance comparison of MoA-VR with other all-in-one methods, using the first frame as an illustration.

TABLE III
COMPARISON OF MOA-VR WITH ALL-IN-ONE METHODS FOR COMPLEX VIDEO RESTORATION. WE HIGHLIGHT THE BEST AND SECOND-BEST VALUES FOR EACH METRIC.

	Multi-Order Degradation				Real-World Degradation			
	MANIQA↑	CLIP-IQA↑	MUSIQ↑	TOPIQ↑	MANIQA↑	CLIP-IQA↑	MUSIQ↑	TOPIQ↑
BasicVSR++ [9]	0.1708	0.2713	27.80	0.3233	0.2054	0.2564	42.35	0.3533
AverNet [12]	0.1569	0.2485	23.45	0.3091	0.2001	0.3028	41.15	0.3451
MoA-VR (Ours)	0.1801	0.2816	38.41	0.3556	0.2238	0.3108	43.31	0.3611

BasicVSR++ [9].

Fig. 9 presents quantitative comparison, which reveal MoA-VR's robustness on real-world videos. Compared with other all-in-one restorers, MoA-VR handles real-world scenarios more robustly due to its integration of various tools. It removes compression artifacts, noise, and blur in a step-by-step manner, while others typically solve single degradation, leaving residual compression artifacts (e.g. AverNet [12]) or noise (e.g. BasicVSR++ [9]).

Both quantitative and quantitative results demonstrate that our method generalizes robustly to more complex multi-order and real-world degradations.

C. Evaluation

Degradation Identification Accuracy. To assess both the zero-shot capability of VLMs in degradation recognition, we benchmark several open-source VLMs (e.g., Qwen, and LLaVA) on the classification task using MoA-VD. We adopt accuracy per degradation type and overall average accuracy as the evaluation metric. As shown in Table IV, VLMs exhibit strong generalization in identifying complex, mixed degradations from multimodal inputs. Among all evaluated models, the Qwen series consistently outperforms others, achieving an average accuracy of **51.18%** on degradation level prediction. Given its superior performance, we fine-tune Qwen on the MoA-VD training set and adopt it as the backbone for our degradation identification agent $\mathcal{A}_i$. After fine-tuning, our ($\mathcal{A}_i$) achieves an average accuracy of **87.2%** across all degradation types. Specifically, it performs best on blur (95.5%), haze (95.0%), and noise (94.0%), while maintaining expressive accuracy on artifact-heavy degradations like low-resolution and low-light ($\geq$80%).

Routing and Restoration Strategy. To examine how the order of degradation removal affects final restoration quality, we design and evaluate six routing strategies. These include reverse-order restoration, random restoration, and a fixed expert-defined order, as well as three variants of our Routing and Restoration Agent ($\mathcal{A}_r$). The first three serve as baselines, ranging from naive to human-guided sequences, while the latter three evaluate the contributions of learned experience and rollback.

Specifically, $\mathcal{A}_r$-Zero-Shot leverages LLM predictions without any prior experience or rollback; $\mathcal{A}_r$-Experience Only builds on training-set feedback to predict restoration sequences

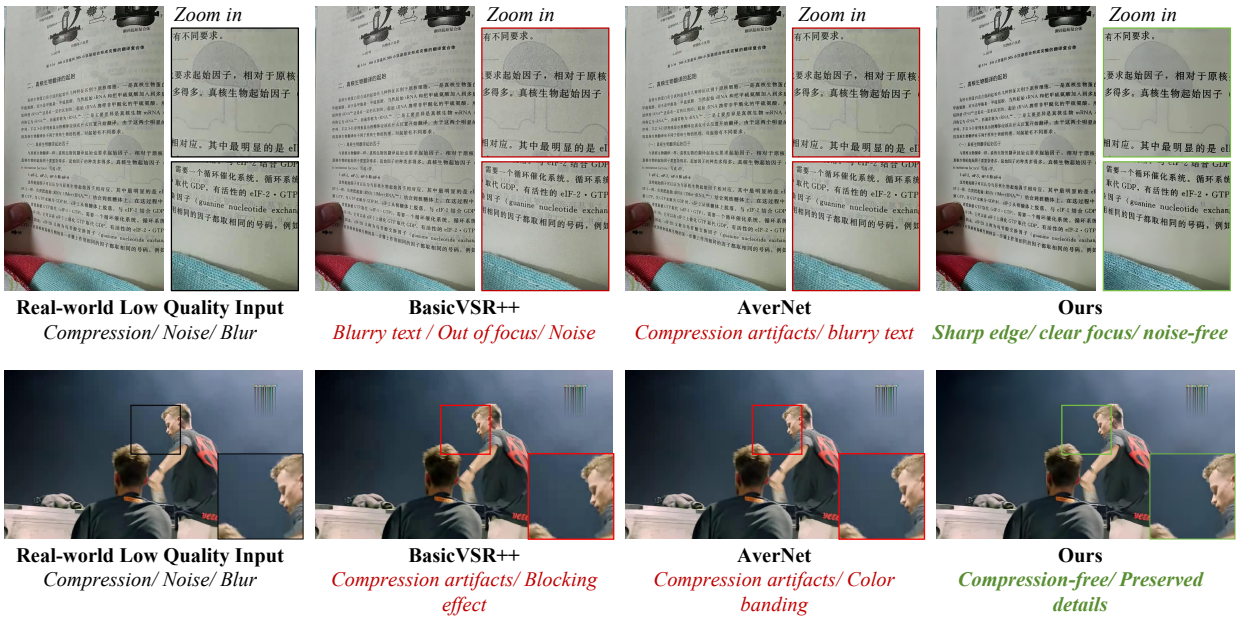


Fig. 9. Quantitative comparison of MoA-VR with other all-in-one video restoration methods on real world video restoration, using first frame as an illustration.

