ENDOTME: PHYSICS-INSPIRED DEGRADATION-AWARE TRAJECTORY MEMORY FOR ENDOSCOPIC VIDEO ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Endoscopic video enhancement is crucial for improving surgical visibility and supporting reliable intraoperative visual perception. However, existing video restoration methods commonly rely on generic temporal feature propagation, which suffers from two key limitations. First, they tend to indiscriminately aggregate historical frame information that may already be distorted or degraded. Second, they lack explicit awareness of endoscopy-specific degradations, making them less effective in challenging surgical scenes involving blood haze, specular reflection, motion blur, low contrast, and temporal instability. To address these issues, we propose EndoTME, an endoscopic video enhancement framework that improves the reliability of temporal information and the adaptivity to complex degradations. Specifically, we develop a reliable trajectory memory retrieval mechanism to selectively exploit trustworthy historical features while suppressing unreliable temporal cues caused by severe image degradation or unstable motion. Such reliability-aware memory selection reduces the cross-frame propagation of corrupted temporal information and improves the temporal coherence of enhanced videos. Meanwhile, we introduce a degradation conditioning mechanism based on physics-inspired visual cues, which explicitly characterizes typical endoscopic degradations and facilitates adaptation of the restoration process to different local degradation patterns. Experiments on a paired synthetic dataset and two real-world endoscopic video datasets demonstrate that EndoTME outperforms representative image and video restoration baselines.

1 INTRODUCTION

Endoscopic video serves as the primary visual interface in minimally invasive surgery, providing essential guidance for anatomical observation, instrument manipulation, and intraoperative decision-making. However, real surgical videos are frequently affected by smoke, blood or fluid contamination, specular reflection, motion blur, insufficient illumination, and unstable camera motion (Wu et al., 2024; Chen et al., 2024; Liang et al., 2026; Bolkar et al., 2018). These degradations may coexist and vary rapidly across space and time, obscuring fine tissue boundaries and clinically relevant structures while reducing the temporal stability of the visual field (Pan et al., 2022; Wang et al., 2024). Therefore, restoring clear and temporally coherent endoscopic videos is important for reliable surgical perception, yet remains challenging under diverse and complex real-world degradations.

A natural solution is to exploit complementary observations from neighboring or historical frames. Representative video restoration methods align and propagate multi-frame features using deformable alignment, recurrent propagation, or attention-based interaction (Wang et al., 2019; Chan et al., 2021; 2022a; Isobe et al., 2020; Liu et al., 2022; Xu et al., 2023). Surgical video methods such as SelfSVD and STANet also demonstrate the value of temporal information for recovering smoke-obscured structures (Wu et al., 2024; Liang et al., 2026). These methods are effective when the propagated history contains clearer and correctly aligned evidence. However, RealBasicVSR reports that severe real-world degradations can be exaggerated through long-term propagation, and therefore introduces a cleaning stage before temporal propagation (Chan et al., 2022b). This observation is especially relevant to endoscopic videos, where a motion-aligned historical region may itself contain blood haze, saturated reflection, blur, or inaccurate motion correspondence. Consequently,

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.

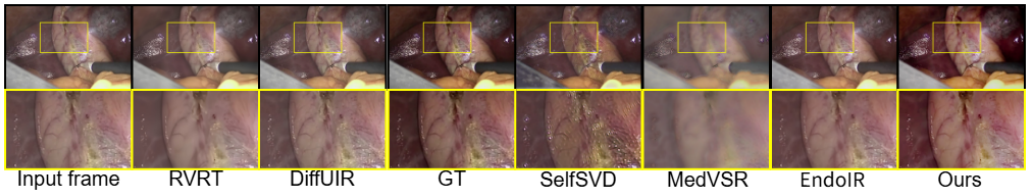


Figure 1: Endoscopic video suffers from blood haze, specular reflection, low contrast, and blurred tissue boundaries. Existing restoration methods may leave residual degradation, over-smooth anatomical structures, or introduce unnatural appearance. In contrast, EndoTME restores clearer local structures and yields results closer to the clean reference.

motion alignment answers *where* historical evidence can be retrieved, but not *whether* that evidence is sufficiently reliable to be used.

