BeTAR:Belief-Guided Trajectory Learning with Agentic Reinforcement Learning for Multi-Degradation Surgical Image Restoration

Anonymous authors
Paper under double-blind review

Abstract

Surgical images are frequently affected by multiple coupled degradations, including smoke, low illumination, specular highlights, blur, water droplets, and low resolution, making fixed restoration pipelines difficult to generalize across complex clinical scenes. To address this challenge, we propose BeTAR, a degradation-aware agentic framework that integrates structured degradation perception with belief-guided restoration planning for multi-degradation surgical image restoration. First, a vision-language model is trained to identify degradation types and severity levels and organize them into an interpretable belief state, providing structured environmental information for subsequent decision making. Second, conditioned on this belief state, we formulate restoration planning under a Partially Observable Markov Decision Process (POMDP) framework and optimize an Agentic-RL policy to generate complete open-loop restoration trajectories, allowing the agent to jointly determine tool selection, execution order, and termination without relying on a predefined restoration chain. The policy is further optimized using a reference-free trajectory-level reward that combines no-reference image quality assessment with tool-use constraints, enabling training without paired clean images. Experiments on both in-distribution and independent surgical image test sets show that BeTAR achieves an F1 score of 0.908 for degradation recognition and ranks first on five of six restoration metrics on both test sets.

1 Introduction

Minimally invasive and robotic-assisted surgery relies heavily on high-quality endoscopic imaging to provide clear anatomical structures and reliable visual cues. However, surgical images acquired in real clinical environments are frequently degraded by smoke, insufficient illumination, specular highlights, motion blur, water droplets, and limited spatial resolution. More importantly, these degradations often coexist and interact with each other. In such cases, restoration performance depends not only on selecting appropriate tools, but also on determining their execution order. An inappropriate sequence may introduce redundant processing, amplify artifacts, or suppress useful visual information required by subsequent restoration operations. Therefore, adaptive tool trajectory is a key challenge for multi-degradation surgical image restoration.

Most existing surgical image restoration methods focus on individual degradation types, such as smoke removal, low-light enhancement, and specular reflection removal (Pan et al., 2022; Chen et al., 2024b; Li et al., 2024). Recent all-in-one approaches further address multiple degradations within a unified framework (Yang et al., 2024; Chen et al., 2026), but typically rely on a monolithic restoration process without explicitly modeling the dependencies and ordering among different restoration operations. Meanwhile, recent agent-based image restoration methods, including RestoreAgent, AgenticIR, JarvisIR, and Q-Agent, introduce vision-language models to perceive degradations and invoke specialized restoration tools (Chen et al., 2024a; Zhu et al., 2025; Lin et al., 2025; Zhou et al., 2025). However, most of these methods are designed for natural or outdoor images and pay limited attention to how multiple restoration tools should be ordered under surgery-specific

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.

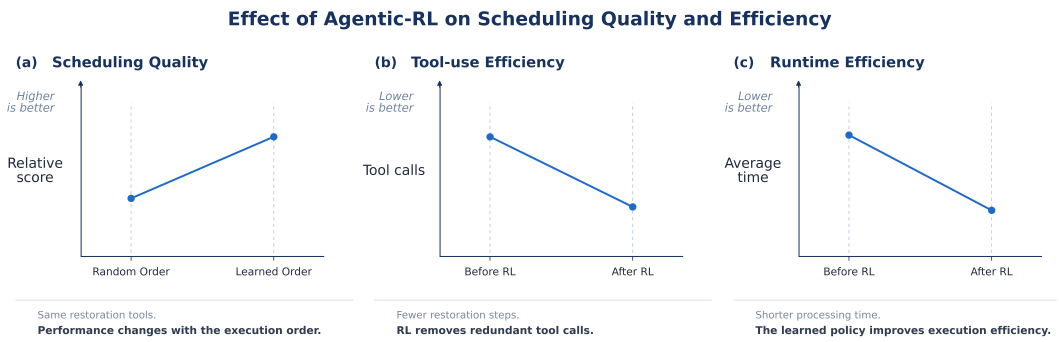

Figure 1: Effect of Agentic-RL on scheduling quality and efficiency. The learned policy improves restoration quality, reduces tool invocations, and lowers average processing time.

mixed degradations. As different operation sequences can lead to substantially different restoration outcomes, learning an effective restoration order remains a key challenge.

To address this problem, we propose a degradation-aware agentic framework that formulates surgical image restoration as belief-guided tool trajectory planning. We train Qwen2.5-VL-7B-Instruct with supervised fine-tuning (SFT) and Group Relative Policy Optimization (GRPO) (Wu et al., 2025a; Shao et al., 2024) to predict degradation types and severity levels, which are organized as an interpretable belief state. Conditioned on this belief, the agent generates a complete open-loop restoration trajectory under a POMDP formulation (Kaelbling et al., 1998), jointly determining tool selection, execution order, and termination. The trajectory policy is optimized through Agentic-RL with reference-free image-quality feedback, enabling the model to learn restoration sequences that improve final image quality.

We further conduct controlled experiments to examine the effect of tool scheduling. As shown in Figure 1,when the same restoration tools are used with different execution orders, the learned trajectory consistently produces better restoration quality than random ordering, showing that the gain comes not only from tool selection but also from tool composition. After Agentic-RL optimization, the agent also uses fewer restoration operations and requires less processing time, while maintaining better final restoration quality. In addition, comparisons among fixed tool usage, random tool selection, and VLM-guided autonomous selection further demonstrate the importance of degradation-aware planning for effective tool scheduling.