TABLE IV
BENCHMARK OF STATE-OF-THE-ART VLMS AND THE PROPOSED DEGRADATION IDENTIFICATION AGENT ON DEGRADATION LEVEL PREDICTION TASK. ♢, AND ♡ DENOTE THE ZERO-SHOT VLMS AND FINE-TUNED VLMS, RESPECTIVELY. THE BEST RESULTS ARE HIGHLIGHTED IN BEST, AND THE SECOND-BEST RESULTS ARE HIGHLIGHTED IN SECOND-BEST. THE EVALUATION METRIC IS PREDICTION ACCURACY.

Model	Noise	Compre.	Blur	Low Light	Rain	Haze	Low Res.	Low Fra.	Avg.
InternVL2 (1B) [23]	0.3800	0.5150	0.5575	0.2600	0.5325	0.4300	0.3550	0.4125	0.4303
InternVL2 (2B) [23]	0.5250	0.4925	0.4775	0.4050	0.5825	0.4350	0.4025	0.5850	0.4881
InternVL3 (1B) [24]	0.5375	0.5100	0.4050	0.4900	0.5875	0.3375	0.4125	0.4400	0.4650
LLaVA-NEXT-Video (7B) [75]	0.4075	0.2550	0.3800	0.2600	0.3350	0.2750	0.3450	0.2100	0.3084
mPLUG-Owl3 (1B) [76]	0.3750	0.2500	0.4475	0.2700	0.5200	0.2500	0.3150	0.1875	0.3269
mPLUG-Owl3 (7B) [76]	0.6450	0.4600	0.5175	0.4400	0.5550	0.4925	0.4350	0.5025	0.5059
Qwen2.5VL (3B) [26]	0.5250	0.2625	0.5900	0.2900	0.5525	0.4175	0.6450	0.6825	0.4956
Qwen2.5VL (7B) [26]	0.5700	0.4150	0.5450	0.3950	0.6500	0.4650	0.5750	0.4800	0.5118
$\mathcal{A}_i$-Ours	0.9400	0.8600	0.9550	0.7525	0.8850	0.9500	0.8000	0.8350	0.8722

but cannot dynamically correct mistakes; and $\mathcal{A}_r$-Ours integrates both experience learning and rollback to actively refine routing decisions. The expert-defined order consists of a set of static sequences, each tailored for a specific combination of degradation types. Table V shows the quantitative results. Note that all other model parameters and settings are held constant across experiments to ensure fair comparison.

The results reveal several important insights. Reverse-order restoration performs worst, confirming that the degradation process is not trivially reversible. Expert-defined ordering improves over naive strategies but remains suboptimal under mixed degradations due to its rigidness. Notably, the zero-shot LLM-based routing already surpasses handcrafted rules, demonstrating the potential of data-driven sequence planning. Furthermore, incorporating experience from training data leads to further gains.

Random and Reverse restoration orders yield the worst results, indicating that restoration order has a significant impact on final quality, and an incorrect order can leave degradations unremoved or even amplify them. Although the Expert Order outperforms random and reverse orders, it is still notably inferior to the automatic routing strategies, showing that fixed human-crafted orders are less effective than data-driven dynamic adjustments. Adapting the order to each video's specific degradation profile consistently outperforms any fixed order. Among the three dynamic-order variants, $\mathcal{A}_r$-Ours consistently surpasses $\mathcal{A}_r$-Experience, demonstrating that an "initial planning + real-time adjustment" approach is more capable of handling diverse degradation combinations and generalizes better than relying solely on past experience. In the Triple Degradation setting, the improvement in perceptual quality metrics is even larger than in the Double Degradation case, highlighting that the routing strategy becomes increasingly critical as degradation complexity increases.

Finally, when introducing both experience based routing and roll-back based rerouting, our full agent $\mathcal{A}_r$-Ours achieves the highest performance across all metrics. For instance, compared with non-intelligent methods, it yields a PSNR improvement of **19.89+1.76** dB and SSIM gain of **0.634+0.01** over the expert-defined strategy on the test set. The accumulated experience enables the agent to learn globally optimal routing patterns, while rollback acts as a corrective mechanism that prevents

TABLE V
COMPARISON OF OUR ROUTING STRATEGY WITH OTHER DEGRADATION REMOVAL STRATEGIES FOR MULTI-DEGRADED VIDEO RESTORATION. WE HIGHLIGHT BEST AND SECOND-BEST VALUES FOR EACH METRIC.