The above limitation also indicates that reliable temporal selection alone is insufficient: the restoration process must understand the degradation affecting each region. Degradation-aware restoration methods have therefore been developed to adapt a unified model to different corruption types and severities (Valanarasu et al., 2022; Park et al., 2023; Zhang et al., 2023; Ai et al., 2024; Luo et al., 2025). AirNet learns contrastive degradation representations for unknown corruptions (Li et al., 2022), while PromptIR uses degradation-specific prompts to dynamically guide restoration (Potlapalli et al., 2023). PromptRestorer further observes that degradation information may be gradually forgotten during restoration and introduces perceived degradation priors from global and local perspectives (Wang et al., 2023a). More recently, EndoIR extends all-in-one restoration to co-occurring endoscopic degradations (Chen et al., 2026). Although these studies establish the importance of degradation conditioning, most existing formulations operate primarily at the image or latent-feature level. They do not explicitly associate interpretable endoscopic degradation distributions with region-specific restoration behaviors. This distinction is important because different areas within the same frame may require substantially different operations: veil-contaminated regions require suppression, blurred tissue boundaries require structure recovery, specular regions require conservative highlight handling, and motion-unstable regions require temporally stable refinement.

To address these challenges, we propose **EndoTME**, an endoscopic video enhancement framework that selectively exploits reliable temporal evidence and adapts restoration to spatially heterogeneous degradations. Specifically, to address the discrepancy between motion correspondence and information reliability, we propose Reliable Trajectory Memory Retrieval (RTMR), which prevents unreliable historical evidence from interfering with the current restoration while preserving useful temporal information. To accommodate the distinct restoration requirements of different local degradations, we further propose Physics-Inspired Degradation Conditioning (PIDC), which associates interpretable endoscopic degradation cues with spatially adaptive restoration behaviors. As illustrated in Fig. 1, EndoTME produces clearer tissue structures and a more natural visual appearance than representative restoration methods.

The main contributions of this work are summarized as follows:

- We propose Reliable Trajectory Memory Retrieval, which distinguishes information reliability from motion correspondence and selectively exploits trustworthy historical evidence for temporally stable endoscopic video enhancement.
- We propose Physics-Inspired Degradation Conditioning, which connects interpretable degradation cues with spatially varying restoration behaviors to accommodate heterogeneous artifacts within the same endoscopic frame.
- Extensive experiments on paired and real-world endoscopic videos demonstrate favorable restoration fidelity, perceptual quality, temporal consistency, and computational efficiency.

2 Related Work

Temporal Alignment and Propagation for Video Restoration. Temporal redundancy is widely exploited in video restoration by aligning neighboring frames and propagating historical features. Representative methods employ deformable alignment, recurrent propagation, or attention-based aggregation to establish temporal correspondence and enlarge the effective receptive field (Wang et al., 2019; Chan et al., 2021; 2022a; Isobe et al., 2020; Liu et al., 2022; Xu et al., 2023). However, medical and endoscopic videos present additional temporal challenges caused by camera shake, rapid viewpoint changes, tissue deformation, surgical instrument motion, and frame-dependent degradation. MedVSR reports that camera shake, noise, and abrupt frame transitions in medical videos can produce substantial optical-flow errors and alignment difficulties, and therefore selectively propagates consistent and informative features (Liu et al., 2025). For laparoscopic video desmoking, SelfSVD uses the clear frame captured before smoke generation as supervision and temporal reference information for subsequent smoky frames (Wu et al., 2024). STANet further observes that diffusion smoke and ambient smoke exhibit different motion patterns and spatio-temporal characteristics, and models them separately during laparoscopic video desmoking (Liang et al., 2026). Recent real-time endoscopic video enhancement methods also estimate and propagate degradation representations across neighboring frames to exploit short-term temporal continuity (Xu et al., 2026). Nevertheless, RealBasicVSR shows that severe degradations may be amplified during long-term propagation (Chan et al., 2022b), while the findings of MedVSR indicate that motion correspondence can become unreliable in challenging medical videos (Liu et al., 2025). Existing methods therefore focus mainly on how neighboring or historical information is aligned and propagated, and commonly treat aligned observations as useful temporal evidence. They do not explicitly determine whether each aligned historical observation is sufficiently clean, informative, and reliable for subsequent feature propagation.

