REVRA: REFLECTIVE VIDEO RESTORATION AGENT WITH SELF-CORRECTIVE DYNAMIC PIPELINES

Anonymous authors
Paper under double-blind review

ABSTRACT

Degradations in real-world user-generated videos are inherently complex and multifaceted, making monolithic restoration tools struggle to decouple and resolve multiple, overlapping artifacts. While recent agentic restoration systems offer greater flexibility by chaining specialized restoration tools into pipelines, they suffer from three limitations: planning relies solely on the input video and static tool priors, making it difficult to select appropriate toolboxes; execution follows a fixed pipeline that cannot adapt to unexpected artifacts introduced by tools; and evaluation relies on absolute quality scores that poorly align with human perception. We propose ReVRA, a novel **Re**flective **V**ideo **R**estoration **A**gent that constructs adaptive restoration tool sequences through dynamic self-correction. ReVRA decomposes complex restoration into self-corrective stages that re-examine the current video state to adapt tool usage to unexpected artifacts. To improve planning, ReVRA replaces one-shot prediction with toolbox decisions guided by instance-aware evolving memory and verified by execution outcomes, which also refine the memory to provide increasingly accurate priors. Moreover, ReVRA evaluates restoration results via relative comparisons, better aligning with the human preference selection process. Experiments across various benchmarks spanning synthetic and real-world degradation scenarios demonstrate that ReVRA achieves superior restoration quality across a wide spectrum of complex degradations, validating the effectiveness of the reflective agentic framework for real-world video restoration.

1 INTRODUCTION

User-generated content (UGC) has become a cornerstone of digital media. However, due to imperfect capture conditions and cascaded media processing pipelines, UGC videos often suffer from a complex mixture of intertwined degradations, typically including compression artifacts, blur, noise, and low resolution. Traditional restoration methods (Xu et al., 2022; Zamir et al., 2022) are tailored to isolated degradations commonly studied in image restoration, such as rain, haze, low-light conditions, and underwater distortion. These methods are ill-suited to the degradation characteristics of UGC videos and cannot handle complex coupled degradations. Although all-in-one methods (Janjua et al., 2026; Zhao et al., 2024) address multiple degradation types, their static architectures still limit adaptability. Consequently, restoring real-world UGC videos remains highly challenging.

Recently, the emergence of Vision-Language Model (VLM)-based agentic systems has offered a new perspective on restoration. With this design, agent-based methods (Zhu et al., 2025; Zuo et al., 2025) have achieved notable success in image restoration, yet the distinct degradation distribution of UGC videos prevents direct transfer. Moreover, existing restoration agents share the following limitations:

Toolbox-fixed planning amplifies restoration errors. As shown in Fig. 1(a), existing agents (Chen et al., 2024; Liu et al., 2025; Zuo et al., 2025) typically determine the complete toolbox (with tools of the same restoration type) sequence solely from the input video all at once, at most adjusting the execution order via rollback without modifying the toolbox types. However, restoration effects are inherently unpredictable: a toolbox may introduce new artifacts (e.g., ringing after super-resolution) that subsequent toolboxes further amplify, or may inadvertently resolve additional degradations that render subsequent toolboxes redundant. Consequently, a pipeline fixed at initialization progressively deviates from the desired outcome as restoration advances.

One-shot prediction from static priors is inherently unreliable. In existing agents, the toolbox sequence is solely dictated by the VLM, which can only reference static tool priors (whether injected via retrieval from memory or encoded through training), making it difficult to predict the optimal plan.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

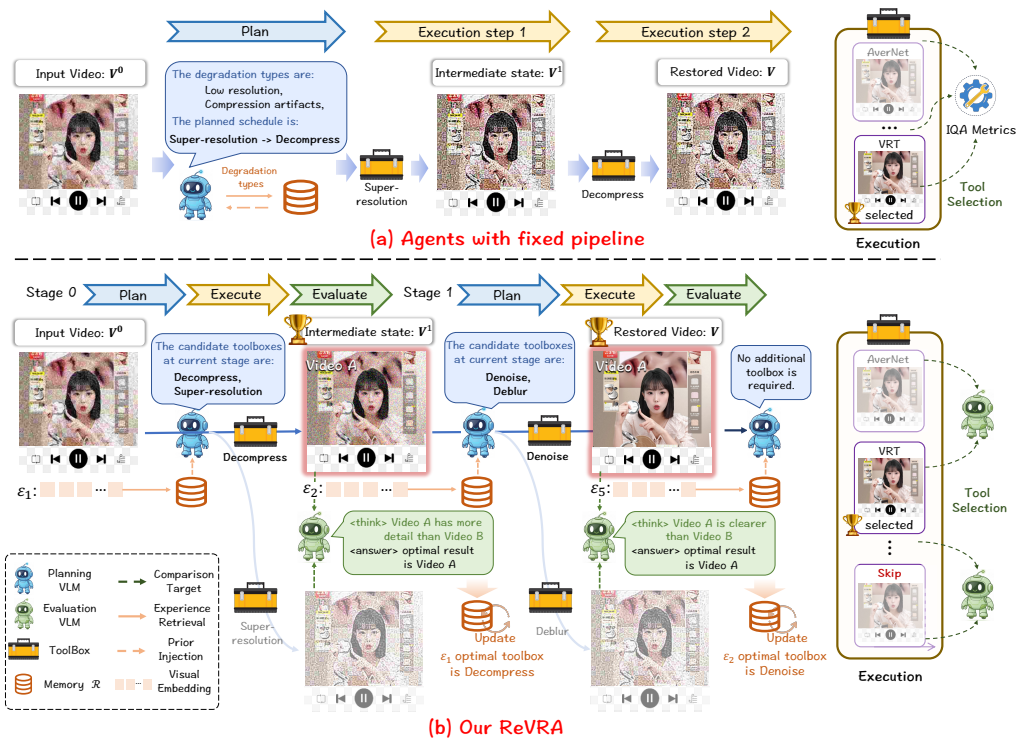

Figure 1: Comparison between existing agents with fixed pipeline and the ReVRA framework. (a) Existing agents determine the entire toolbox sequence at once based on the input video and select the best tool within each toolbox using absolute quality scores. (b) ReVRA consists of multiple plan–execute–evaluate stages, where each stage selects candidate toolboxes conditioned on the current video state and identifies the optimal tool via pairwise comparison as the stage output.

When the input video differs significantly from the memory records or deviates from the training distribution, the VLM's judgment may falter, leading to suboptimal planning.

Absolute quality scores misalign with human perception. Most agentic systems (Zhou et al., 2025; Zuo et al., 2025) rely on absolute IQA/VQA scores to select the optimal tool from each toolbox or determine the pipeline execution order. However, assessing video quality in isolation without reference is prone to high variance (Mittal et al., 2012a; Wen et al., 2026), fails to accurately capture artifact textures in the content, and is therefore difficult to align with human preferences.

In general, let tool f denote a restoration network as a basic operation unit of restoration. Assume we have a tool set $\boldsymbol{\mathcal{S}}$ that can be partitioned into M toolboxes $\{\boldsymbol{\mathcal{F}}_i\}_{i=1}^{M}$, where each toolbox is defined as $\boldsymbol{\mathcal{F}}_i = \{f_{i,j} \mid 1 \le j \le M_i\}$. Thus, $\boldsymbol{\mathcal{S}}$ can be expressed as $\boldsymbol{\mathcal{S}} = \{f_{i,j} \mid 1 \le i \le M, 1 \le j \le M_i\}$, where i and j denote the toolbox index and the tool index, respectively. Note that in this context, a toolbox $\boldsymbol{\mathcal{F}}_i$ often contains restoration tools of the same type. The restoration pipeline is defined as $\mathcal{T} := \prod f_t$, representing a sequence of tools, where t denotes the index to the tool set $\boldsymbol{\mathcal{S}}$. An optimal pipeline may be achieved by searching the best tool for each element of the sequence. Here, f_t may correspond to f_{i_t,j_t}, which is indexed first by i_t to the toolbox $\boldsymbol{\mathcal{F}}_{i_t}$ and then by j_t to the specific tool within it. To reduce the search space, such a two-dimensional indexing method can be utilized. To get the final output video, e.g., during inference, a sequence of the tools as indicated by $\mathcal{T}$ will be executed sequentially from the input video.