Our main contributions are summarized as follows: (1) We formulate multi-degradation surgical image restoration as a tool trajectory planning problem and introduce a POMDP-guided Agentic-RL framework that jointly optimizes tool selection, execution order, trajectory length, and termination. (2) We develop a structured VLM perception model through SFT and GRPO, whose degradation predictions are explicitly used as the belief state for downstream Agentic-RL planning. (3) We conduct controlled scheduling experiments showing that learned tool ordering improves restoration quality while reducing unnecessary tool invocations and processing cost.

2 Related Work

2.1 Surgical Image Restoration

Image degradation is common in endoscopic and minimally invasive surgery due to the complex imaging environment and surgical operations. Existing studies have developed dedicated restoration methods for individual degradation types. For example, DeSmoke-LAP employs unpaired image-to-image translation for laparoscopic smoke removal (Pan et al., 2022), while LighTDiff introduces a lightweight diffusion framework for surgical low-light enhancement (Chen et al., 2024b). EndoSRR addresses specular reflection through reflection detection and subsequent region inpainting (Li et al., 2024). These task-specific approaches achieve strong performance for their corresponding degradations, but require the degradation type to be known or implicitly fixed by the model.

Recent research has begun to move toward more general surgical restoration. SurgClean provides a benchmark covering multiple laparoscopic restoration tasks (Pei et al., 2026), while EndoIR

develops a degradation-agnostic all-in-one diffusion framework for multiple endoscopic degradations (Chen et al., 2026). Nevertheless, unified restoration models generally perform degradation handling within a single network and do not explicitly reason about the composition, severity, and processing order of coexisting degradations. In contrast, BeTAR separates degradation perception from restoration planning and dynamically composes specialized restoration tools according to the predicted degradation belief.

2.2 Agentic Image Restoration

Multimodal instruction tuning (Liu et al., 2023a) and language-model tool use, exemplified by ReAct (Yao et al., 2022) and Toolformer (Schick et al., 2023), provide foundations for autonomous visual agents. In image restoration, these capabilities enable systems to evolve from fixed neural mappings toward tool-based processing. RestoreAgent uses an MLLM to identify image degradations, determine restoration tasks, select appropriate models, and optimize their execution sequence (Chen et al., 2024a). AgenticIR further introduces perception, scheduling, execution, reflection, and rescheduling for complex real-world restoration (Zhu et al., 2025). JarvisIR employs a VLM-based controller for adverse-weather restoration in autonomous-driving scenarios (Lin et al., 2025), while Q-Agent incorporates degradation perception and image-quality-guided restoration planning (Zhou et al., 2025). More recently, HybridAgent integrates multiple restoration modes to improve the handling of mixed distortions (Li et al., 2026).

Despite their flexibility, existing restoration agents are predominantly trained and evaluated on natural or outdoor imagery, whose degradation distributions differ substantially from surgical scenes. Directly transferring such agents does not explicitly account for surgery-specific degradations or their clinical visual characteristics. BeTAR instead learns degradation perception from surgical images and introduces a structured belief state to explicitly represent degradation type and severity. This belief-conditioned formulation provides a more interpretable basis for Agentic-RL-based tool selection and trajectory optimization in multi-degradation surgical restoration.

3 Method

3.1 Overview

Real-world surgical images often suffer from multiple coupled degradations. Since different restoration operations can interact with each other, their execution order can substantially affect the final restoration quality. This makes fixed restoration pipelines suboptimal due to varying degradation types, severities, and tool dependencies. We propose BeTAR, a two-stage degradation-aware restoration agent. First, SFT and GRPO train the perception model to identify degradation types and severity levels, which are represented as a structured belief state. This belief provides the degradation-aware state representation for the subsequent POMDP. Conditioned on the belief, Agentic-RL optimizes a complete restoration trajectory, jointly determining tool selection, execution order, and termination using reference-free image quality rewards and tool-use constraints. The overall framework is shown in Figure 3.

3.2 Degradation-Aware Visual Perception

Degradation-Aware Belief Construction. As shown in Figure 2,we employ Qwen2.5-VL-7B-Instruct (Wu et al., 2025a) as the visual-language backbone and introduce a five-stage structured reasoning template inspired by chain-of-thought (CoT) prompting (Wei et al., 2022). Given an input image x, the model progressively reasons from semantic understanding to degradation perception:

$$\mathbf{h} = \left(h^{\text{scene}}, h^{\text{content}}, h^{\text{global}}, h^{\text{local}}, h^{\text{deg}}\right). \tag{1}$$

Specifically, h^{scene} first summarizes the overall surgical scene and provides a coarse semantic context for subsequent quality assessment. Then, h^{content} identifies the major anatomical tissues, surgical instruments, and their spatial relationships, helping the model distinguish normal scene content from degradation-induced appearance changes. Based on these semantic cues, h^{global} evaluates image-level visual quality, including illumination, visibility, contrast, and sharpness, to capture degradations that affect a large portion of the image. In contrast, h^{local} focuses on spatially