Degradation-Aware Image Restoration. Degradation-aware restoration aims to condition a shared model on the degradation affecting an input image. General restoration methods learn degradation representations or prompts to distinguish different corruption types and severities (Li et al., 2017; Valanarasu et al., 2022; Park et al., 2023; Zhang et al., 2023; Ai et al., 2024; Luo et al., 2025). AirNet learns contrastive degradation representations (Li et al., 2022), PromptIR introduces degradation-specific prompts (Potlapalli et al., 2023), and PromptRestorer models degradation priors from global and local perspectives (Wang et al., 2023a). In medical and endoscopic imaging, existing studies have primarily addressed individual or predefined groups of surgical degradations. LighTDiff focuses on low-light enhancement for surgical endoscopic images (Chen et al., 2024), while SelfSVD and STANet are designed primarily for laparoscopic smoke removal (Wu et al., 2024; Liang et al., 2026; Bolkar et al., 2018; Pan et al., 2022; Wang et al., 2024). EndoIR extends endoscopic image restoration to co-occurring low illumination, smoke, and bleeding by using degradation-aware conditioning and noise-aware routing (Chen et al., 2026). SurgClean establishes a real-world benchmark covering desmoking, defogging, and desplashing and reports a substantial gap between existing restoration methods and clinical requirements (Pei et al., 2026). These studies motivate explicit degradation modeling, but they do not address the reliability of trajectory-aligned historical evidence in video restoration. EndoTME differs from prior degradation-aware image restoration in two complementary aspects: RTMR evaluates each trajectory-aligned historical candidate before aggregation, whereas PIDC uses endoscopy-oriented analytical cues to condition spatially adaptive restoration. The two modules therefore address temporal evidence admission and local degradation adaptation within a unified video-enhancement framework.

3 Method

3.1 Overview

As shown in Fig. 2, given a degraded video clip $\mathcal{I}_{LQ}^{t-T:t} = \{\mathbf{I}_{LQ}^{t-T}, \ldots, \mathbf{I}_{LQ}^{t}\}$, a shared encoder extracts the corresponding feature sequence $\{\mathbf{F}^{t-T}, \ldots, \mathbf{F}^{t}\}$. For each time step τ, Reliable Trajectory Memory Retrieval (RTMR) traces historical features along estimated motion trajectories, evaluates their reliability, and aggregates the top-ranked memories into $\mathbf{F}_{traj}^{\tau}$. Physics-Inspired Degradation Conditioning (PIDC) then encodes degradation cues into type-level tokens and a dense spatial embedding. The analytical cue maps are shared across the two modules: RTMR uses them to assess

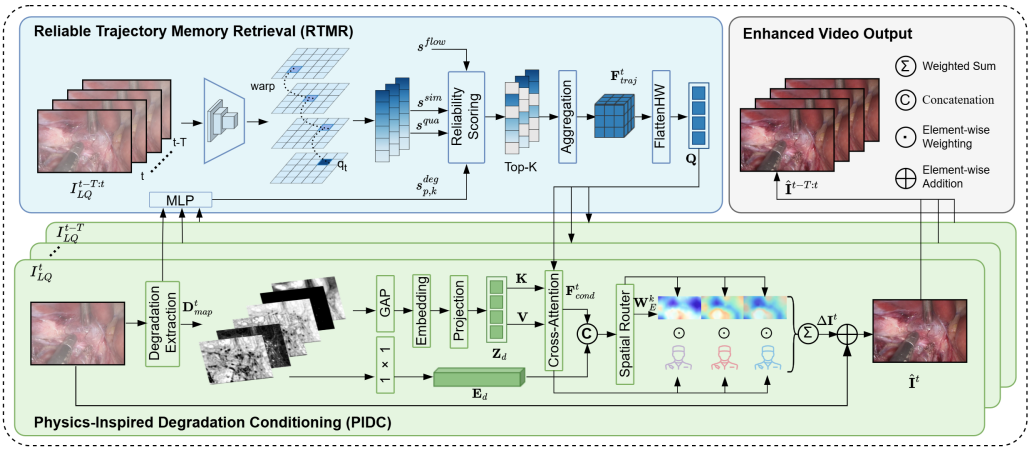


Figure 2: Overview of EndoTME. RTMR retrieves reliable historical memories, while PIDC performs degradation-aware conditioning and spatial expert routing. The reconstruction head is omitted for clarity.

the degradation risk of historical candidates, whereas PIDC encodes them for degradation-aware conditioning and spatial routing. The type-level tokens condition trajectory features through cross-attention, while the dense embedding guides spatial routing among restoration experts. A shared reconstruction head predicts a residual $\Delta\mathbf{I}^\tau$, yielding $\hat{\mathbf{I}}^\tau = \mathbf{I}^\tau_{LQ} + \Delta\mathbf{I}^\tau$.

3.2 Reliable Trajectory Memory Retrieval

Motion alignment establishes temporal correspondence but does not guarantee information reliability. RTMR therefore evaluates each trajectory-aligned historical candidate before temporal aggregation.