We propose ReVRA to address the above limitations. As illustrated in Fig. 1(b), ReVRA decomposes the restoration into multiple stages, each performing a complete plan–execute–evaluate workflow for tool selection and execution. Unlike existing methods that determine the entire toolbox sequence all at once, ReVRA re-selects the candidate toolboxes conditioned on the intermediate state at each stage, identifies the optimal toolbox based on execution outcomes, and updates the memory with this experience to guide future planning. This closed-loop design is essential: the effect of any restoration tool (including artifacts it may inadvertently introduce) cannot be reliably predicted before execution, making dynamic re-planning indispensable. Additionally, we introduce a skip option in each toolbox,

which performs no operation and allows the agent to abstain when none of the available tools is expected to improve quality, thereby enhancing system fault tolerance.

To improve inference efficiency, we leverage historical experience to refine the scope of candidate toolboxes. Specifically, we design an Instance-aware Evolving Memory (IEM) from which the most relevant toolbox is retrieved given the current state at each planning step, providing a reference for the planning VLM. Unlike existing methods (Liu et al., 2025; Zhu et al., 2025; Zuo et al., 2025) shown in Fig. 1(a), the IEM is indexed by video embeddings rather than coarse degradation categories, enabling differentiated guidance for videos that share the same degradation type but have distinct restoration needs. Moreover, the IEM is not static but can be automatically updated online. By introducing a contrastive update mechanism, the IEM continuously improves its retrieval accuracy during inference, empowering the agent with self-evolution capability.

Furthermore, we adopt a pairwise comparison strategy, where each candidate tool results and pipeline are sequentially evaluated against the current best via VLM judgment. This reformulates quality assessment into relative judgments of superiority, which are generally more stable and better aligned with human perception than absolute scoring.

Finally, existing benchmarks (Chan et al., 2022c; Lu et al., 2024) suffer from biased distributions and limited categories, we thus extend these datasets to construct a comprehensive UGC video restoration benchmark, covering 7 scenario types and 4 degradation categories. Our main contributions are:

- **ReVRA**, a reflective video restoration agent that formulates restoration as a recursive, state-conditioned decision process, achieving closed-loop self-correction by grounding each toolbox selection in the current intermediate state rather than the fixed initial input; a per-toolbox skip mechanism further improves fault tolerance.
- **Prior-informed and outcome-verified planning**. At each stage, the agent proposes only a candidate toolbox set, with the actual selection determined by observed execution outcomes. The result then updates the Instance-aware Evolving Memory, refining the prior and narrowing the search space for subsequent planning.
- **Comparison-based evaluation strategy**, which replaces absolute scoring with relative judgment to align naturally with how humans perceive visual quality, achieving more consistent restoration quality evaluation.
- **UGC video restoration benchmark**, a real-world UGC video benchmark that covers diverse scenarios and various degradation types for comprehensive evaluation.

2 Related Work

2.1 Video Restoration

Video restoration aims to reconstruct high-quality videos from degraded inputs, and existing methods broadly fall into degradation-specific and all-in-one approaches. Degradation-specific methods perform strongly on individual degradations within their narrow training distributions but fail to generalize. For instance, BasicVSR++ (Chan et al., 2022a) introduces enhanced propagation and deformable alignment for video super-resolution, while VRT (Liang et al., 2024) leverages transformers to model long-range temporal dependencies. More recently, diffusion-based methods have yielded perceptually compelling results: Vivid-VR (Bai et al., 2026) distills generative priors from text-to-video models for photorealistic restoration, and FlashVSR (Zhuang et al., 2025) improves efficiency via a single-step diffusion design. However, these methods remain confined to their training distributions. Among all-in-one approaches, AverNet (Zhao et al., 2024) handles multiple unknown degradations within a single model, and MoCE-IR (Zamfir et al., 2025) adopts mixture-of-complexity-experts for adaptive capacity allocation. Yet a unified static architecture remains fundamentally inflexible for the compound, heterogeneous degradations of UGC videos, motivating agentic restoration systems.

2.2 Restoration Agent

Agentic restoration systems offer greater flexibility by leveraging VLMs to dynamically analyze degradations and schedule restoration tools. RestoreAgent (Chen et al., 2024) pioneers this direction for image restoration by composing toolchains from degradation descriptions. Q-Agent (Zhou et al., 2025) improves restoration planning by searching over toolbox orderings and evaluating them via IQA scoring. AgenticIR (Zhu et al., 2025) introduces a memory for pipeline prediction from historical knowledge and a rollback mechanism to handle tool failures. 4KAgent (Zuo et al., 2025) follows a similar scheduling paradigm targeting 4K super-resolution. MoA-VR (Liu et al., 2025) extends this paradigm to video through a three-agent framework that coordinates degradation identification,

pipeline planning, and quality assessment. Recently, VQ-Jarvis (Zhang et al., 2026) proposes a retrieval-augmented video restoration agent with hierarchical scheduling, where optimal pipelines for easy cases are retrieved directly from a static memory based on the initial input, while harder cases are resolved by greedy search. Although these agentic systems have achieved substantial gains in restoration performance, they still struggle in complex UGC video scenarios due to inaccurate retrieval and open-loop pipeline planning.

3 METHOD

Algorithm 1: Overall pipeline of ReVRA

Components: Planning VLM ϕ, Evaluation VLM ψ, Instance-aware Evolving Memory $\mathcal{R}$, toolboxes $\{\boldsymbol{\mathcal{F}}_i\}_{i=1}^{M}$, maximum number of stages L
Input: Degraded video $\boldsymbol{V}^0$
Output: Restored video $\boldsymbol{V}$

for $l = 0, 1, \ldots, L$ **do**
  Extract video embedding of current state: $\boldsymbol{e} \leftarrow \phi$.encode$(\boldsymbol{V}^l)$;
  Retrieve the best-matching toolbox from $\mathcal{R}$ as planning prior: $\boldsymbol{\mathcal{F}}_r \leftarrow$ RETRIEVE$(\mathcal{R}, \boldsymbol{e})$;
  Plan the candidate toolbox set: $\boldsymbol{\mathcal{C}} \leftarrow \phi(\boldsymbol{V}^l, \boldsymbol{\mathcal{F}}_r)$;
  if $\boldsymbol{\mathcal{C}} = \emptyset$ **then break**;
  Initialize the optimal result and corresponding toolbox: $\boldsymbol{V}^* \leftarrow \boldsymbol{V}^l$ and $\boldsymbol{\mathcal{F}}^* \leftarrow \emptyset$;
  foreach $\boldsymbol{\mathcal{F}}_i \in \boldsymbol{\mathcal{C}}$ **do**
    foreach $f \in \boldsymbol{\mathcal{F}}_i$ **do**
      if $\max_\psi(f(\boldsymbol{V}^l), \boldsymbol{V}^*) = f(\boldsymbol{V}^l)$ **then**
        Update the optimal result: $\boldsymbol{V}^* \leftarrow f(\boldsymbol{V}^l)$;
        Update the optimal toolbox: $\boldsymbol{\mathcal{F}}^* \leftarrow \boldsymbol{\mathcal{F}}_i$
  if $\boldsymbol{\mathcal{F}}^* \neq \boldsymbol{\mathcal{F}}_r$ **then**
    Update the memory $\mathcal{R}$;
  if $\boldsymbol{\mathcal{F}}^* = \emptyset$ **then break**;
  Advance to next state: $\boldsymbol{V}^{l+1} \leftarrow \boldsymbol{V}^*$;
Update restored video: $\boldsymbol{V} = \boldsymbol{V}^l$;
return $\boldsymbol{V}$

3.1 OVERALL FRAMEWORK

As illustrated in Fig. 1 and formalized in Algorithm 1, ReVRA organizes restoration as an iterative Plan-Execute-Evaluate loop. At each stage l, Planning is first performed: Planning VLM ϕ retrieves the most relevant toolbox $\boldsymbol{\mathcal{F}}_r$ from the IEM $\mathcal{R}$ based on the video embedding $\boldsymbol{e}$ of current video $\boldsymbol{V}^l$, where $\boldsymbol{e}$ is extracted by the visual encoder of ϕ, and uses it as a prior to plan a candidate toolbox set $\boldsymbol{\mathcal{C}}$. Subsequently, the agent executes all tools in the candidate toolboxes in sequence. Evaluation is then performed, where the outputs produced by these tools, together with the original $\boldsymbol{V}^l$, are compared pairwise by the evaluation VLM ψ to select the best result. The selected result $\boldsymbol{V}^*$ is used both as the output of the current stage and as the input to the next stage, while the toolbox $\boldsymbol{\mathcal{F}}_i$ to which the best-performing tool belongs is recorded as optimal toolbox $\boldsymbol{\mathcal{F}}^*$. If $\boldsymbol{\mathcal{F}}^* \neq \boldsymbol{\mathcal{F}}_r$, the IEM $\mathcal{R}$ is updated to correct the experience bias. The retrieval and update mechanisms of the IEM are detailed in Section 3.2. The pipeline terminates when $\boldsymbol{\mathcal{F}}^* = \emptyset$, indicating that the skip operation yields the best quality, or when the candidate toolbox set $\boldsymbol{\mathcal{C}}$ is empty. The output video of the current stage is then returned as the final restoration result.