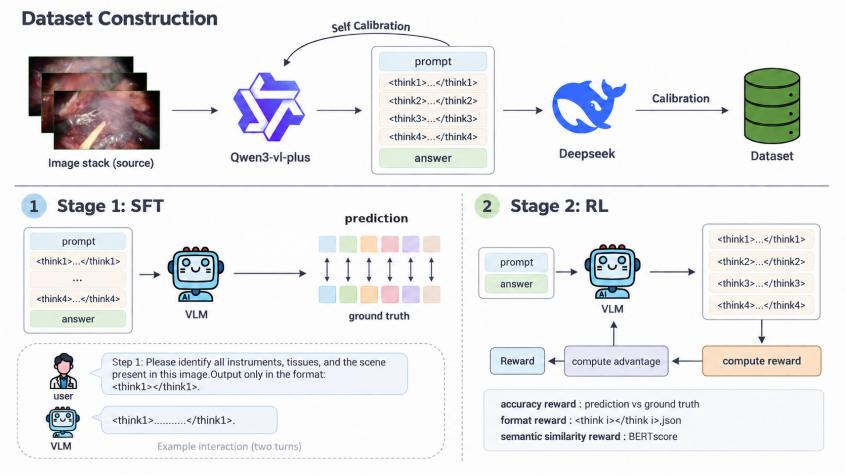

Figure 2: Training pipeline of the degradation-aware visual perception model. Structured surgical-image annotations are first constructed and used for SFT initialization, followed by reward-guided reinforcement learning to refine the structured reasoning and degradation prediction outputs.

localized quality abnormalities, such as regional blur, specular highlights, water droplets, or non-uniform smoke, and estimates their affected regions. Finally, h^{deg} aggregates the semantic and perceptual evidence from the preceding stages and converts it into structured degradation types together with their corresponding severity levels. This coarse-to-fine reasoning process guides the model from scene understanding to explicit degradation assessment, providing richer intermediate evidence for more stable, fine-grained, and interpretable degradation perception.The complete perception-training pipeline is illustrated in Figure 2.

GRPO-Based Perception Refinement. We first perform supervised fine-tuning (SFT) on approximately 2,000 structured image–response pairs to initialize the model with stable structured reasoning and basic degradation recognition capabilities, and then further refine the perception policy using GRPO (Shao et al., 2024). Since degradation perception relies on a multi-stage reasoning process rather than only the final prediction, supervising the outcome alone may allow semantically inconsistent intermediate reasoning. We therefore employ two outcome-based rewards together with one process-based reward in GRPO. The distinction between outcome and process supervision follows prior work on step-level reasoning feedback (Lightman et al., 2024), while our process reward specifically encourages semantic consistency between the generated reasoning stages and their reference descriptions. For each image, the policy samples G candidate responses $\mathbf{y}^{(i)}{}_{i=1}^{G}$.

The outcome-based accuracy reward $r_{\text{acc}}^{(i)}$ evaluates whether the predicted degradation types agree with the ground truth, while the format reward $r_{\text{fmt}}^{(i)}$ assesses whether the response follows the required five-stage CoT structure and produces a valid, parseable degradation output. These outcome rewards constrain the correctness and structural validity of the final response, but do not explicitly verify whether the intermediate reasoning is supported by the visual evidence. Therefore, process-level supervision is introduced to further regularize the semantic quality and consistency of the structured reasoning process.

To provide process-level supervision, we compute the semantic similarity between the predicted and ground-truth reasoning contents in the first four stages:

$$r_{\text{sem}}^{(i)} = \frac{1}{4}\sum_{m=1}^{4} \cos\left(\phi\left(\hat{h}_m^{(i)}\right), \phi\left(h_m^*\right)\right), \tag{2}$$

where $\phi(\cdot)$ denotes the semantic encoder, while $\hat{h}_m^{(i)}$ and h_m^* are the predicted and ground-truth contents of the m-th reasoning stage, respectively.

The overall perception reward is defined as

$$r_{\text{perc}}^{(i)} = \lambda_{\text{acc}} r_{\text{acc}}^{(i)} + \lambda_{\text{fmt}} r_{\text{fmt}}^{(i)} + \lambda_{\text{sem}} r_{\text{sem}}^{(i)}. \tag{3}$$