Trajectory-aligned candidate construction. Let $\mathbf{q}^t_p = \mathbf{F}^t(p)$ denote the current feature token at position p. The corresponding candidate from the j-th historical feature map is

$$\mathbf{m}^t_{p,j} = \mathcal{W}_F\left(\mathbf{F}^{t-j}, \pi^t_{p,j}\right), \qquad j = 1, \ldots, M, \tag{1}$$

where $\pi^t_{p,j}$ is the trajectory-guided sampling position and $\mathcal{W}_F(\cdot)$ denotes differentiable feature warping.

Multi-cue reliability scoring. RTMR evaluates each candidate using appearance similarity, local quality, flow reliability, and degradation risk. Specifically, $s^{\text{sim}}_{p,j} = \cos(\mathbf{q}^t_p, \mathbf{m}^t_{p,j})$, while $s^{\text{qua}}_{p,j} = \phi_{\text{qua}}(\mathbf{u}^{\text{qua}}_{p,j})$ and $s^{\text{flow}}_{p,j} = \phi_{\text{flow}}(\mathbf{u}^{\text{flow}}_{p,j})$ encode local-quality and trajectory-reliability statistics.

The risk of propagating degraded historical evidence is predicted by

$$s^{\text{deg}}_{p,j} = \sigma\left(\text{MLP}_{\text{deg}}\left(\text{Concat}\left[\mathbf{d}^t_p, \mathbf{d}^{\text{hist}}_{p,j}, \left|\mathbf{d}^t_p - \mathbf{d}^{\text{hist}}_{p,j}\right|\right]\right)\right), \tag{2}$$

where $\mathbf{d}^t_p$ and $\mathbf{d}^{\text{hist}}_{p,j}$ are sampled from the current and trajectory-aligned historical cue maps before PIDC encoding. The final candidate reliability is

$$r^t_{p,j} = \alpha_s s^{\text{sim}}_{p,j} + \alpha_q s^{\text{qua}}_{p,j} + \alpha_f s^{\text{flow}}_{p,j} - \alpha_d s^{\text{deg}}_{p,j}, \tag{3}$$

where the four score terms are normalized to comparable ranges and the coefficients balance content consistency, local quality, motion reliability, and degradation risk. The coefficients are fixed after selection on the paired validation split; detailed definitions are provided in the supplementary material.

Top-K retrieval and aggregation. Only the K highest-scoring candidates are retained:

$$\Omega^t_p = \text{TopK}\left(\left\{r^t_{p,j}\right\}_{j=1}^{M}, K\right), \qquad K \le M, \tag{4}$$

with normalized weights $a_{p,j}^t = \exp(r_{p,j}^t)/\sum_{l\in\Omega_p^t}\exp(r_{p,l}^t)$. The selected memories and current query are jointly aggregated:

$$\mathbf{f}_{traj,p}^t = \mathcal{A}_{traj}\left(\left[\mathbf{q}_p^t; \left\{a_{p,j}^t \mathbf{m}_{p,j}^t\right\}_{j\in\Omega_p^t}\right]\right), \tag{5}$$

where $\mathcal{A}_{traj}(\cdot)$ produces the trajectory-enhanced token. The resulting feature is folded into $\mathbf{F}_{traj}^t$ and flattened to obtain query tokens $\mathbf{Q}^t = \mathrm{Flatten}_{HW}(\mathbf{F}_{traj}^t)\mathbf{W}_Q$.

3.3 Physics-Inspired Degradation Conditioning

PIDC represents spatially heterogeneous degradations at both type and spatial levels, allowing different regions to adopt different restoration behaviors.

Physics-inspired degradation cues. PIDC extracts six cue maps $\{\mathbf{D}_m^t\}_{m=1}^6$ from the degraded frame and motion information. They characterize red-veil contamination, specular reflection, local blur, weak contrast, structural response, and flow magnitude, and serve as interpretable indicators rather than parameters of a strict physical image-formation model He et al. (2011); Li et al. (2017); Wang et al. (2024); Xu et al. (2023). Exact cue definitions and normalization are provided in the supplementary material.