	Double Degradation						Triple Degradation					
	PSNR ↑	SSIM ↑	LPIPS ↓	MANIQA↑	CLIP-IQA↑	MUSIQ↑	PSNR ↑	SSIM ↑	LPIPS ↓	MANIQA↑	CLIP-IQA↑	MUSIQ↑
Reverse Order	18.24	0.6181	0.4719	0.1813	0.2405	27.72	17.32	0.5914	0.5102	0.1347	0.2124	22.56
Random Order	19.63	0.6208	0.4139	0.1857	0.2566	25.64	18.94	0.6150	0.4523	0.1482	0.2203	23.87
Expert Order	19.89	0.6340	0.3915	0.1855	0.3047	33.73	19.75	0.6321	0.4010	0.1860	0.2787	27.25
$\mathcal{A}_r$-Zero-Shot	20.31	0.6382	0.3798	0.1914	0.3125	34.02	20.12	0.6367	0.3894	0.1913	0.2776	27.92
$\mathcal{A}_r$-Experience	20.98	0.6415	0.3687	0.1976	0.3198	34.89	20.23	0.6442	0.3721	0.1928	0.2842	28.11
$\mathcal{A}_r$-Ours	21.65	0.6463	0.3544	0.2042	0.3263	35.43	21.11	0.6483	0.3713	0.2081	0.3012	29.74

TABLE VI
PERFORMANCE OF STATE-OF-THE-ART MODELS AND THE PROPOSED QUALITY ASSESSMENT AGENT ON OUR ESTABLISHED RES-VQ DATABASE IN TERMS OF QUALITY PREDICTION. ♠, ♢, AND ♡ DENOTE THE TRADITIONAL IQA METHODS, TRADITIONAL VQA METHODS, AND DNN-BASED VQA METRICS, RESPECTIVELY. IT SHOULD BE NOTED THAT ALL THE DNN-BASED MODELS ARE RE-TRAINED WITH THEIR DEFAULT SETTINGS. THE BEST RESULTS ARE HIGHLIGHTED IN BEST, AND THE SECOND-BEST RESULTS ARE HIGHLIGHTED IN SECOND-BEST.

Methods / Metrics	SRCC↑	KRCC↑	PLCC↑	RMSE↓
♠ NIQE [77]	0.3196	0.2222	0.1475	N/A
♠ QAC [78]	0.0501	0.0317	0.0046	20.18
♠ HOSA [79]	0.3699	0.2584	0.3259	23.36
♢ BMPRI [43]	0.2872	0.2056	0.3339	28.32
♢ VIDEVAL [44]	0.7137	0.5387	0.7586	9.66
♢ RAPIQUE [45]	0.8166	0.6284	0.8402	8.41
♡ DOVER* [80]	0.7751	0.5787	0.7589	19.20
♡ SimpleVQA* [47]	0.8007	0.6032	0.7844	13.81
♡ VSFA [81]	0.7576	0.5612	0.7196	11.03
♡ FAST-VQA [48]	0.7592	0.5554	0.7160	11.09
♡ DOVER [49]	0.7877	0.5834	0.7669	10.06
♡ SimpleVQA [47]	0.8596	0.6671	0.8593	7.85
♡ GSTVQA [50]	0.8902	0.7144	0.8938	7.62
$\mathcal{A}_a$-Ours	0.9165	0.7510	0.9284	5.65

the model from being trapped in suboptimal sequences. This combination proves essential in adapting to diverse and complex degradation scenarios, ultimately leading to superior restoration fidelity.

Quality Assessment Performance. To evaluate the effectiveness of our proposed quality assessment agent $\mathcal{A}_a$, we conduct comprehensive experiments on the Res-VQ dataset. We benchmark $\mathcal{A}_a$ against 11 representative no-reference VQA models, categorized into: (1) traditional NR-IQA methods (NIQE [77], QAC [78], HOSA [79]), (2) traditional NR-VQA models (TLVQA [82], VIDEVAL [44], RAPIQUE [45]), and (3) deep learning-based NR-VQA methods (VSFA [81], GSTVQA [50], SimpleVQA [47], Fast-VQA [48], Dover [49]). All models are retrained on Res-VQ with an 80:20 training/testing split. Evaluation metrics include SRCC, PLCC, KRCC, and RMSE, assessing correlation with human MOS scores and prediction accuracy.

As shown in Table VI, traditional NR-IQA models exhibit poor alignment with human perception on restored video quality (e.g., NIQE: SRCC = 0.320), mainly due to the lack of temporal modeling. NR-VQA models offer moderate improvements (e.g., VIDEVAL: SRCC = 0.713), but they still struggle to handle diverse restoration artifacts. Deep learning-based methods such as Simple-VQA and GSTVQA also perform poorly in a zero-shot setting (SRCC = 0.80 and 0.77, respectively), largely because these models are predominantly trained on User-Generated Content (UGC) videos, which differ substantially from the intermediate restored outputs of our agent-based system. Consequently, their generalization to restored video quality assessment is limited. After being trained on our constructed Res-VQ dataset, deep learning-based models demonstrate significant improvements (Simple-VQA: SRCC = 0.859; GSTVQA: SRCC = 0.890), which validates the importance of building a dedicated database for restored video quality assessment.

Building on this, our agent $\mathcal{A}_a$ achieves the highest performance across all metrics (SRCC = **0.9165**, PLCC = **0.9284**, KRCC = **0.7510**, RMSE = **5.65**). The results highlight the superior capability of LLM-enhanced architectures in fusing visual and textual cues to assess quality from a human-centric perspective. By leveraging the semantic richness of prompt-aware inputs and the cross-modal reasoning ability of the LLM backbone, $\mathcal{A}_a$ effectively aligns predicted scores with human opinions. Moreover, it enables quality-aware routing guidance in our restoration pipeline, providing not only accurate scoring but also practical value in downstream decision-making.