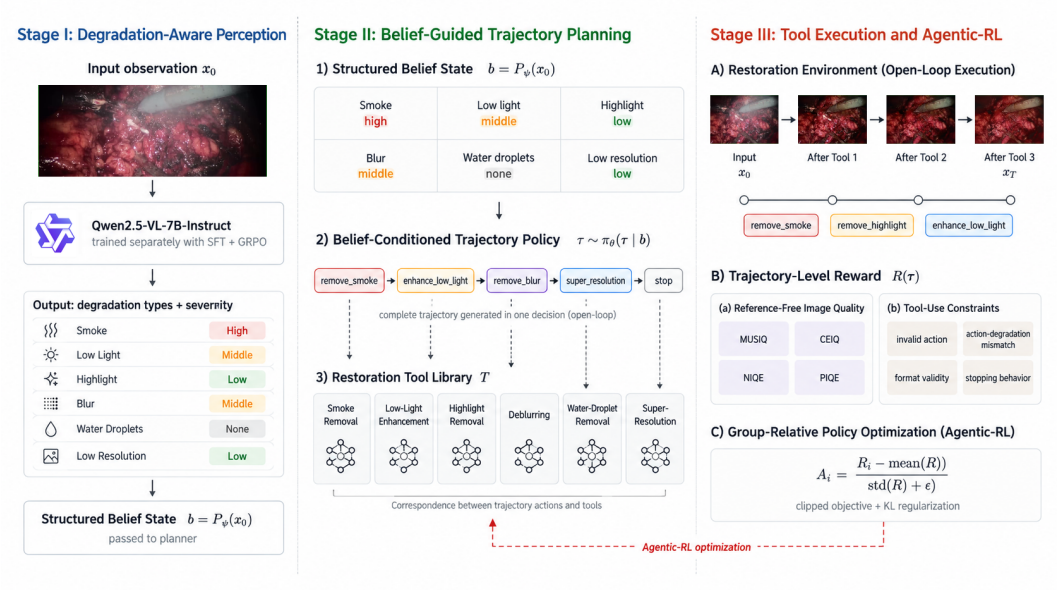

Figure 3: Overall architecture of BeTAR. The framework first performs degradation-aware perception to construct a structured belief state, then generates a belief-conditioned open-loop restoration trajectory, executes specialized restoration tools, and optimizes the trajectory policy with trajectory-level Agentic-RL rewards.

3.3 Belief-Guided Agentic Policy Optimization

After degradation-aware perception training, we formulate multi-degradation surgical image restoration as a partially observable Markov decision process (POMDP) (Kaelbling et al., 1998):

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \Omega, \mathcal{R}, \gamma)\,. \tag{4}$$

Here, $\mathcal{S}$ represents the latent image and degradation state, $\mathcal{A}$ contains the restoration tools and the terminal action, $\mathcal{O}$ denotes the observable surgical image, and $\mathcal{T}$ describes image transitions caused by tool execution. $\mathcal{R}$ is the trajectory-level reward and $\gamma = 1$ since the reward is assigned after the complete trajectory is executed.

Given the initial image x_0, the perception model $\mathcal{P}_\psi$, trained through SFT and GRPO, predicts degradation types and severity levels and constructs the structured belief

$$b = \mathcal{P}_\psi(x_0) = \left[(\hat{y}_k, \hat{\ell}_k)\right]_{k=1}^{K}, \tag{5}$$

where $\hat{y}_k$ and $\hat{\ell}_k$ denote the presence and severity of the k-th degradation, respectively. We consider $K = 6$ degradation categories: smoke, low light, highlight, blur, water droplets, and low resolution. The belief state therefore serves as the interface between visual perception and restoration planning.

Belief-Guided Trajectory Planning. Conditioned on b, the policy jointly determines which restoration tools should be selected, their execution order, and when the restoration process should terminate. The probability of a complete trajectory is factorized as

$$\pi_\theta(\tau \mid b) = \prod_{t=1}^{T} \pi_\theta(a_t \mid b, a_{<t}), \tag{6}$$

where $\tau = (a_1, \ldots, a_T)$, $a_T = a_{\text{stop}}$, and $T \le T_{\max}$. Each non-terminal action invokes a fixed restoration tool $\mathcal{F}_{a_t}$ and updates the image as $x_{t+1} = \mathcal{F}_{a_t}(x_t)$. Unlike independent tool classification or predefined restoration chains, this formulation explicitly models dependencies among restoration actions and learns an image-adaptive execution order.

Table 1: Quantitative comparison of surgical image restoration methods on two test sets. The best results are shown in bold and the second-best results are underlined.

Method	Independent Test Set (1,003 Images)						In-Distribution Test Set (1,500 Images)					
	MANIQA↑	MUSIQ↑	PI↓	LIQE↑	FADE↓	IS↑	MANIQA↑	MUSIQ↑	PI↓	LIQE↑	FADE↓	IS↑
EndoCaver	0.1175	14.9772	8.6736	1.0079	0.7108	1.9789	0.0883	18.7703	8.3194	1.0027	0.5793	1.7414
LightDiff	0.1808	43.7101	5.1243	2.0667	0.5420	2.4946	0.1933	40.4734	5.1191	2.0671	0.5246	1.9701
AMIR	0.1852	43.1367	5.5206	1.8046	0.4963	2.2674	0.1592	31.8606	6.5313	1.4881	0.4525	1.5483
AST	0.2107	40.2131	4.9520	1.7671	0.4534	2.0099	0.2148	34.2848	5.8776	1.5172	0.4355	1.9601
MambaIRv2	0.1996	44.0908	5.9134	1.9462	0.5228	2.1498	0.1812	34.7547	6.4655	1.5818	0.5072	1.5112
SFHformer	0.1997	46.0328	4.3238	2.1756	**0.4117**	2.5136	0.1996	44.0908	5.9134	1.9462	0.5228	1.7578
Desmoke-LAP	0.1849	42.9366	4.5363	1.6995	0.9006	2.4855	0.2019	38.6217	**4.8646**	1.4173	1.1490	1.6098
MDA-SR	0.1776	31.3016	6.6571	1.0238	0.4551	1.8837	0.1545	24.3437	6.5561	1.0074	0.5188	1.5945
BeTAR	**0.2303**	**48.8669**	**4.1058**	**2.2932**	0.4455	**2.5708**	**0.2460**	**44.6037**	4.8876	**2.1893**	**0.4343**	**2.0997**

During training, the policy samples G candidate trajectories:

$$\tau_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid b), \qquad i = 1, \ldots, G. \tag{7}$$

The restoration environment executes each trajectory sequentially and produces the corresponding final image $x_T^{(i)}$. Since the policy is open-loop, intermediate images are transferred only between restoration tools and are not used for belief updating or replanning.