Dual-granularity degradation representation. Each cue is pooled and projected into a type-level token:

$$\mathbf{Z}_d^t = \left[\mathcal{P}_1\left(\mathrm{GAP}(\mathbf{D}_1^t)\right); \ldots; \mathcal{P}_6\left(\mathrm{GAP}(\mathbf{D}_6^t)\right)\right], \tag{6}$$

which preserves degradation-type information. The cue maps are also projected into a dense spatial embedding:

$$\mathbf{E}_d^t = \mathrm{Conv}_{1\times1}\left(\mathrm{Concat}\left[\mathbf{D}_1^t, \ldots, \mathbf{D}_6^t\right]\right), \tag{7}$$

where $\mathbf{Z}_d^t$ indicates which degradation types are present and $\mathbf{E}_d^t$ preserves where they occur.

Type-level degradation interaction. The RTMR feature provides queries $\mathbf{Q}^t$, while $\mathbf{K}^t = \mathbf{Z}_d^t\mathbf{W}_K$ and $\mathbf{V}^t = \mathbf{Z}_d^t\mathbf{W}_V$. Degradation information is injected through residual cross-attention:

$$\widetilde{\mathbf{Q}}^t = \mathrm{LN}\left(\mathbf{Q}^t + \mathrm{Softmax}\left(\frac{\mathbf{Q}^t(\mathbf{K}^t)^\top}{\sqrt{d}}\right)\mathbf{V}^t\mathbf{W}_O\right), \tag{8}$$

and is reshaped into $\mathbf{F}_{cond}^t$ for spatial routing.

Spatial strategy routing and reconstruction. The conditioned feature and dense spatial embedding jointly predict three location-dependent expert weights:

$$\left\{\mathbf{W}_e^t\right\}_{e=1}^3 = \mathrm{Softmax}_e\left(\mathcal{G}\left(\mathrm{Concat}\left[\mathbf{F}_{cond}^t, \mathbf{E}_d^t\right]\right)\right), \tag{9}$$

where $\mathcal{G}(\cdot)$ denotes the spatial router and the weights sum to one across experts at each location (Shazeer et al., 2017; Zamfir et al., 2025; He et al., 2026). The expert features are spatially fused as

$$\mathbf{F}_{pidc}^t = \sum_{e=1}^{3}\mathbf{W}_e^t \odot \mathcal{E}_e\left(\mathbf{F}_{cond}^t\right), \tag{10}$$

and the shared reconstruction head predicts $\Delta\mathbf{I}^t = \mathcal{H}_{rec}(\mathbf{F}_{pidc}^t)$.

3.4 Overall Training Loss

EndoTME is optimized using reconstruction, perceptual, and temporal consistency losses:

$$\mathcal{L}_{rec} = \frac{1}{T+1}\sum_{\tau=t-T}^{t}\left\|\hat{\mathbf{I}}^\tau - \mathbf{I}_{GT}^\tau\right\|_1, \tag{11}$$

$$\mathcal{L}_{per} = \frac{1}{T+1}\sum_{\tau=t-T}^{t}\sum_{l}\left\|\Phi_l(\hat{\mathbf{I}}^\tau) - \Phi_l(\mathbf{I}_{GT}^\tau)\right\|_1, \tag{12}$$

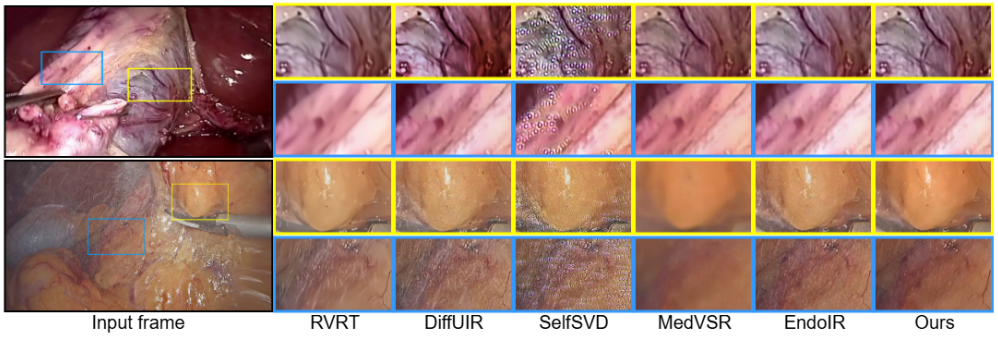


Figure 3: Qualitative comparison on paired and real-world endoscopic video datasets. The upper row is from STSVD, where the clean reference is available, and the lower row is from LSVD, where no clean reference is provided. Yellow and blue boxes indicate two regions of interest with different local degradations. Compared with existing methods, EndoTME better suppresses blood haze and specular artifacts while preserving clearer tissue boundaries and more natural local appearance.