Ablation Study on Degradation Identification. To assess the generalization ability of the identification agent, we performed additional experiments by modifying the degradation generation pipeline. We constructed an additional test set by replacing all degradation generation procedures so that they differ from those used in the original MoA-VD test set (Group A). Specifically, we introduced impulse noise, motion blur, H.265 compression, and adopted the pipeline of [83] to synthesize low-light and rain. For low resolution and low frame rate, we randomly varied their scaling factors. Degradations were applied in both single-order and multi-order combinations, yielding 400 new test samples (denoted as Group B). As seen in Fig. 10 and Table VII, on this unseen dataset, our agent achieved an average recognition accuracy of 84%. In particular, the recognition accuracy for noise and blur exceeded 90%, while low-resolution was more easily confused with blur, and low-light performance varied depending on video capture conditions. These results indicate that our degradation identification agent learns generalizable degradation features rather than merely memorizing patterns, and remains robust when facing various degradations.

Ablation Study on Rollback and Experience. To vali-

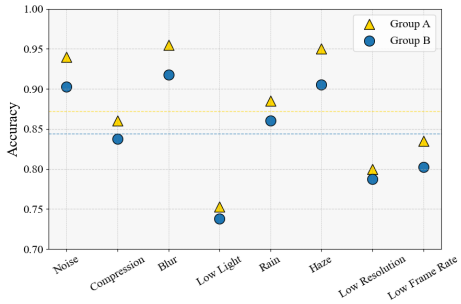


Fig. 10. Ablation study on degradation identification. Group A is the original test set and Group B is an additional test set generated with an altered degradation pipeline.

TABLE VII
PREDICTION ACCURACY OF THE DEGRADATION IDENTIFICATION AGENT ACROSS DEGRADATION TYPES. GROUP A USES THE ORIGINAL TEST SET (IN-DOMAIN), WHILE GROUP B DENOTES OUT-OF-DOMAIN DATA. HIGHER IS BETTER.

Dataset	Noise	Compre.	Blur	Low Light	Rain	Haze	Low Res.	Low Fra.	Avg.
Group A (In-domain)	0.9400	0.8600	0.9550	0.7525	0.8850	0.9500	0.8000	0.8350	0.8722
Group B (Out-of-domain)	0.9025	0.8375	0.9175	0.7375	0.8600	0.9050	0.7875	0.8025	0.8438

date the effectiveness of key components in our MoA-VR framework, we conduct ablation studies on the experience mechanism and the rollback mechanism, as illustrated in Fig. 12 and Fig. 11.

Without the experience mechanism, the agent selects restoration sequences purely based on its innate policy, which often results in suboptimal decisions. For instance, in Fig. 12(a), the agent mistakenly applies denoising before decompression, treating compression artifacts as noise and causing over-smoothing and amplified residuals. In contrast, Fig. 12(b) demonstrates that with the experience mechanism, the agent first performs decompression followed by denoising, yielding improved perceptual quality and structural fidelity.

Similarly, the rollback mechanism is crucial for correcting poor routing decisions even when guided by experience. Fig. 11(a) shows that performing deblurring before super-resolution, despite mild blur, induces over-smoothing, compromising the effectiveness of super-resolution. With the rollback mechanism (Fig. 11(b)), the agent receives degradation feedback, rolls back, and reroutes to super-resolution first, resulting in better detail preservation.

These results underscore the necessity of both mechanisms. The experience mechanism facilitates better initial decision-making based on prior observations, while the rollback mechanism provides the flexibility to revise routes dynamically. This dual capability empowers MoA-VR with expert-like adaptability, critical for robust restoration across diverse real-world degradations.

Cross-dataset Validation of Quality Assessment Agent. We conduct cross-dataset validation to assess generalization ability of our quality assessment agent. To the best of our knowledge, Res-VQ is the first VQA dataset specifically designed for restored videos. Hence, we choose existing UGC VQA database as validation. We train SimpleVQA [47], GSTVQA [48], and our quality assessment agent on Res-VQ-train and evaluate on Fine-VD-test [55], reporting SRCC, PLCC, KRCC, and RMSE. As shown in Table VIII, our quality assessment agent consistently outperforms state-of-the-art VQA methods (e.g., **+0.23** SRCC and a **13.6** RMSE reduction relative to SimpleVQA [47]). This indicates that our quality assessment agent exhibits strong generalization to novel content not seen during training.

TABLE VIII
THE CROSS-DATASET EVALUATION RESULTS. THE MODELS ARE TRAINED ON OUR RES-VQ AND TESTED ON OTHER UGC VQA DATASET.

Test on:	Fine-VD			
Train on: Res-VQ	SRCC↑	PLCC↑	KRCC↑	RMSE↓
SimpleVQA [47]	0.3597	0.3305	0.2493	24.82
GSTVQA [50]	0.3345	0.3238	0.2251	18.83
Ours	**0.5884**	**0.6211**	**0.4195**	**11.22**

Time Complexity Analysis. We analyze the computational complexity of different routing strategies. For a video with n degradation types,

(1) Full Search (Exhaustive): Full search strategy evaluates all possible restoration orders. Hence, the time complexity of full search is

$$T_{\text{full search}} = O(n \times n!) \tag{8}$$

which grows factorially with the number of degradations and becomes intractable for large n.