Trajectory-Level Reward. Each candidate trajectory is evaluated using a reference-free reward:

$$R_i = \lambda_{\text{IQA}} R_{\text{IQA}}\left(x_T^{(i)}\right) + \lambda_{\text{tool}} R_{\text{tool}}(\tau_i, b) + \lambda_{\text{valid}} R_{\text{valid}}(\tau_i). \tag{8}$$

Here, R_{IQA} aggregates direction-aligned and normalized MANIQA (Yang et al., 2022), MUSIQ (Ke et al., 2021), NIQE (Mittal et al., 2012b), BRISQUE (Mittal et al., 2012a), and PIQE (Venkatanath et al., 2015) scores. The term R_{tool} penalizes unnecessary, repeated, invalid, or degradation-mismatched actions, while R_{valid} evaluates whether the generated trajectory follows the required action format. The reward therefore jointly considers final image quality and trajectory efficiency without requiring paired clean references.

Agentic Trajectory Optimization. Following GRPO (Shao et al., 2024), for trajectories sampled from the same input image, the relative advantage is computed as $\widehat{A}_i = \frac{R_i - \overline{R}}{\sigma_R + \epsilon}$, where $\overline{R}$ and σ_R denote the mean and standard deviation of the group rewards, respectively. This group-relative normalization eliminates the need for an additional value network.

The trajectory policy is optimized using a PPO-style clipped surrogate objective (Schulman et al., 2017) with the group-relative advantages and KL regularization of GRPO (Shao et al., 2024):

$$\mathcal{L}_{\text{agent}} = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\tau_i|}\sum_{n=1}^{|\tau_i|}\min\Big(\rho_{i,n}\widehat{A}_i, \text{clip}(\rho_{i,n}, 1-\varepsilon, 1+\varepsilon)\widehat{A}_i\Big)\right] + \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \tag{9}$$

where $\rho_{i,n}$ is the probability ratio between the current and old policies for the n-th token of trajectory τ_i, and π_{ref} is the frozen reference policy. By assigning a trajectory-level advantage to all tokens in the same trajectory, the agent jointly optimizes tool selection, execution order, trajectory length, and stopping behavior. As shown in Algorithm 1, the agent first constructs the degradation belief state, samples multiple restoration trajectories, and then updates the trajectory policy using group-relative advantages.

4 Experiments

4.1 Experiment Settings

Datasets. We construct separate datasets for degradation perception and restoration policy optimization. For perception training, Qwen3-VL-Plus (Bai et al., 2025) generates five-stage structured annotations covering scene understanding, image content, global/local quality assessment, and

Algorithm 1 Agentic-RL Training of BeTAR

Require: Training set $\mathcal{D}$, policy π_θ, perception model $\mathcal{P}_\psi$, tools $\{\mathcal{F}_a\}$, group size G

for $x_0 \in \mathcal{D}$ **do**
  $b_0 \leftarrow \mathcal{P}_\psi(x_0)$
  for $i = 1$ to G **do**
    $\tau_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid b_0)$; $x^{(i)} \leftarrow x_0$
    for $a_t \in \tau_i$ **do**
      if $a_t = \texttt{stop}$ **then**
        break
      end if
      $x^{(i)} \leftarrow \mathcal{F}_{a_t}(x^{(i)})$
    end for
    $R_i \leftarrow R_{\text{traj}}(x_0, \tau_i, x^{(i)})$
  end for
  Compute $\{\widehat{A}_i\}_{i=1}^{G}$ and update θ with $\mathcal{L}_{\text{agent}}$
end for
return Optimized policy π_θ

degradation prediction. After self-calibration and independent verification with DeepSeek, approximately 10,000 image–response pairs are obtained. For restoration planning, Agentic-RL is used to optimize the policy that generates complete tool trajectories conditioned on the perceived degradation state. For Agentic-RL training, we use unlabeled surgical images, with the primary dataset split at approximately 8:2 into over 8,000 training images and about 1,500 test images. An additional independent set of 1,003 images is used to evaluate cross-dataset generalization.

Implementation Details. We adopt Qwen2.5-VL-7B-Instruct (Wu et al., 2025a) as the visual-language backbone. The degradation perception model is first initialized through supervised fine-tuning and subsequently refined using perception-oriented GRPO. The SFT stage is conducted on four NVIDIA RTX 4090 GPUs for three epochs, while both the perception-level GRPO and trajectory-level Agentic-RL stages are trained on two NVIDIA H100 GPUs for one epoch.

Restoration Tools. The restoration tool library contains six pretrained models: PhySe-RPO (Fang et al., 2026) for smoke removal, WTNet (Wu et al., 2025b) for low-light enhancement, EndoSRR (Li et al., 2024) for highlight removal, MPT-CataBlur (Zhang et al., 2024) for deblurring, XRestormer (Chen et al., 2024c) from the Surgical-Image-Restoration benchmark (Pei et al., 2026) for water-droplet removal, and DCSSR (Zhang et al., 2022) for super-resolution. These tools remain fixed during Agentic-RL training and are invoked through the corresponding agent actions.