Table 1: Quantitative comparison on the paired STSVD dataset. The best results are shown in bold.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	tWE↓	tOF↓
RVRT (Liang et al., 2022)	22.03	0.8381	0.2205	0.1868	0.1579	0.2241
DiffUIR (Zheng et al., 2024)	24.33	0.8813	0.0952	0.0908	0.1551	0.2386
SelfSVD (Wu et al., 2024)	23.86	0.8714	0.3428	0.2235	0.2793	0.2576
MedVSR (Liu et al., 2025)	20.60	0.8083	0.4047	0.3089	0.1335	0.2125
EndoIR (Chen et al., 2026)	24.61	0.8767	0.0865	0.1286	0.1996	0.2317
Ours	**25.88**	**0.8922**	**0.0840**	**0.0787**	**0.1287**	**0.1736**

where $\Phi_l(\cdot)$ denotes a fixed perceptual feature extractor.

Temporal consistency is enforced through image-domain warping:

$$\mathcal{L}_{temp} = \frac{1}{T} \sum_{\tau=t-T+1}^{t} \left\| \hat{\mathbf{I}}^{\tau} - \mathcal{W}_I \left(\hat{\mathbf{I}}^{\tau-1}, \mathbf{O}_{\tau-1\to\tau} \right) \right\|_1, \tag{13}$$

Only valid warped pixels contribute to the temporal loss; invalid boundary coordinates are excluded as detailed in the supplementary material.

The overall objective is

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{temp}\mathcal{L}_{temp}. \tag{14}$$

Additional implementation details and the reference-free adaptation objective are provided in the supplementary material.

4 Experimental Results

4.1 Experimental Setup

Datasets. We evaluate EndoTME on three endoscopic video datasets under a paired-to-real protocol. STSVD (Liang et al., 2026) contains 120 paired smoky–clean videos, each comprising 100 frames at 720×1080 resolution. LSVD (Wu et al., 2024) contains 486 real laparoscopic video clips collected from 40 patients, with 20–50 frames per clip at 1080×1920 resolution. EndoReal is an internally collected real-world surgical endoscopic video collection comprising 5,691 clips from 320 patients. Most videos are recorded at 1080×1920 or 720×1080 resolution. As frame-aligned clean references are unavailable, EndoReal is used for real-world adaptation and held-out no-reference evaluation. Detailed protocols are provided in the supplementary material.

Table 2: Quantitative comparison on real-world endoscopic video datasets without clean references.

Method	LSVD					EndoReal				
	NIQE ↓	BRISQUE ↓	CLIP-IQA ↑	tWE ↓	tOF ↓	NIQE ↓	BRISQUE ↓	CLIP-IQA ↑	tWE ↓	tOF ↓
RVRT	7.28	56.08	0.2558	0.1926	0.3450	13.19	104.26	0.4963	0.1385	0.2418
DiffUIR	6.48	42.08	0.3254	0.1413	0.3809	7.67	80.80	0.5469	0.1209	0.1962
SelfSVD	5.65	49.71	0.4473	0.3076	0.1938	7.71	57.92	0.3103	0.1447	0.1533
MedVSR	9.74	57.96	0.3654	0.1832	0.2119	9.75	91.38	0.5216	0.1117	0.1136
EndoIR	8.50	42.35	0.4510	0.1751	0.5599	9.13	41.45	0.4162	0.1678	0.5825
Ours	**5.32**	**31.95**	**0.4791**	**0.1378**	**0.1843**	**6.42**	**40.94**	**0.5511**	**0.1041**	**0.1015**