3.2 INSTANCE-AWARE EVOLVING MEMORY

To enhance planning precision, we propose an Instance-aware Evolving Memory $\mathcal{R} = \{\boldsymbol{p}_i\}_{i=1}^{M+1}$, which maintains $M + 1$ prototypes corresponding to the M toolboxes and one skip operation. Each prototype $\boldsymbol{p}_i$ is the mean video embedding of all historical videos associated with $\boldsymbol{\mathcal{F}}_i$, where n is the number of associated records, and $\boldsymbol{e}$ is the video embedding. Compared to coarse degradation labels, such visual embeddings capture instance-level features that enable more accurate retrieval. At each planning step, $\boldsymbol{e}$ is matched against all prototypes, and the name of the nearest toolbox $\boldsymbol{\mathcal{F}}_r$ is injected into the VLM prompt as a planning prior. After execution, if $\boldsymbol{\mathcal{F}}_r$ mismatches the optimal toolbox $\boldsymbol{\mathcal{F}}^*$, the corresponding prototype is updated via an incremental mean update. As more inferences accumulate, $\mathcal{R}$ progressively self-evolves to supply increasingly accurate planning priors.

3.3 Comparison-based Visual Quality Assessment

The evaluation VLM ψ adopts a pairwise comparison strategy to select the highest-quality video from a set of candidates. Given two candidate videos $\boldsymbol{V}_a$ and $\boldsymbol{V}_b$, we define the comparison operator $\max_\psi(\boldsymbol{V}_a, \boldsymbol{V}_b)$ as returning the one with superior visual quality. We implement ψ with a two-stage prompt: it first localizes perceptual differences between $\boldsymbol{V}_a$ and $\boldsymbol{V}_b$, such as texture degradation, ringing artifacts, and structural distortions, and then makes an overall preference judgment. This design follows human perceptual assessment and is more sensitive to localized restoration artifacts than aggregate VQA/IQA scores. To further align ψ with human aesthetic preferences, we fine-tune it via direct preference optimization (DPO) (Rafailov et al., 2023). Specifically, the training data consists of video pairs $\mathcal{D} = \{(\boldsymbol{x}_i, y_i^+, y_i^-)\}_{i=1}^N$ annotated with human preferences. Here, $\boldsymbol{x}_i$ denotes the video input and the text prompts, y_i^+ denotes the response consistent with human preference, and y_i^- denotes the response inconsistent with human preference. In terms of the training strategy, we adopt LoRA for parameter-efficient fine-tuning. Finally, the optimization objective can be formulated as follows:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(\boldsymbol{x},y^+,y^-)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\frac{\log\pi_\theta(y^+|\boldsymbol{x})}{\log\pi_{ref}(y^+|\boldsymbol{x})} - \beta\frac{\log\pi_\theta(y^-|\boldsymbol{x})}{\log\pi_{ref}(y^-|\boldsymbol{x})}\Big)\Big], \tag{1}$$

where π_θ denotes the policy model to be optimized, π_{ref} denotes the frozen reference model, $\sigma(\cdot)$ is the sigmoid function, and β is a temperature parameter which is set to 0.1 during training.

3.4 Construction of a UGC Video Restoration Benchmark

Existing video restoration benchmarks predominantly construct low-quality inputs via synthetic degradation pipelines, failing to capture the complexity of real-world UGC content. While real-world test sets such as VideoLQ (Chan et al., 2022c) and CRVD (Yue et al., 2020) exist, they are limited in scope: CRVD covers only noise degradation, and VideoLQ is biased toward super-resolution with few videos. To this end, we introduce **TaoUGC**, a real-world UGC video restoration benchmark collected from authentic short-video platforms. TaoUGC comprises 150 clips spanning resolutions from 720P to 4K, covering four common degradation categories — compression artifacts, blur, noise, and low resolution — with most videos exhibiting compound degradations that better reflect the entangled nature of real-world UGC content. Videos are sampled across diverse content categories including natural scenes, human activities, and AIGC content, providing a more rigorous and realistic testbed for UGC video restoration.

4 Experiments

4.1 Experimental Setting

Implementation Details: ReVRA targets four representative UGC degradation types: compression artifacts, blur, low resolution, and noise, with corresponding degradation-specific restoration toolboxes $\{\boldsymbol{\mathcal{F}}_i\}_{i=1}^M, M = 4$. The IEM is reset for each dataset. All training and inference are conducted on two NVIDIA L40 GPUs. See the Supplementary Material for prompt and hyperparameter details.

Toolbox: The restoration toolbox is organized as follows: denoising toolbox includes BasicVSR++ (Chan et al., 2022a), AverNet (Zhao et al., 2024), and VRT (Liang et al., 2024); deblurring toolbox includes BasicVSR++, DSTNet+ (Pan et al., 2025), and VRT; compression artifact reduction toolbox includes BasicVSR++ and Zheng et al. (Li et al., 2026); super-resolution toolbox includes VRT, Vivid-VR (Bai et al., 2026), FlashVSR (Zhuang et al., 2025), and Zheng et al. (Li et al., 2026).

Training Data for ψ: We adopt Qwen3-VL-8B-Instruct(Bai et al., 2025) as both planning VLM ϕ and as the base model for evaluation VLM ψ. To construct the preference dataset, we synthesize degraded videos from the YouHQ training set (Zhou et al., 2024) using the four degradation types above and restore them with the corresponding tools in the toolbox, excluding the skip operation. In total, nearly 6,000 video pairs are obtained. The resulting video pairs are annotated by human subjects through pairwise preference comparison, and labels are aggregated to obtain the final preference annotations. We then use GPT-5 (Singh et al., 2025) to generate rationales for each annotated pair, forming the final dataset for DPO fine-tuning of ψ.

Test Datasets: We evaluate ReVRA on seven datasets spanning synthetic and real-world scenarios. For standard-resolution evaluation, we use KVQ (Lu et al., 2024) and TaoUGC (real-world compound-degradation), GoPro (Nah et al., 2017) and DVD (Su et al., 2017) (synthetic blur and compression), and CRVD (Yue et al., 2020) (real-world noise). For low-resolution, we use YouHQ40 (Zhou et al., 2024) (synthetic, 240P) and VideoLQ (Chan et al., 2022c) (real-world, 240P).

Table 1: Quantitative comparison with state-of-the-art methods on the standard-resolution benchmark. The best and second performances are marked in red and blue, respectively.