Evaluation Metrics. We evaluate degradation perception using Accuracy, Recall, and F1 score, which measure the correctness and completeness of the predicted degradation annotations. The evaluation considers the model's ability to recognize multiple coexisting degradation types and estimate their corresponding severity levels. Since clean reference images are generally unavailable for real surgical data, restoration performance is evaluated using MANIQA (Yang et al., 2022), PI (Blau et al., 2018), FADE (Choi et al., 2015), MUSIQ (Ke et al., 2021), and LIQE (Zhang et al., 2023), together with the dataset-level Inception Score (IS) (Salimans et al., 2016).

4.2 Experimental Results

4.2.1 Quantitative Analysis

Surgical Image Restoration. Table 1 compares BeTAR with EndoCaver (Wu et al., 2026), LighT-Diff (Chen et al., 2024b), AMIR (Yang et al., 2024), AST (Zhou et al., 2024), MambaIRv2 (Guo et al., 2025), SFHformer (Jiang et al., 2024), DeSmoke-LAP (Pan et al., 2022), and MDA-SR (Liu et al., 2023b). On the independent test set, BeTARranks first on five of six metrics, achieving 0.2303 MANIQA and 48.8669 MUSIQ. It maintains this advantage on the in-distribution test set, reaching 0.2460 MANIQA and 2.0997 IS. These results demonstrate consistent restoration performance and cross-dataset generalization.

Degradation Perception. We first compare the degradation perception capability of BeTAR with representative general-purpose and medical vision-language models. As shown in Table 2, our model achieves the best performance across all three metrics, obtaining an Acc of 0.633, a Recall of 0.889, and an F1 score of 0.908. These results demonstrate that structured reasoning supervision and reward-guided refinement substantially improve the recognition of coexisting degradation types and severity levels. In particular, the strong recall and F1 performance suggest that the model is

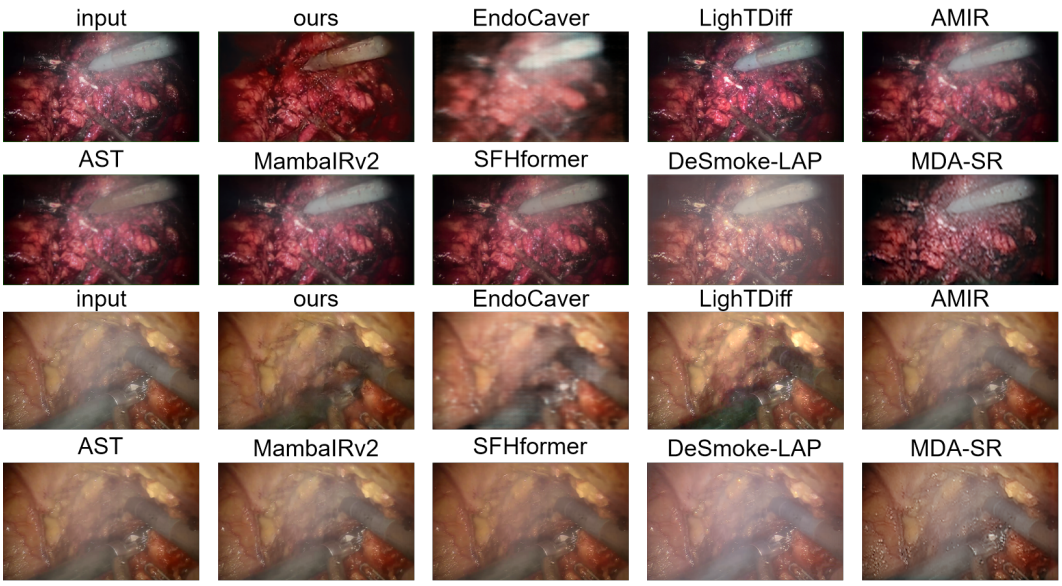

Figure 4: Qualitative comparison on representative multi-degradation surgical images. The examples illustrate restoration behavior under different combinations of smoke, illumination degradation, blur, highlight artifacts, and other surgery-specific distortions.

Table 2: Comparison of degradation perception performance.

Model	Acc↑	Recall↑	F1↑
HuatuoGPT-Vision-7B	0.251	0.559	0.637
Qwen3.5-27B	0.398	0.831	0.712
Lingshu-7B	0.217	0.609	0.693
InternVL3.5-8B	0.258	0.476	0.539
Hulu-Med-7B	0.111	0.306	0.477
MedGemma-1.5-4B	0.171	0.406	0.336
BeTAR	**0.633**	**0.889**	**0.908**

Table 3: Ablation of SFT and GRPO.

Model	Acc↑	Recall↑	F1↑
Qwen2.5-VL-7B-Instruct	0.247	0.654	0.670
+ SFT	0.459	0.782	0.836
+ SFT + GRPO	**0.633**	**0.889**	**0.908**

Table 4: Agentic-RL efficiency.

Method	Tool Calls↓	Time/Image↓
Before Agentic-RL	2.72	9.27 s
After Agentic-RL	**2.44**	**7.75 s**

more robust to mixed degradations, where multiple corruption factors may overlap spatially or exhibit similar visual characteristics. This provides a more reliable degradation belief for subsequent restoration planning and reduces the risk of missing critical degradations during tool selection.