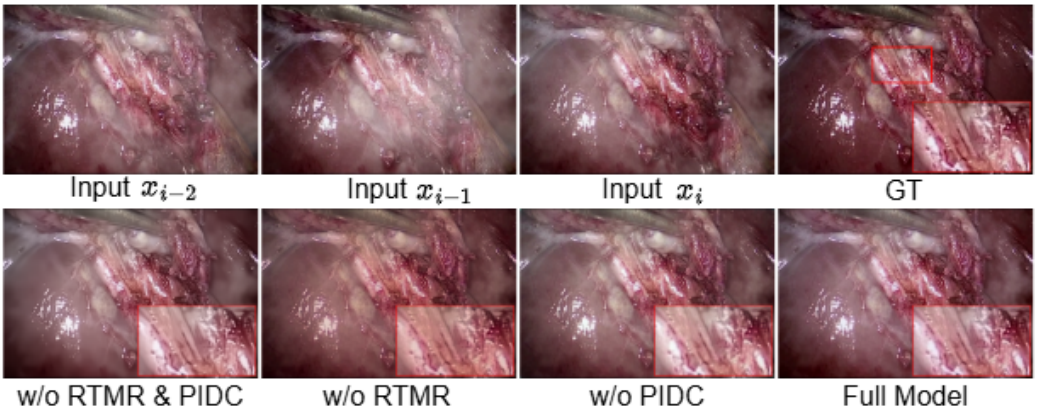


Figure 4: Visual comparison of the component ablation study. The first three images show consecutive frames from the input clip, while the remaining panels present the ground truth and restoration results for the target frame x_i. Enlarged regions are provided for closer comparison.

Evaluation Metrics. We report PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), NIQE (Mittal et al., 2013), BRISQUE (Mittal et al., 2012), CLIP-IQA (Wang et al., 2023b), and the temporal metrics tWE and tOF (Bonneel et al., 2015; Chu et al., 2020).

Implementation Details. EndoTME is implemented in PyTorch and optimized using Adam with an initial learning rate of 1×10^{-4} and a cosine decay schedule. Random cropping and flipping are applied during training. Public datasets follow their official splits, while EndoReal is partitioned at the patient level to ensure that no patient appears in more than one split. For real-world evaluation, EndoTME uses only the designated adaptation and validation splits for adaptation and checkpoint selection; held-out test clips are not used for model selection. All methods are evaluated using the same input protocol and metric implementations on NVIDIA RTX 3090 GPUs. Additional optimization, baseline, and adaptation details are provided in the supplementary material.

4.2 Comparison with State-of-the-Art Methods

Compared Methods. We compare EndoTME with representative image and video restoration methods, including RVRT (Liang et al., 2022), DiffUIR (Zheng et al., 2024), SelfSVD (Wu et al., 2024), MedVSR (Liu et al., 2025), and EndoIR (Chen et al., 2026). All restored sequences are evaluated using the same evaluation pipeline.

Quantitative Comparison. Tables 1 and 2 report the quantitative comparisons. On paired endoscopic videos, EndoTME achieves the best results across all six metrics. In particular, it obtains a PSNR of 25.88 dB, outperforming the second-best result by 1.27 dB, while improving SSIM from

Table 3: Module-level factorial ablation on STSVD. RTMR and PIDC are independently enabled or disabled, and the last row corresponds to the full EndoTME model.

RTMR	PIDC	PSNR↑	SSIM↑	LPIPS↓	tWE↓
×	×	23.12	0.8576	0.1017	0.1749
√	×	24.34	0.8808	0.0878	0.1408
×	√	24.25	0.8770	0.0854	0.1537
√	√	**25.88**	**0.8922**	**0.0840**	**0.1287**

Table 4: Internal ablation of RTMR on STSVD with PIDC enabled and fixed. Reliability cues are cumulatively incorporated from trajectory alignment to degradation-risk assessment.

RTMR Configuration	PSNR↑	SSIM↑	tWE↓	tOF↓
Trajectory Alignment Only	24.35	0.8776	0.1524	0.2304
+ Feature Similarity	24.71	0.8793	0.1491	0.2236
+ Local Quality	24.63	0.8831	0.1376	0.1954
+ Flow Reliability	25.02	0.8865	0.1329	0.1948
+ Degradation Risk (Full RTMR)	**25.88**	**0.8922**	**0.1287**	**0.1736**

0.8813 to 0.8922. EndoTME also reduces DISTS from 0.0908 to 0.0787. The lowest tWE and tOF further indicate that the enhanced frames maintain more consistent temporal appearance.

On real-world videos without clean references, EndoTME obtains the best reported values across all metrics under the evaluation protocol used here. On LSVD, it reduces BRISQUE from 42.08 to 31.95 and improves CLIP-IQA from 0.4510 to 0.4791. On EndoReal, it obtains a NIQE of 6.42, a CLIP-IQA of 0.5511, a tWE of 0.1041, and a tOF of 0.1015. These results indicate improved visual quality and temporal stability under the reported real-world evaluation protocol.

Qualitative Comparison. Figure 3 presents representative visual comparisons. RVRT and MedVSR tend to produce over-smoothed local structures, while SelfSVD introduces noticeable patterned artifacts in severely degraded regions. DiffUIR and EndoIR recover part of the tissue appearance but leave residual degradation or weakened local boundaries. In contrast, EndoTME more effectively suppresses blood-related veiling and specular artifacts while preserving fine tissue structures, vascular details, and natural color appearance. The enlarged regions further demonstrate its advantage in restoring spatially varying degradations without excessive smoothing.