Datasets	Metrics	BasicVSR++	MoCE-IR	AverNet	AgenticIR	4KAgent	Ours
TaoUGC	NIQE↓	4.847	5.358	5.517	5.436	5.373	5.135
	CLIPIQA↑	0.485	0.376	0.394	0.374	0.394	0.490
	MUSIQ↑	61.51	55.66	54.35	55.76	57.09	63.76
	DOVER↑	0.818	0.753	0.753	0.747	0.758	0.818
KVQ	NIQE↓	5.686	6.054	6.210	6.182	5.954	6.033
	CLIPIQA↑	0.498	0.371	0.397	0.379	0.403	0.522
	MUSIQ↑	59.62	54.63	54.24	55.99	57.41	63.74
	DOVER↑	0.668	0.576	0.577	0.577	0.599	0.697
Gopro	PSNR↑	24.11	29.54	23.93	25.52	26.36	27.27
	NIQE↓	5.448	5.459	5.786	5.088	4.931	4.697
	CLIPIQA↑	0.198	0.227	0.197	0.255	0.256	0.286
	MUSIQ↑	29.81	38.18	25.36	43.82	46.01	47.26
	DOVER↑	0.536	0.570	0.407	0.650	0.660	0.679
	tOF↓	1.518	0.587	1.984	0.526	0.388	0.461
DVD	PSNR↑	27.67	30.08	28.56	30.48	29.86	31.36
	NIQE↓	4.737	4.602	4.895	4.557	4.597	4.167
	CLIPIQA↑	0.278	0.265	0.236	0.289	0.297	0.391
	MUSIQ↑	44.50	43.77	34.59	44.95	45.28	56.10
	DOVER↑	0.760	0.664	0.574	0.680	0.681	0.734
	tOF↓	0.371	0.233	0.322	0.216	0.234	0.196
CRVD	NIQE↓	4.252	4.992	4.360	5.907	5.350	3.407
	CLIPIQA↑	0.318	0.243	0.330	0.296	0.325	0.485
	MUSIQ↑	30.30	24.45	33.40	34.87	35.31	55.70
	DOVER↑	0.490	0.311	0.490	0.552	0.483	0.736

Evaluation Metrics: To better reflect real-world UGC video scenarios, we adopt a set of widely used no-reference VQA/IQA metrics for evaluation, including DOVER(Wu et al., 2023), NIQE(Mittal et al., 2012b), CLIP-IQA(Wang et al., 2023), and MUSIQ(Ke et al., 2021). For the GoPro and DVD datasets where ground-truth references are available, we additionally evaluate reference-based metrics, with PSNR and tOF(Chu et al., 2020) reported in the main paper and the remaining results provided in the supplementary material. Among these metrics, DOVER and tOF are VQA metrics that measure overall video quality and temporal consistency, respectively; NIQE, CLIP-IQA, and MUSIQ are IQA metrics assessing perceptual quality; and PSNR quantifies pixel-level fidelity. IQA scores are averaged over all frames.

4.2 Results on Standard-Resolution Scenarios

We compare ReVRA against representative methods spanning three categories: the degradation-specific method BasicVSR++ (Chan et al., 2022a), all-in-one methods MoCE-IR (Zamfir et al., 2025) and AverNet (Zhao et al., 2024), and agent-based methods AgenticIR (Zhu et al., 2025) and 4KAgent (Zuo et al., 2025). Among them, BasicVSR++ adopts the variant designed for same-resolution enhancement(Chan et al., 2022b). Since all-in-one methods lack super-resolution capability, we restrict this comparison to standard-resolution benchmarks to ensure a fair evaluation. Experiments are conducted on the real-world degradation benchmarks TaoUGC and CRVD, and the synthetic degradation benchmarks DVD, GoPro and KVQ.

Quantitative results are reported in Table 1. ReVRA achieves the best performance across all five benchmarks on most metrics. On the UGC benchmarks TaoUGC and KVQ, where degradation types are complex and tightly coupled, ReVRA significantly outperforms existing agent-based methods, highlighting the advantage of recursive state-conditioned planning over static pipelines that commit to a fixed plan without intermediate feedback. On the other degradation benchmarks (GoPro, DVD, CRVD), ReVRA also achieves top performance, even when only one degradation type is present, the closed-loop design avoids over-processing artifacts that fixed pipelines may introduce. Qualitative comparisons in Fig. 2 further show that ReVRA recovers finer texture details and suppresses residual artifacts more effectively across both real-world and synthetic scenarios. See the supplementary material for more comparisons.

Table 2: Quantitative comparison with state-of-the-art methods on the low-resolution benchmark. The best and second performances are marked in red and blue, respectively.

Datasets	Metrics	Non-agent super-resolution methods					Agent-driven restoration methods			
		BasicVSR++	SeedVR	SeedVR2	FlashVSR	Vivid-VR	AgenticIR	4KAgent	VQ-Jarvis	Ours
YouHQ40	NIQE↓	9.431	4.249	4.576	3.320	3.410	3.969	3.397	-	3.460
	CLIPIQA↑	0.294	0.362	0.336	0.591	0.447	0.460	0.480	0.542	0.596
	MUSIQ↑	22.17	61.52	59.34	70.64	70.55	62.43	71.29	67.79	70.93
	DOVER↑	0.275	0.886	0.872	0.875	0.901	0.857	0.863	0.572	0.886
VideoLQ	NIQE↓	6.161	4.900	5.593	3.822	4.217	4.806	4.290	-	3.890
	CLIPIQA↑	0.258	0.233	0.223	0.443	0.354	0.350	0.374	-	0.449
	MUSIQ↑	26.12	46.68	43.30	56.56	63.93	61.22	60.58	-	59.30
	DOVER↑	0.426	0.678	0.668	0.776	0.793	0.490	0.453	-	0.795

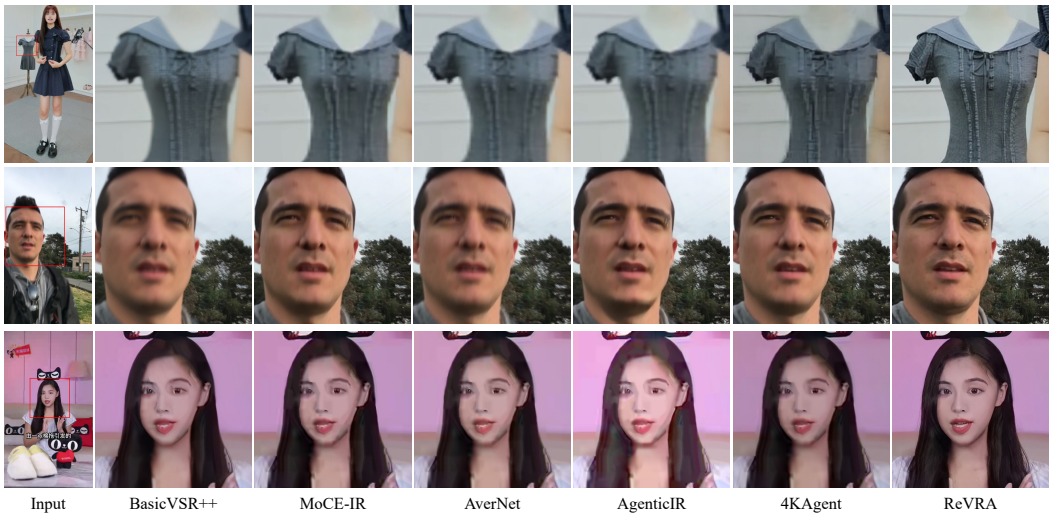

Figure 2: Qualitative comparison results on TaoUGC (first row), DVD (second row), and KVQ (third row) benchmarks. The proposed ReVRA produces the frames with better visual quality.

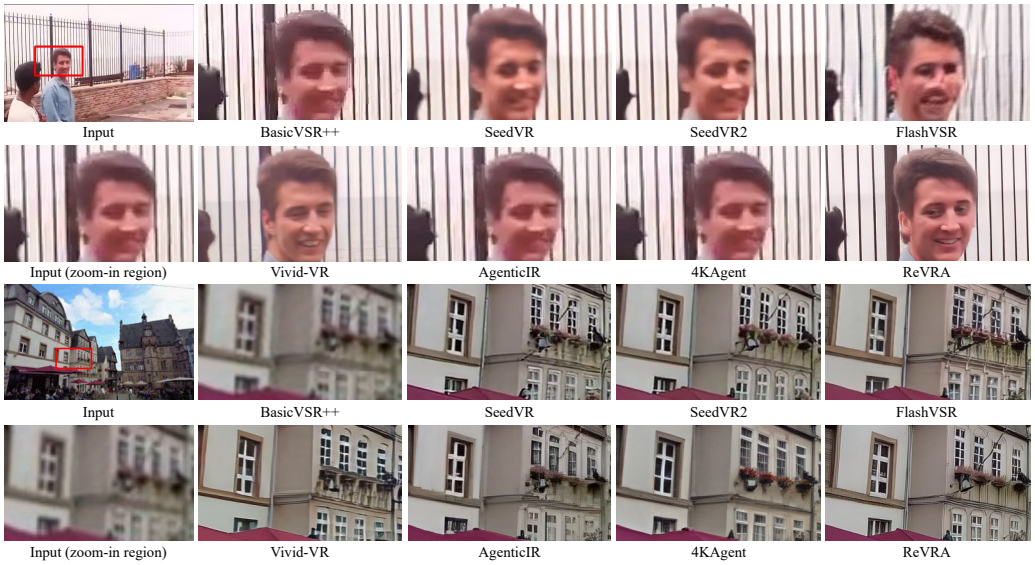

Figure 3: Qualitative comparison results on low-resolution benchmark VideoLQ (top) and YouHQ40 (bottom). The proposed ReVRA produces the frames with better visual quality.