4.2.2 Qualitative Analysis

Figure 4 presents qualitative comparisons on representative surgical images with multiple coexisting degradations. Compared with existing methods, several baselines still exhibit residual blur, illumination inconsistency, color deviation, or over-enhancement artifacts. In contrast, BeTAR achieves more balanced restoration, with clearer tissue structures and improved local contrast. Notably, the restored images preserve anatomical boundaries and tissue appearance while avoiding excessive enhancement in already well-exposed regions. Such behavior is particularly important in mixed-degradation scenarios, where independently applying multiple restoration operations can easily introduce secondary artifacts. These results indicate that degradation-aware perception and belief-conditioned trajectory planning enable the agent to coordinate multiple restoration tools more effectively for complex surgical scenes.

4.3 Ablation Study

Ablation on Key Components. We evaluate the contribution of the key components in BeTAR in Tables 3 and 5. For degradation perception, SFT provides a strong initialization for structured

degradation understanding, while perception-level GRPO further improves recognition by jointly refining prediction accuracy, output structure, and reasoning consistency. For restoration planning, trajectory-level Agentic-RL consistently improves restoration quality on both test sets, indicating that optimizing complete restoration trajectories is beneficial for learning effective tool dependencies and execution orders. These results verify the complementary contributions of structured perception learning and trajectory-level policy optimization.

Table 5: Ablation study of POMDP and trajectory-level Agentic-RL on two test sets.

Method	Independent Test Set (1,003 Images)						In-Distribution Test Set (1,500 Images)					
	MANIQA↑	MUSIQ↑	PI↓	LIQE↑	FADE↓	IS↑	MANIQA↑	MUSIQ↑	PI↓	LIQE↑	FADE↓	IS↑
BeTAR w/o Agentic-RL & POMDP	0.1870	42.1433	5.2720	1.7143	0.5196	2.2441	0.1855	41.9440	5.4717	1.7845	0.4983	1.7612
BeTAR w/o POMDP	0.1920	44.0464	4.7277	1.9807	0.4543	2.4633	0.1980	39.7764	5.1145	1.9679	0.4577	1.8889
BeTAR w/o Agentic-RL	0.2015	45.3157	4.9625	1.8854	0.4877	2.4560	0.2252	42.5369	4.9162	2.0711	0.4759	2.0836
BeTAR	**0.2303**	**48.8669**	**4.1058**	**2.2932**	**0.4455**	**2.5708**	**0.2460**	**44.6037**	**4.8876**	**2.1893**	**0.4343**	**2.0997**

Ablation on Tool Selection Strategies. We further analyze the influence of tool selection strategies in Tables 4 and 6. Compared with fixed or random tool selection, degradation-aware selection guided by the structured belief state consistently produces better restoration results, demonstrating that explicit degradation information provides useful guidance for choosing appropriate restoration operations. Meanwhile, the learned agent avoids unnecessary tool invocations and produces more compact trajectories, leading to lower computational cost and shorter processing time. These observations suggest that effective restoration depends not only on selecting suitable tools, but also on adapting their combination, ordering, and stopping behavior to the degradation characteristics of each input image.

Table 6: Ablation of different tool-selection strategies on the independent and in-distribution test sets.

Method	Independent Test Set (1,003 Images)						In-Distribution Test Set (1,500 Images)					
	MANIQA↑	MUSIQ↑	PI↓	LIQE↑	FADE↓	IS↑	MANIQA↑	MUSIQ↑	PI↓	LIQE↑	FADE↓	IS↑
Fixed Tools	0.2022	43.6613	6.3946	1.6284	0.5384	2.2692	0.1928	39.3624	5.7205	2.1885	0.4907	1.8070
Random Tools	0.1996	41.2457	5.3753	1.9329	0.5077	2.3165	0.2053	40.6206	5.4717	1.9696	0.4791	1.7775
VLM-Guided Selection	**0.2303**	**48.8669**	**4.1058**	**2.2932**	**0.4455**	**2.5708**	**0.2460**	**44.6037**	**4.8876**	**2.1893**	**0.4343**	**2.0997**

4.4 WHY POMDP-GUIDED AGENTIC-RL?

Direct Agentic-RL maps the input image to a restoration trajectory without explicitly representing the underlying degradation state, making tool selection difficult to interpret when multiple degradations coexist. We therefore use the POMDP formulation to convert the initial image into a structured belief state containing degradation types and severity levels. Conditioned on this belief, the agent generates the complete restoration trajectory in a single decision. Each selected tool can thus be traced to a corresponding degradation estimate, while severity information provides a basis for tool priority and termination. POMDP offers an interpretable decision representation, whereas Agentic-RL optimizes tool selection, ordering, trajectory length, and stopping behavior.