(2) Tree Search: Tree search improves upon full search by eliminating the outer n factor. Its complexity can be expanded as

$$T_{\text{tree search}} = O\left(n! \sum_{k=0}^{n-1} \frac{1}{k!}\right) \tag{9}$$

As $n \to \infty$, the summation

$$\sum_{k=0}^{n-1} \frac{1}{k!} \longrightarrow e,$$

which is a constant. Therefore, the asymptotic time complexity is

$$T_{\text{tree}} = O(n!). \tag{10}$$

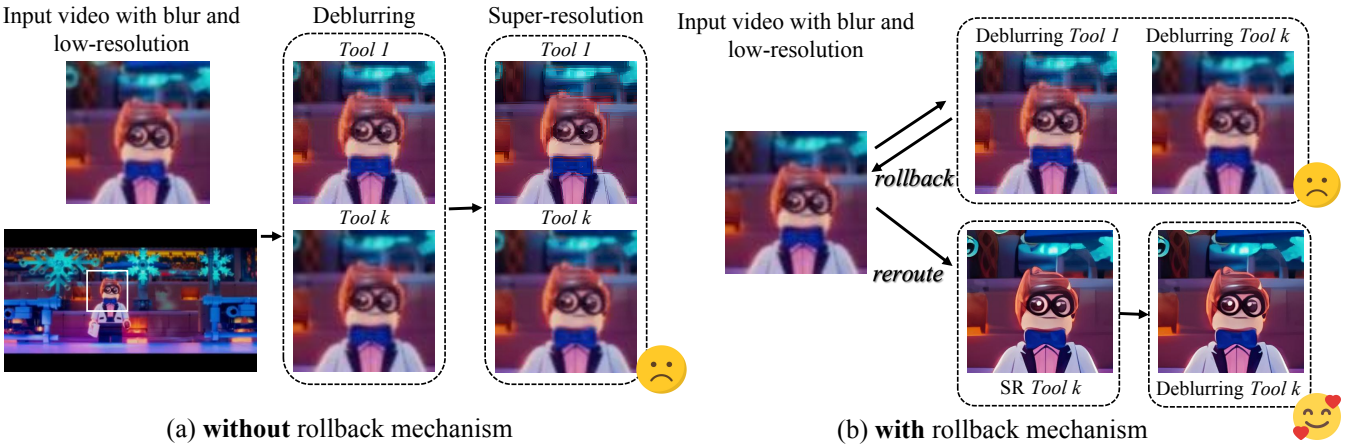
(3) Ours: Building on tree search, our method leverages an LLM-based predictor to propose a highly probable restoration sequence. At each step, the predictor succeeds with probability $p \in (0, 1]$. When a wrong decision is made, rollback occurs, leading to an expected cost of $(1-p) \cdot i$ at depth i. Summing over all n steps yields an overall expected complexity of

$$T_{\text{ours}} = O!\left(\frac{n}{p} + (1-p)n^2\right), \tag{11}$$

which remains polynomial and significantly lower than the factorial complexity of exhaustive or tree search.

Fig. 13 visualizes the time complexity comparison among three routing algorithms, indicating that our routing strategy achieves significantly lower complexity, especially as the number of degradations increases.

Runtime and Tool Invocation Experiments. We conduct additional experiments comparing our routing agent strategy



(a) **without** rollback mechanism (b) **with** rollback mechanism

Fig. 11. Ablation study on the rollback mechanism. (a) Without rollback, the agent is unable to revise suboptimal decisions, leading to degraded results. (b) Enabled by rollback, the agent re-evaluates and adjusts its restoration route, producing clearer and more accurate outputs.

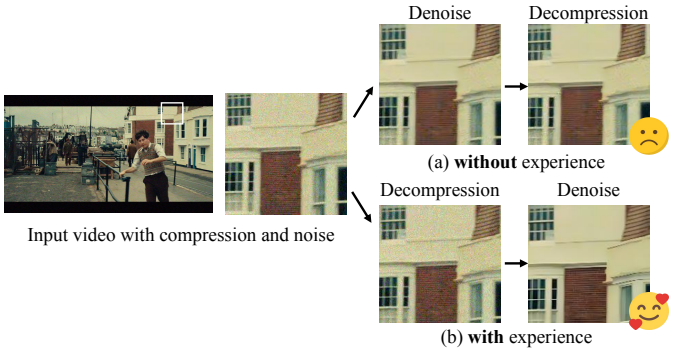


Fig. 12. Ablation study on the experience mechanism. (a) Without experience guidance, the agent misorders the operations, resulting in significant quality degradation. (b) Leveraging the experience mechanism, the agent selects an effective restoration sequence, yielding improved perceptual quality and structure preservation.