4.3 Results on Low-Resolution Scenarios

To assess restoration capability under compound degradation with low resolution, we further evaluate ReVRA on two super-resolution benchmarks: synthetic YouHQ40 and real-world VideoLQ. We

Table 3: Performance and efficiency comparison with existing agents on the GoPro dataset. Asterisks (*) denote results under the aligned tool pool; unmarked entries use original tool configurations.

	PSNR↑	NIQE↓	MUSIQ↑	CLIPIQA↑	DOVER↑	tOF↓	Inference Time↓	Tool calls↓
AgenticIR*	26.94	5.284	37.34	0.216	0.588	0.932	524s	5.2
AgenticIR	25.52	5.088	43.82	0.255	0.650	0.526	409s	4.2
4KAgent*	27.79	5.022	40.85	0.248	0.596	0.676	583s	2.8
4KAgent	26.36	4.931	46.01	0.256	0.660	0.388	1342s	6.2
Ours	27.27	4.697	47.26	0.286	0.679	0.461	769s	3.8

compare against a broader set of baselines, including the discriminative model BasicVSR++ (Chan et al., 2022a), the diffusion-based methods SeedVR (Wang et al., 2025), SeedVR2 (Wang et al., 2026), FlashVSR (Zhuang et al., 2025), Vivid-VR (Bai et al., 2026), and agent-based methods AgenticIR (Zhu et al., 2025), 4KAgent (Zuo et al., 2025), and VQ-Jarvis (Zhang et al., 2026).

Quantitative results are reported in Table 2. ReVRA achieves leading performance on the majority of metrics across both benchmarks. On the YouHQ40 dataset, ReVRA outperforms the latest video restoration agent, VQ-Jarvis, on all metrics. On the real-world VideoLQ dataset, ReVRA surpasses all restoration agents and specialized video super-resolution methods on multiple metrics. These results highlight that ReVRA's iterative re-planning is particularly effective in low-resolution scenarios. Qualitative comparisons in Figure 3 further illustrate that ReVRA achieves superior detail reconstruction on both real-world and synthetic low-resolution videos, recovering finer textures and sharper structures that competing methods fail to restore.

4.4 Controlled Comparison with Existing Agents

To accurately assess the effectiveness of the ReVRA pipeline, we compare its performance and efficiency with existing agent-based methods under aligned tool settings. Results are presented in Table 3, where *Inference Time* and *Tool Calls* denote the per-video inference latency and the number of tool invocations, respectively. As can be seen, compared with the baselines under aligned tool sets (Rows 1 and 4), ReVRA achieves superior overall performance with comparable complexity. When the baselines adopt our tool set, none of the metrics improve except PSNR. These results demonstrate that ReVRA's performance gains stem from the proposed dynamic, self-corrective scheduling strategy rather than the adoption of a more powerful tool set.

4.5 Ablation Studies

Reflective Planning vs. Static Pipeline: We validate the closed-loop re-planning design by replacing it with two open-loop baselines: a fixed pipeline (Zuo et al., 2025; Zhu et al., 2025) that selects the toolboxes to be invoked from the input but uses a predefined execution order, and a static pipeline (Zhang et al., 2026; Jiang et al., 2026) that determines both toolboxes and order solely from the input. We compare these settings on the VideoLQ dataset, with the results shown in Table 4. ReVRA consistently outperforms both baselines across the major metrics, confirming that state-conditioned re-planning is essential for compound degradation scenarios.

Instance-aware Evolving Memory Assessment: At each stage, the planning VLM determines the candidate toolboxes by consulting the IEM for more precise decisions. We conduct the following ablation studies on TaoUGC to assess the effectiveness of the planning VLM and our memory mechanism: (1) both disabled, i.e., brute-force search over all toolboxes; (2) memory disabled; (3) planning VLM disabled; (4) static memory, i.e., not updated during inference; (5) category-aware memory, i.e., retrieving by degradation category (Liu et al., 2025; Zhu et al., 2025; Zuo et al., 2025); and (6) our IEM. To control against the static setting, all IEM variants are pre-initialized on KVQ, GoPro, and VideoLQ. Results are reported in Table 6, where *Inference Time* and *Tool Calls* denote per-video inference latency and the number of tool invocations, respectively. Compared with the memory-free setting (Row 2), enabling the IEM (Row 6) improves restoration quality across all metrics with fewer tool calls, confirming that the memory effectively helps the planning VLM select more suitable candidate toolboxes. Compared with static (Row 4) and category-aware (Row 5) memory, the IEM consistently outperforms across all metrics, validating that (i) the dynamic update mechanism continuously refines the planning prior, and (ii) embedding-based retrieval captures video features more precisely. Using the IEM alone (Row 3) yields poor results, as the memory prior requires VLM interpretation and cannot be directly applied without validation. Finally, brute-force search (Row 1) achieves the best quality scores but incurs prohibitive computational cost. Since ReVRA selects outputs via pairwise quality comparison, we conducted a user study with 15 judges

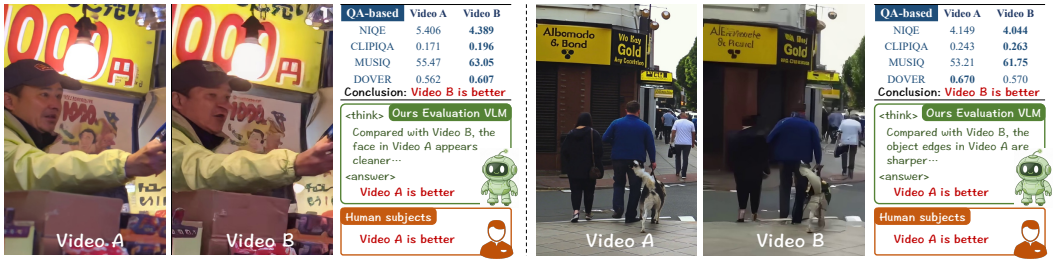

Figure 4: Comparison between the selection results of VQA-based method and our evaluation VLM.

(see the supplementary material for details), where brute-force and ReVRA were preferred 53% and 50% of the time respectively, with 3% ties, indicating a negligible perceptual difference. Together, these results demonstrate IEM's effectiveness, enabling ReVRA to match brute-force search in perceptual quality with 8.7× fewer tool calls and 4.6× less time.

Table 4: Ablation studies on reflective planning scheme. Fixed: predefined toolbox execution order; Static: toolboxes and order both determined from input; Ours: re-planned at each stage.

Pipeline	NIQE↓	CLIPIQA↑	MUSIQ↑	DOVER↑
Fixed	3.938	0.456	58.27	0.769
Static	4.098	0.458	59.25	0.752
Ours	3.890	0.449	59.30	0.795

Table 5: Comparison of video selection accuracy, measured as the percentage of cases where each method's selection matches human subject preference. 'QA-based' refers to selection via averaged VQA and IQA scores.

Dataset	Qwen3-VL	Gemini3.1-pro	QA-based	Ours
KVQ	62.06%	64.25%	92.51%	97.84%
VideoLQ	63.91%	62.96%	64.40%	74.36%

Comparison-based Visual Quality Assessment: ReVRA employs the evaluation VLM ψ with a pairwise comparison scheme to select the best result from candidate videos, thereby determining which tool to invoke at each stage. To evaluate this design, we compare ψ against the baseline VLM Qwen3-VL-8B-Instruct, the proprietary VLM Gemini3.1-Pro, and the QA-based method from (Zhou et al., 2025; Zuo et al., 2025), which selects based on average VQA/IQA scores. Experiments are conducted on KVQ and VideoLQ, where each video is processed by different pipelines to produce multiple versions. In KVQ, the differences lie mainly in overall visual quality, while in VideoLQ they arise from temporal inconsistencies and erroneous frame textures. Human subjects provide preference annotations as ground truth. See the supplementary material for details. As reported in Table 5, our evaluation VLM achieves the highest judgment accuracy on both datasets, surpassing the baseline by 23%. Compared with the QA-based approach, it improves markedly on VideoLQ, indicating that pairwise comparison better identifies perceptual distortions. Fig. 4 presents representative cases: Video B exhibits severe degradation in facial details and edge textures, respectively. Such issues are hard to capture with QA-based methods, whereas our evaluation VLM identifies them accurately, yielding judgments consistent with human subjects.