5 CONCLUSION

We presented BeTAR, a degradation-aware framework that combines structured VLM perception with POMDP-guided Agentic-RL for multi-degradation surgical image restoration. By constructing an interpretable degradation belief and optimizing trajectory-level rewards, the agent learns adaptive tool selection, execution order, and termination without paired clean images. Experiments demonstrate strong degradation recognition and restoration performance across both in-distribution and independent test sets. One limitation of the current framework is that restoration is performed with a fixed tool library under an open-loop planning strategy, which limits its ability to adapt to unexpected intermediate restoration outcomes. Future work will therefore explore more diverse and specialized restoration tools, closed-loop trajectory adjustment based on intermediate feedback, and task-aware restoration objectives incorporating downstream clinical tasks such as segmentation and instrument recognition.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript formatting, language editing, literature sourcing, and reference formatting. All AI-assisted outputs were reviewed and revised by the authors, who take responsibility for the final content, claims, and artifacts presented in this work.

REFERENCES

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.

Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *European Conference on Computer Vision*, pp. 334–355. Springer, 2018.

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. *Advances in Neural Information Processing Systems*, 37:110643–110666, 2024a.

Tong Chen, Qingcheng Lyu, Long Bai, Erjian Guo, Huxin Gao, Xiaoxiao Yang, Hongliang Ren, and Luping Zhou. Lightdiff: surgical endoscopic image low-light enhancement with t-diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 369–379. Springer, 2024b.

Tong Chen, Xinyu Ma, Long Bai, Wenyang Wang, Yue Sun, and Luping Zhou. Endoir: Degradation-agnostic all-in-one endoscopic image restoration via noise-aware routing diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3047–3055, 2026.

Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024c.

Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik. Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Transactions on Image Processing*, 24(11): 3888–3901, 2015.

Zining Fang, Cheng Xue, Chunhui Liu, Bin Xu, Ming Chen, and Xiaowei Hu. Physe-rpo: Physics and semantics guided relative policy optimization for diffusion-based surgical smoke removal. *arXiv preprint arXiv:2603.22844*, 2026.

Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. In *2025 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 28124–28133. IEEE, 2025.

Xingyu Jiang, Xiuhui Zhang, Ning Gao, and Yue Deng. When fast fourier transform meets transformer for image restoration. In *European conference on computer vision*, pp. 381–402. Springer, 2024.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Bingchen Li, Xin Li, Yiting Lu, and Zhibo Chen. Hybrid agents for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22636–22647, 2026.

Wei Li, Fucang Jia, and Wenjian Liu. Endosrr: a comprehensive multi-stage approach for endoscopic specular reflection removal. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):1203–1211, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Yunlong Lin, Zixu Lin, Haoyu Chen, Panwang Pan, Chenxin Li, Sixiang Chen, Kairun Wen, Yeying Jin, Wenbo Li, and Xinghao Ding. Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22369–22380. IEEE, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Tianbao Liu, Zefeiyun Chen, Qingyuan Li, Yusi Wang, Ke Zhou, Weijie Xie, Yuxin Fang, Kaiyi Zheng, Zhanpeng Zhao, Side Liu, et al. Mda-sr: multi-level domain adaptation super-resolution for wireless capsule endoscopy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 518–527. Springer, 2023b.

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.

Yirou Pan, Sophia Bano, Francisco Vasconcelos, Hyun Park, Taikyeong Ted Jeong, and Danail Stoyanov. Desmoke-lap: improved unpaired image-to-image translation for desmoking in laparoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 17(5): 885–893, 2022.

Jialun Pei, Diandian Guo, Donghui Yang, Zhixi Li, Yuxin Feng, Long Ma, Bo Du, and Pheng-Ann Heng. Benchmarking endoscopic surgical image restoration and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37413–37422, 2026.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Narasimhan Venkatanath, Debbabi Praneeth, S Channappayya Sumohana, S Medasani Swarup, et al. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pp. 1–6. IEEE, 2015.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Zhijian Wu, Hong Wang, Yuxuan Shi, Dingjiang Huang, and Yefeng Zheng. A prior-driven lightweight network for endoscopic exposure correction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 13–23. Springer, 2025b.

Zhuoyu Wu, Wenhui Ou, Pei-Sze Tan, Jiayan Yang, Wenqi Fang, Zheng Wang, and Raphaël C-W Phan. Endocaver: Handling fog, blur and glare in endoscopic images via joint deblurring-segmentation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6981–6985. IEEE, 2026.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1190–1199. IEEE, 2022.

Zhiwen Yang, Haowei Chen, Ziniu Qian, Yang Yi, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. All-in-one medical image restoration via task-adaptive routing. In *International conference on medical image computing and computer-assisted intervention*, pp. 67–77. Springer, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Tianyi Zhang, Yun Gu, Xiaolin Huang, Jie Yang, and Guang-Zhong Yang. Disparity-constrained stereo endoscopic image super-resolution. *International Journal of Computer Assisted Radiology and Surgery*, 17(5):867–875, 2022.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081. IEEE, 2023.

Yuelin Zhang, Pengyu Zheng, Wanquan Yan, Chengyu Fang, and Shing Shin Cheng. A unified framework for microscopy defocus deblur with multi-pyramid transformer and contrastive learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11125–11136. IEEE, 2024.

Shihao Zhou, Duosheng Chen, Jinshan Pan, Jinglei Shi, and Jufeng Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2952–2963. IEEE, 2024.

Yingjie Zhou, Jiezhang Cao, Farong Wen, Zicheng Zhang, Yu Zhou, Yue Shi, Xiaohong Liu, Radu Timofte, Luc Van Gool, 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 *International Conference on Learning Representations*, volume 2025, pp. 57985–58013, 2025.