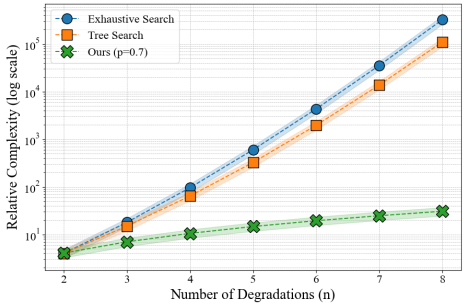


Fig. 13. Time complexity comparison between three routing algorithms.

with exhaustive search to further explore the routing efficiency of our approach. We fix a tool for each degradation removal and compared the time complexity of our routing strategy and the exhaustive search. The runtime per frame and tool invocations are reported in Fig. 14. It can be observed that, compared with exhaustive search, our method achieves a 66% reduction in runtime on triple degradations.

VQA-driven Optimization Baseline. We augment BasicVSR++ [9] with a differentiable VQA loss derived from

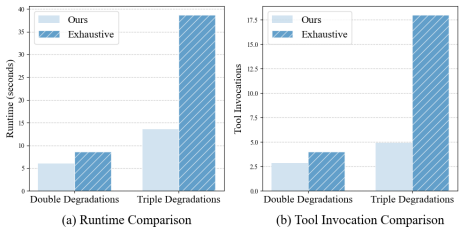


Fig. 14. Additional experiments in runtime and tool invocation.

our quality assessment agent. Specifically, in addition to the pixel-level Charbonnier reconstruction loss $\mathcal{L}_{\text{pixel}}$, we introduce a VQA loss that aligns the predicted perceptual quality of the restored video with that of the ground-truth video:

$$\mathcal{L}_{\text{vqa}} = \frac{1}{B}\sum_{b=1}^{B}\left|\Phi(\hat{V}^{(b)}) - \Phi(V_{\text{gt}}^{(b)})\right|,$$

where Φ denotes the frozen VQA network, B is the batch size, and $b \in \{1, \ldots, B\}$ indexes the training samples. During training, gradients are back-propagated only to the restoration backbone, while the VQA model remains fixed. The overall objective is then defined as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{pixel}}\mathcal{L}_{\text{pixel}} + \lambda_{\text{vqa}}\mathcal{L}_{\text{vqa}},$$

with $\lambda_{\text{vqa}} = 10^{-4}$. We follow the standard BasicVSR++ fine-tuning protocol and schedule, using MoA-VD as training data.

Table IX shows that in both double- and triple-degradation settings, the VQA-driven baseline increases perceptual quality but reduces pixel-level fidelity (e.g., under double degradation: ΔMUSIQ $+2.77$, ΔPSNR $-0.33\,\text{dB}$), and it remains clearly inferior to MoA-VR overall. Notably, the VQA-driven baseline lacks explicit degradation disentanglement and routing, which limits its ability to robustly handle compound degradations.Notably, the VQA-driven baseline lacks explicit degradation disentanglement and routing, which limits its ability to robustly handle compound degradations.

TABLE IX
COMPARISON OF MOA-VR WITH ALL-IN-ONE METHODS FOR MULTI-DEGRADED VIDEO RESTORATION. WE HIGHLIGHT **BEST** IN **BOLD**.

	Double Degradation				Triple Degradation			
	PSNR ↑	SSIM ↑	LPIPS ↓	MUSIQ↑	PSNR ↑	SSIM ↑	LPIPS ↓	MUSIQ↑
BasicVSR++ [9]	20.45	0.6049	0.5146	25.65	17.09	0.5026	0.5579	24.65
BasicVSR++ + VQA loss	20.12	0.5981	0.4013	28.42	16.95	0.4897	0.4126	26.31
Δ ↑ (vs. BasicVSR++)	-0.33	-0.0068	-0.1133	+2.77	-0.14	-0.0129	-0.1453	+1.66
MoA-VR-Ours	**23.47**	**0.6852**	**0.3386**	**36.14**	**22.94**	**0.6497**	**0.3074**	**30.43**

V. CONCLUSION

We introduce MoA-VR, a novel framework for all-in-one video restoration especially in complex mixed degradation scenarios. This system is built upon a mixture-of-agent architecture, comprising three collaborative agents: degradation identification agent, routing and restoration agent, and restoration quality assessment agent. These agents work together in a closed-loop process to emulate the reasoning of human experts. Specifically, we utilize a VLM to develop the degradation identification agent and evaluate its predictive performance. Additionally, we incorporate an agent driven by an LLM that autonomously adapts routing strategies by observing the effect of degradation removal. To assess the quality of the output, we create a restoration video quality (Res-VQ) dataset and develop a VQA model focused on restoration tasks. Extensive experiments validate that MoA-VR successfully handles diverse and compound degradation issues, consistently outperforming existing benchmarks in both objective metrics and perceptual quality. These findings emphasize the potential of integrating multimodal intelligence and modular reasoning into general-purpose video restoration systems.

REFERENCES

[1] H. Wu, Y. Yang, H. Xu, W. Wang, J. Zhou, and L. Zhu, "Rainmamba: Enhanced locality learning with state space models for video deraining," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, p. 7881–7890.

[2] J. Xu, X. Hu, L. Zhu, Q. Dou, J. Dai, Y. Qiao, and P.-A. Heng, "Video dehazing via a multi-range temporal alignment network with physical prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] J. Pan, B. Xu, J. Dong, J. Ge, and J. Tang, "Deep discriminative spatial and temporal network for efficient video deblurring," in *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, Feb 2023.

[4] W. Li, G. Wu, W. Wang, P. Ren, and X. Liu, "Fastllve: Real-time low-light video enhancement with intensity-aware lookup table," in *Proceedings of the 31th ACM International Conference on Multimedia*, 2023.