5 Conclusion

In this paper, we propose ReVRA, a **Re**flective **V**ideo **R**estoration **A**gent for real-world UGC scenarios. ReVRA formulates complex restoration as a multi-stage *plan–execute–evaluate* process. At each stage, it generates candidate tools conditioned on the current video state and selects the best one through pairwise evaluation, allowing adaptive planning and closed-loop self-correction. ReVRA further incorporates a dynamically updated, embedding-indexed memory for precise retrieval of relevant historical priors. Experiments on seven benchmarks covering diverse scenarios demonstrate that ReVRA achieves state-of-the-art restoration performance and underscores the potential of agentic systems for real-world video restoration.

Table 6: Ablation studies on the planning VLM and memory mechanism on TaoUGC dataset.

Planning VLM	Memory	Tool calls↓	Inference Time↓	NIQE↓	MUSIQ↑	CLIPIQA↑	DOVER↑
✗	✗	26.52	2926s	4.525	70.66	0.601	0.852
✓	✗	3.21	671s	5.135	63.76	0.490	0.818
✗	IEM	3.16	633s	5.221	63.65	0.482	0.816
✓	Static	3.06	645s	5.171	63.86	0.492	0.820
✓	Category	3.07	647s	5.198	63.76	0.489	0.815
✓	IEM	3.04	642s	5.055	64.86	0.500	0.826

AI USE STATEMENT

We used generative AI tools for language translation and polishing, as well as for drafting parts of the experimental code. We did not use them to generate synthetic datasets, propose or refine hypotheses, design or evaluate the research methodology or experiments, implement the core methods, conduct qualitative or thematic analysis, or interpret the results. Applications such as developing theoretical models, formulating or proving mathematical claims, and cleaning or reformatting datasets are not applicable to this work. All AI-assisted output was reviewed by the authors: the translated and polished text was checked for accuracy and clarity, and the AI-generated code was tested to verify its correctness. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Haoran Bai, Xiaoxu Chen, Canqian Yang, Zongyao He, Sibin Deng, and Ying Chen. Vivid-vr: Distilling concepts from text-to-video diffusion transformer for photorealistic video restoration. In *International Conference on Learning Representations (ICLR)*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5972–5981, 2022a.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. On the generalization of basicvsr++ to video deblurring and denoising. *arXiv preprint arXiv:2204.05308*, 2022b.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5962–5971, 2022c.

Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. Restoreagent: Autonomous image restoration agent via multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110643–110666, 2024.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75–1, 2020.

Muhammad Kamran Janjua, Amirhosein Ghasemabadi, Kunlin Zhang, Mohammad Salameh, Chao Gao, and Di Niu. Grounding degradations in natural language for all-in-one video restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5734–5743, 2026.

Xu Jiang, Gehui Li, Bin Chen, and Jian Zhang. Multi-agent image restoration. *International Journal of Computer Vision*, 134(5):205, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Xin Li, Jiachao Gong, Xijun Wang, Shiyao Xiong, Bingchen Li, Suhang Yao, Chao Zhou, Zhibo Chen, Radu Timofte, Yuxiang Chen, et al. Ntire 2026 challenge on short-form ugc video restoration in the wild with generative models: Datasets, methods and results. *arXiv preprint arXiv:2604.10551*, 2026.

Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 33: 2171–2182, 2024.

Lu Liu, Chunlei Cai, Shaocheng Shen, Jianfeng Liang, Weimin Ouyang, Tianxiao Ye, Jian Mao, Huiyu Duan, Jiangchao Yao, Xiaoyun Zhang, et al. Moa-vr: A mixture-of-agents system towards all-in-one video restoration. *IEEE Journal of Selected Topics in Signal Processing*, 2025.

Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25963–25973, 2024.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012b.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3883–3891, 2017.

Jinshan Pan, Long Sun, Boming Xu, Jiangxin Dong, and Jinhui Tang. Learning efficient deep discriminative spatial and temporal networks for video deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1279–1288, 2017.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2161–2172, 2025.

Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, Xuefeng Xiao, Chen Change Loy, and Lu Jiang. Seedvr2: One-step video restoration via diffusion adversarial post-training. In *ICLR*, 2026.

Wen Wen, Tianwu Zhi, Kanglong Fan, Yang Li, Xinge Peng, Yabin Zhang, Yiting Liao, Junlin Li, et al. Self-evolving vision-language models for image quality assessment via voting and ranking. In *International Conference on Learning Representations*, volume 2026, pp. 39444–39457, 2026.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20144–20154, 2023.

Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17714–17724, 2022.

Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *European conference on computer vision*, pp. 224–242. Springer, 2024.

Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2301–2310, 2020.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12753–12763, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

Xuanyu Zhang, Weiqi Li, Qunliang Xing, Jingfen Xie, Bin Chen, Junlin Li, Li Zhang, Jian Zhang, and Shijie Zhao. Vq-jarvis: Retrieval-augmented video restoration agent with sharp vision and fast thought. *arXiv preprint arXiv:2603.22998*, 2026.

Haiyu Zhao, Lei Tian, Xinyan Xiao, Peng Hu, Yuanbiao Gou, and Xi Peng. Avernet: All-in-one video restoration for time-varying unknown degradations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 127296–127316, 2024.

Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2535–2545, 2024.

Yingjie Zhou, Jiezhang Cao, Zicheng Zhang, Farong Wen, Yanwei Jiang, Jun Jia, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Q-agent: quality-driven chain-of-thought image restoration agent through robust multimodal large language model. *arXiv preprint arXiv:2504.07148*, 2025.

Kaiwen Zhu, Jinjin Gu, Zhiyuan You, Yu Qiao, and Chao Dong. An intelligent agentic system for complex image restoration problems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3RLxccFPHz.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747*, 2025.

Yushen Zuo, Qi Zheng, Mingyang Wu, Xinrui Jiang, Renjie Li, Jian Wang, Yide Zhang, Gengchen Mai, Lihong V. Wang, James Zou, Xiaoyu Wang, Ming-Hsuan Yang, and Zhengzhong Tu. 4kagent: Agentic any image to 4k super-resolution. In *Advances in Neural Information Processing Systems*, 2025.

A APPENDIX

A.1 USER PROMPTS

In the proposed ReVRA, both the planning and evaluation processes are driven by their corresponding VLMs. The prompts are shown in Fig. 5. Before generating a plan, the planning VLM retrieves the toolbox prototype most relevant to the current video from the instance-aware evolving memory and incorporates it into the prompt as prior knowledge.

Planning VLM ϕ

System prompt: You are a video restoration expert. Your task is to diagnose visual degradations in the input video. You will be provided with reference cases of similar videos and their degradation analysis results to help you make accurate judgments.
User Prompt: Analyze the above video for the following degradation types: Noise, Motion Blur, Compression Artifacts, and Low Resolution. For reference, the video shows the highest similarity to the standard profile of ***<nearest prototype t^r>***. INSTRUCTION: Treat the reference as a tentative suggestion only. Rely strictly on visual evidence from the video frames. If the video content disagrees with the reference, trust the video content. Based on the input video and the above references, output your analysis result in valid JSON format with boolean values (true/false). Do not include markdown code blocks.Required format:{ \"noise\": true/false,\n \"motion_blur\": true/false,\n \"compression_artifacts\": true/false,\n \"low_resolution\": true/false\n}.

Evaluation VLM ψ

System prompt: You are a video quality assessment expert. Your task is to compare two video sequences (Video A and Video B) and determine which one has better visual quality. Consider factors such as sharpness, clarity, noise level, motion blur, compression artifacts, and overall detail preservation.
User Prompt: Compare video A and video B. Determine which video has better overall visual quality. First, provide a brief analysis for each video. Then, conclude which one has better overall visual quality.Format your response as follows: Analysis: [Your analysis here] Conclusion: video A OR video B

Figure 5: Prompts for the planning and evaluation VLMs.

A.2 ADDITIONAL QUANTITATIVE AND QUALITATIVE COMPARISONS

In this section, we present additional quantitative comparisons in terms of reference-based metrics on datasets with ground-truth references (GoPro and DVD), as well as qualitative comparisons on the UGC video benchmarks KVQ (Lu et al., 2024) and the proposed TaoUGC. The quantitative results are reported in Table7. As can be seen, on DVD, ReVRA achieves the best result on all metrics; on GoPro, it ranks first on DISTS, second on PSNR and tOF, remaining highly competitive with methods specialized for this degradation. These results demonstrate that the details recovered by ReVRA are faithful to the ground truth.

The qualitative comparisons are shown in Fig. 6. ReVRA more effectively removes compression artifacts (first and third rows), better restores fine details and textures such as those around facial regions in portraits (second row), and yields overall clearer results (fourth row). Moreover, compared with existing agent-based methods such as AgenticIR (Zhu et al., 2025) and 4KAgent (Zuo et al., 2025), ReVRA produces more faithful colors without introducing over-sharpening artifacts.

Table 7: Additional quantitative comparisons on the GoPro and DVD datasets. The best and second performances are marked in red and blue, respectively.

datasets	metrics	Basicvsr++	MoCE-IR	Avernet	AgenticIR	4KAgent	Ours
GoPro	LPIPS↓	0.289	0.111	0.276	0.095	0.082	0.101
	DISTS↓	0.183	0.095	0.187	0.099	0.090	0.088
	PSNR↑	24.11	29.54	23.93	25.52	26.36	27.27
	SSIM↑	0.737	0.891	0.741	0.857	0.879	0.870
	tOF↓	1.518	0.587	1.984	0.526	0.388	0.461
DVD	LPIPS↓	0.190	0.119	0.176	0.106	0.104	0.072
	DISTS↓	0.118	0.065	0.103	0.063	0.063	0.050
	PSNR↑	27.67	30.08	28.56	30.48	29.86	31.36
	SSIM↑	0.828	0.888	0.848	0.900	0.901	0.922
	tOF↓	0.371	0.233	0.322	0.216	0.234	0.196

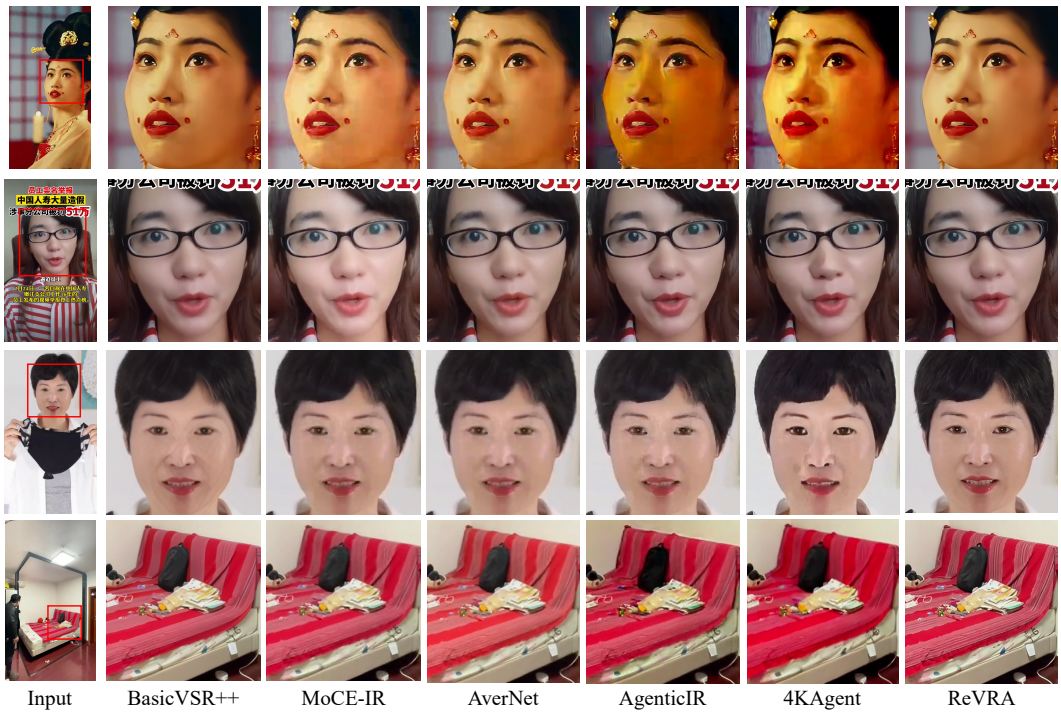

Figure 6: Additional qualitative comparisons on the UGC video benchmark, where the first two rows are from KVQ and the last two rows are from TaoUGC. The proposed ReVRA produces the frames with better visual quality. **Zoom-in for best view**

A.3 Visualization of Intermediate Results in the Pipeline

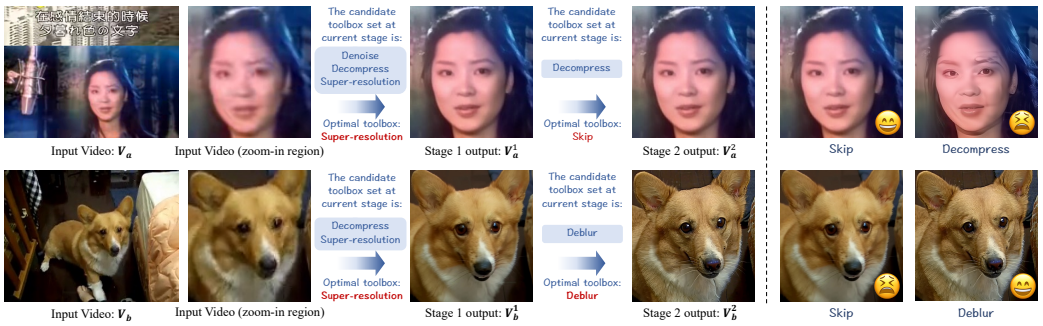

Figure 7: Visualization of intermediate results in the pipeline.

To illustrate how our dynamic pipeline is formulated, we visualize the intermediate results of the restoration process, as shown in Fig. 7. For Video V_a, in the first stage, the planning VLM generates three candidate toolboxes: denoising, decompression, and super-resolution. The evaluation VLM then selects super-resolution as the optimal toolbox. After the super-resolution is executed, the output of Stage 1 V_a^1 shows that the video no longer exhibits noticeable noise or low-resolution degradation. Consistent with this observation, the planning VLM in stage 2 reaches the same conclusion and updates the candidate toolbox set accordingly, dynamically removing denoising and super-resolution from the candidate toolbox set. Upon further applying the decompression toolbox, erroneous facial textures can be observed in the processed frames (right panel). After assessment by the evaluation VLM, the skip option is judged to yield better visual quality at this stage. As a result, the final output V_a^2 of stage 2 is the skipped version rather than the decompressed one. This result demonstrates

that the skip mechanism and the dynamic pipeline strategy together enable effective self-correction, thereby preventing the propagation of errors.

For Video $\boldsymbol{V}_b$, in stage 1, the planning VLM identifies compression artifacts and low resolution as the most prominent degradations and accordingly formulates the corresponding candidate toolboxes. After the super-resolution is performed, the Stage 1 output $\boldsymbol{V}_b^1$ shows that compression and low-resolution degradations have been removed, but a new blurring issue emerges. The planning VLM reaches the same conclusion and replans the candidate toolboxes set in stage 2, removing decompression and super-resolution while introducing a new deblurring toolbox. It can be observed that, after deblurring, the video (right panel) exhibits clearer textures and substantially better visual quality than the skip alternative, leading to the selection of deblur as the optimal toolbox for the second stage. Through iterative evaluation and replanning, our dynamic pipeline can effectively handle unexpected issues that arise during processing, thereby achieving superior restoration performance.

A.4 TaoUGC Benchmark Details

As shown in Table 8, we construct TaoUGC for evaluating video restoration in real-world UGC scenarios. TaoUGC is constructed via a systematic pipeline to ensure broad coverage across quality levels, content categories, and degradation types: (1) Quality Distribution: Thousands of candidate videos are sampled from short-video platforms and filtered using no-reference metrics (such as CLIP-IQA, DOVER) to span a diverse spectrum of visual quality. (2) Content Diversity: Typical UGC content categories are identified, and a two-stage process — combining automated content labeling with manual selection — ensures comprehensive scenario coverage. (3) Degradations: Manual inspection prioritizes samples exhibiting entangled, compound distortions representative of real-world UGC content. Among these, 20 high-quality clips are curated to further assess the generalization capability of restoration methods to high-quality scenarios.