[5] Z. Guo, W. Li, and C. C. Loy, "Generalizable implicit motion modeling for video frame interpolation," in *Advances in Neural Information Processing Systems*, 2024.

[6] C. Rota, M. Buzzelli, and J. van de Weijer, "Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models," in *European Conference on Computer Vision*. Springer, 2024, pp. 36–53.

[7] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Investigating tradeoffs in real-world video super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[8] S. Zhou, P. Yang, J. Wang, Y. Luo, and C. C. Loy, "Upscale-A-Video: Temporal-consistent diffusion model for real-world video super-resolution," in *CVPR*, 2024.

[9] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "On the generalization of BasicVSR++ to video deblurring and denoising," *arXiv preprint arXiv:2204.05308*, 2022.

[10] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, and L. Van Gool, "Vrt: A video restoration transformer," *arXiv preprint arXiv:2201.12288*, 2022.

[11] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, and L. Van Gool, "Recurrent video restoration transformer with guided deformable attention," *arXiv preprint arXiv:2206.02146*, 2022.

[12] H. Zhao, L. Tian, X. Xiao, P. Hu, Y. Gou, and X. Peng, "Avernet: All-in-one video restoration for time-varying unknown degradations," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 127 296–127 316.

[13] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[14] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the International conference on machine learning (ICML)*, 2022, pp. 12 888–12 900.

[15] O. Team, "Chatgpt-4o," https://chatgpt.com/, 2024, accessed: 2025-03-08.

[16] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[17] H. Duan, X. Min, S. Wu, W. Shen, and G. Zhai, "Uniprocessor: A text-induced unified low-level image processor," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[18] H. Duan, W. Shen, X. Min, Y. Tian, J.-H. Jung, X. Yang, and G. Zhai, "Develop then rival: A human vision-inspired framework for superimposed image decomposition," *IEEE Transactions on Multimedia*, vol. 25, pp. 4267–4281, 2023.

[19] S. Gao, H. Duan, X. Li, K. Fu, Y. Peng, Q. Xu, Y. Chang, J. Wang, X. Min, and G. Zhai, "Quality-guided skin tone enhancement for portrait photography," 2024. [Online]. Available: https://arxiv.org/abs/2406.15848

[20] Q. Hu, Q. He, H. Zhong, G. Lu, X. Zhang, G. Zhai, and Y. Wang, "Var-fvv: View-adaptive real-time interactive free-view video streaming with edge computing," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2025.

[21] OpenAI, "Gpt-4 technical report," 2024.

[22] A. Meta, "Llama 3.2: Revolutionizing edge ai and vision with open, customizable models," *Meta AI Blog. Retrieved December*, vol. 20, p. 2024, 2024.

[23] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[24] W. Wang, Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, J. Zhu, X. Zhu, L. Lu, Y. Qiao, and J. Dai, "Enhancing the reasoning ability of multimodal large language models via mixed preference optimization," *arXiv preprint arXiv:2411.10442*, 2024.

[25] Z. Zhou, J.-X. Shi, P.-X. Song, X.-W. Yang, Y.-X. Jin, L.-Z. Guo, and Y.-F. Li, "Lawgpt: A chinese legal knowledge-enhanced large language model," 2024.

[26] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[27] M. Wooldridge and N. Jennings, "Intelligent agents: theory and practice the knowledge engineering review," 1995.

[28] F. Hong, Y. Huang, Z. Zhao, Z. Zhou, J. Yao, D. Li, Y. Zhang, and Y. Wang, "Dual-granularity sinkhorn distillation for enhanced learning from long-tailed noisy data," *Machine Learning*, 2025.

[29] X. Cao, M. Xu, X. Yu, J. Yao, W. Ye, S. Huang, M. Zhang, I. W. Tsang, Y.-S. Ong, J. T. Kwok, and H.-T. Shen, "Analytical survey of learning with low-resource data: From analysis to investigation," *ACM Computing Surveys*, 2025.

[30] Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008.

[31] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, ser. ICML'94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, p. 157–163.

[32] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '23. New York, NY, USA: Association for Computing Machinery, 2023.

[33] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 38 154–38 180.

[34] X. Zhu, Y. Chen, H. Tian, C. Tao, W. Su, C. Yang, G. Huang, B. Li, L. Lu, X. Wang, Y. Qiao, Z. Zhang, and J. Dai, "Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory," *arXiv preprint arXiv:2305.17144*, 2023.

[35] X. Li, P. Li, M. Liu, D. Wang, J. Liu, B. Kang, X. Ma, T. Kong, H. Zhang, and H. Liu, "Towards generalist robot policies: What matters in building vision-language-action models," *arXiv preprint arXiv:2412.14058*, 2024.

[36] K. Zhu, J. Gu, Z. You, Y. Qiao, and C. Dong, "An intelligent agentic system for complex image restoration problems," in *The Thirteenth International Conference on Learning Representations*, 2025.

[37] H. Chen, W. Li, J. Gu, J. Ren, S. Chen, T. Ye, R. Pei, K. Zhou, F. Song, and L. Zhu, "Restoreagent: Autonomous image restoration agent via multimodal large language models," 2024.

[38] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[39] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1279–1288.

[40] Z. Zhong, Y. Gao, Y. Zheng, B. Zheng, and I. Sato, "Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 284–301, 2023.

[41] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. Mu Lee, "Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[42] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, p. 1106–1125, Feb. 2019.

[43] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Transactions on Broadcasting (TBC)*, vol. 64, no. 2, pp. 508–517, 2018.

[44] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 4449–4464, 2021.

[45] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Rapique: Rapid and accurate video quality prediction of user generated content," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 425–440, 2021.

[46] H. Duan, X. Min, Y. Zhu, G. Zhai, X. Yang, and P. Le Callet, "Confusing image quality assessment: Toward better augmented reality experience," *IEEE Transactions on Image Processing*, vol. 31, p. 7206–7221, 2022. [Online]. Available: http://dx.doi.org/10.1109/TIP.2022.3220404

[47] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for ugc videos," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2022, pp. 856–865.

[48] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 538–554.

[49] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023, pp. 20 144–20 154.

[50] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, "Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 32, no. 4, pp. 1903–1916, 2021.

[51] H. Duan, X. Min, W. Sun, Y. Zhu, X.-P. Zhang, and G. Zhai, "Attentive deep image quality assessment for omnidirectional stitching," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 6, pp. 1150–1164, 2023.

[52] H. Duan, W. Shen, X. Min, Y. Tian, J.-H. Jung, X. Yang, and G. Zhai, "Develop then rival: A human vision-inspired framework for superimposed image decomposition," *IEEE Transactions on Multimedia*, vol. 25, pp. 4267–4281, 2022.

[53] L. Yang, H. Duan, Y. Zhu, X. Liu, L. Liu, Z. Xu, G. Ma, X. Min, G. Zhai, and P. L. Callet, "Omni^2: Unifying omnidirectional image generation and editing in an omni model," *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

[54] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *International Conference on Computer Vision Workshops (ICCVW)*.

[55] H. Duan, Q. Hu, J. Wang, L. Yang, Z. Xu, L. Liu, X. Min, C. Cai, T. Ye, X. Zhang *et al.*, "Finevq: Fine-grained user generated content video quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[56] L. Liu, H. Duan, Q. Hu, L. Yang, C. Cai, T. Ye, H. Liu, X. Zhang, and G. Zhai, "F-bench: Rethinking human preference evaluation metrics for benchmarking face generation, customization, and restoration," 2024.

[57] Z. Xu, H. Duan, B. Liu, G. Ma, J. Wang, L. Yang, S. Gao, X. Wang, J. Wang, X. Min *et al.*, "Lmm4edit: Benchmarking and evaluating multimodal image editing with lmms," *arXiv preprint arXiv:2507.16193*, 2025.

[58] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-In-One Image Restoration for Unknown Corruption," in *IEEE Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, Jun. 2022.

[59] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[60] H. Chen, W. Li, J. Gu, J. Ren, S. Chen, T. Ye, R. Pei, K. Zhou, F. Song, and L. Zhu, "Restoreagent: Autonomous image restoration agent via multimodal large language models," 2025.

[61] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[62] B. Series, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT*, pp. 500–13, 2012.

[63] H. Fan, Y. Li, B. Xiong, W.-Y. Lo, and C. Feichtenhofer, "Pyslowfast," https://github.com/facebookresearch/slowfast, 2020.

[64] V. Potlapalli, S. W. Zamir, S. Khan, and F. Khan, "Promptir: Prompting for all-in-one image restoration," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[65] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Controlling vision-language models for universal image restoration," *arXiv preprint arXiv:2310.01018*, 2023.

[66] J. Cao, Y. Cao, L. Pang, D. Meng, and X. Cao, "Hair: Hypernetworks-based all-in-one image restoration," 2024. [Online]. Available: https://arxiv.org/abs/2408.08091

[67] Y. Liu, T. Huang, W. Dong, F. Wu, X. Li, and G. Shi, "Low-light image enhancement with multi-stage residue quantization and brightness-aware attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 12 140–12 149.

[68] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Transactions on Image Processing*, vol. 32, pp. 1927–1941, 2023.

[69] I. Team, "Internlm: A multilingual language model with progressively enhanced capabilities," 2023.

[70] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[71] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1191–1200.

[72] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *AAAI*, 2023.

[73] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5128–5137.

[74] C. Chen, J. Mo, J. Hou, H. Wu, L. Liao, W. Sun, Q. Yan, and W. Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE Transactions on Image Processing*, vol. 33, pp. 2404–2418, 2024.

[75] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models," *arXiv preprint arXiv:2407.07895*, 2024.
[76] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou, "mplug-owl3: Towards long image-sequence understanding in multi-modal large language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
[77] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters (SPL)*, vol. 20, no. 3, pp. 209–212, 2012.
[78] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 995–1002.
[79] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing (TIP)*, vol. 25, no. 9, pp. 4444–4457, 2016.
[80] H. Wu, E. Zhang, L. Liao, C. Chen, J. H. Hou, A. Wang, W. S. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.
[81] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2019, pp. 2351–2359.
[82] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 12, pp. 5923–5938, 2019.
[83] Y. Lin, Z. Lin, H. Chen, P. Pan, C. Li, S. Chen, W. Kairun, Y. Jin, W. Li, and X. Ding, "Jarvisir: Elevating autonomous driving perception with intelligent image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.