The resulting 150 clips span 7 content categories, primarily including entertainment, animation, e-commerce, and video snippets. The dominant degradation types are compression artifacts (73%), noise (15%), blur (17%), and low-resolution (3%) introduced by post-processing operations such as interpolation and resizing; the majority of clips exhibit compound degradations. All clips are stored in MP4 format with $H.264$ or $H.265$ encoding. Together, the diverse content, broad quality distribution, and overlapping degradation types make TaoUGC a challenging benchmark for evaluating video restoration methods' generalizability under real-world distortions.

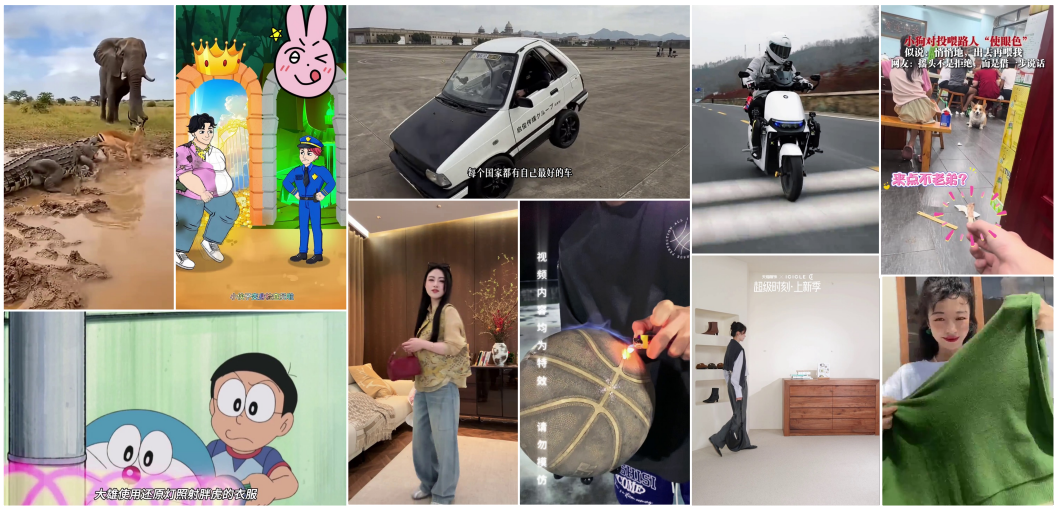

Figure 8: Representative frames from the TaoUGC benchmark.

Table 8: Summary of test datasets and the proposed TaoUGC benchmark.

Datasets	Degradation Source	Degradation Type	Resolution	Clips	Total Frames
GoPro	Synthetic	Blur, Compression	1280×720	11	1111
DVD	Synthetic	Blur, Compression	1280×720	10	1000
YouHQ40	Synthetic	Blur, Compression, Noise, Low-Resolution	256×256 – 455×256	40	1318
KVQ	Synthetic	Blur, Compression, Noise, Low-Resolution	270×480 – 2160×3840	122	28752
CRVD	Real-World	Noise	1920×1080	50	2500
VideoLQ	Real-World	Blur, Compression, Noise, Low-Resolution	320×240 – 720×576	50	4721
TaoUGC	Real-World	Blur, Compression, Noise, Low-Resolution	720×960 – 2160×3840	150	192427

A.5 Further Implementation Details

For the evaluation VLM, a core component of ReVRA, we use Qwen3-VL-8B-Instruct(Bai et al., 2025) as the base model and further align it with human preferences via DPO(Rafailov et al., 2023) fine-tuning. All training video pairs compare outputs restored by different tools from the same degraded video. To enable parameter-efficient adaptation, we adopt LoRA. Specifically, the LoRA rank, scaling factor, and dropout are set to 8, 16, and 0.05, respectively, and LoRA is applied to the projection modules in the last 6 layers of the language model. For DPO training, the preference optimization parameter β is set to 0.1. The model is trained for 3 epochs with an initial learning rate of 5×10^{-6}.

During inference, the max iterations L of the ReVRA restoration chain is set to 4. We evaluate ReVRA on seven datasets spanning synthetic and real-world scenarios. Among them, the KVQ(Lu et al., 2024) benchmark comprises 122 video scenes, each associated with seven compressed versions designed to simulate the processing pipeline of short-video platforms. We randomly sample one version from each scene for testing in this manuscript. For the CRVD(Yue et al., 2020) benchmark, we evaluate on the outdoor scenes without ground truth, as they provide a more realistic setting for real-world video restoration. For simplicity and without loss of generality, all benchmark results are obtained from the first 30 frames of each video.

A.6 Ablation Setup for Comparison-based Visual Quality Assessment

This section details the ablation study settings for the "Comparison-based Visual Quality Assessment" subsection in this manuscript. For the compared QA-based methods, we follow the evaluation protocol proposed in (Zhou et al., 2025; Zuo et al., 2025), where the average VQA/IQA score is used as the assessment criterion. Specifically, the relevant metrics include NIQE(Mittal et al., 2012b), CLIPIQA(Wang et al., 2023), MUSIQ(Ke et al., 2021) and DOVER(Wu et al., 2023), which are consistent with the performance evaluation metrics used in this manuscript. Since higher scores indicate better quality across all metrics except NIQE, the average score is computed as: $(\mathrm{CLIPIQA} + \mathrm{MUSIQ} + \mathrm{DOVER} - \mathrm{NIQE})/4$. Given a pair of videos, the one with the higher average score is regarded as having better quality.

We construct a human-preference test set based on the KVQ(Lu et al., 2024) and VideoLQ(Chan et al., 2022c) datasets. KVQ contains 122 scenes, each associated with seven videos degraded by compression and other distortions at different severity levels, along with human-annotated MOS scores. For each video pair, the video with the higher MOS score is treated as the human-preferred one. VideoLQ contains 50 low-quality videos from different scenes. To simulate potential restoration artifacts, such as erroneous textures and temporal inconsistency, we enhance these videos using multiple generative enhancement tools, including MGLD(Yang et al., 2024), UAV(Zhou et al., 2024), SeedVR(Wang et al., 2025), SeedVR2(Wang et al., 2026), and Vivid-VR(Bai et al., 2026). We then present pairs of enhanced videos from the same scene to human annotators, who select the one with the best subjective quality. The final human preference for each pair is determined by majority voting over the annotations from three subjects.

A.7 Annotation Details of the User Study

We recruited human subjects to perform subjective pairwise visual quality preference annotation, which serves two purposes: (1) constructing the training and test datasets for the evaluation VLM (see the "Comparison-based Visual Quality Assessment" subsection of the main manuscript), and (2) comparing the subjective restoration quality of brute-force search and ReVRA (see the "Instance-aware Evolving Memory Assessment" subsection of the main manuscript). All participants had normal visual acuity and color vision, and the evaluations were conducted under controlled viewing conditions. Each participant was presented with multiple pairs of restored videos, where the two videos in each pair share identical content but are produced by different restoration tools or pipelines. During evaluation, participants could freely switch between the two videos while they were playing and select the visually better one, or indicate that the two were comparable. To minimize bias, both the presentation order of video pairs and the left–right positions of the two videos within each pair were randomized.

For the first application, we recruited 12 participants divided into 4 groups. Participants were instructed to base their selection on their own visual preference, provided that no obvious quality artifacts (e.g., flickering or severe distortion) were observed in the preferred video. The final human preference label for each pair was determined by majority voting among the three subjects within each group. For the second application, we recruited 15 participants and aggregated annotations across all participants to obtain the statistical conclusion.

A.8 Limitations

ReVRA has two limitations. First, the multi-stage plan–execute–evaluate design requires executing and comparing all tools within each candidate toolbox at every stage. As the toolbox includes diffusion-based methods with substantial computational overhead, the overall inference time of ReVRA is non-trivial. Furthermore, the current toolbox composition has not been comprehensively explored; lighter-weight tools that achieve comparable restoration quality with lower inference cost may exist, and identifying such alternatives remains an open direction. Second, the quality of the final restored output depends critically on the judgment of the evaluation VLM ψ: an erroneous preference decision at any stage propagates to subsequent stages and ultimately degrades the final result. Although Table 5 demonstrates that the DPO-finetuned ψ aligns more closely with human preference than existing alternatives, its judgment accuracy has not yet reached optimal—particularly on challenging cases involving subtle temporal inconsistencies and fine-grained texture distortions. Further improving the reliability of ψ remains an important avenue for future work.