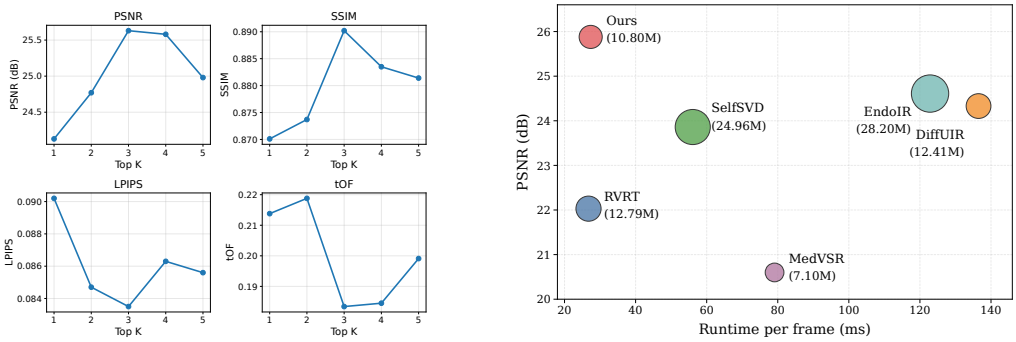
4.3 Ablation Study

Effectiveness of RTMR. Table 3 shows that RTMR improves PSNR from 23.12 to 24.34 dB and reduces tWE from 0.1749 to 0.1408, showing the benefit of incorporating reliability-aware temporal evidence selection. Table 4 further shows that the four reliability cues provide complementary benefits with metric-specific trade-offs, while their complete combination achieves the strongest overall performance. This confirms that motion alignment alone is insufficient for reliable memory retrieval. As illustrated in Fig. 4, removing RTMR weakens local detail recovery in the selected region, whereas the full model produces a result closer to the clean reference.

Effectiveness of PIDC. Table 3 shows that PIDC improves PSNR from 23.12 to 24.25 dB and reduces LPIPS from 0.1017 to 0.0854, indicating the effectiveness of degradation-aware local restoration. Table 5 further demonstrates that replacing degradation-derived tokens with learnable tokens or removing either the type-level tokens or dense spatial embedding degrades performance, validating the complementary roles of the two representations. The visual comparison in Fig. 4 further shows that removing PIDC results in less effective local degradation removal and weaker structural recovery. Combining RTMR and PIDC achieves the best PSNR, SSIM, and tWE, while producing the visual result closest to the reference.

Table 5: Internal ablation of PIDC on STSVD with RTMR enabled and fixed. We compare the degradation-derived representation with learnable-token conditioning and variants removing the type-level or dense spatial branch.

PIDC Configuration	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
Learnable Tokens	24.71	0.8842	0.0863	0.0834
w/o Type-Level Tokens	24.97	0.8857	0.0855	0.0789
w/o Dense Spatial Embedding	25.46	0.8893	0.0851	0.0812
Full PIDC	**25.88**	**0.8922**	**0.0840**	**0.0787**



(a) Top-K sensitivity analysis.

(b) Performance–efficiency analysis.

Figure 5: Further analysis of EndoTME. (a) Sensitivity to the number of retrieved memories K, with all other components fixed to the full model. (b) Accuracy–efficiency comparison on STSVD.

4.4 Further Analysis

Sensitivity to Top-K. Figure 5a evaluates K from 1 to 5. Performance improves as K increases to 3, but decreases when additional memories are retained, indicating that excessive candidates may introduce redundant or unreliable information. We therefore set $K = 3$, which achieves the best results across all reported metrics.

Complexity Analysis. Figure 5b compares PSNR, per-frame runtime, and parameter count. EndoTME achieves the highest PSNR of 25.88 dB with 10.80M parameters and a runtime of 27.36 ms/frame, corresponding to the second-smallest model size and second-fastest inference speed. Compared with the fastest method, RVRT, EndoTME introduces only 0.61 ms/frame additional latency while improving PSNR by 3.85 dB, demonstrating a favorable accuracy–efficiency trade-off.

5 Conclusion

In this paper, we proposed **EndoTME**, a physics-inspired degradation-aware trajectory memory framework for endoscopic video enhancement. Motivated by unreliable temporal propagation and insufficient degradation awareness in complex surgical videos, EndoTME introduces two core designs. First, Reliable Trajectory Memory Retrieval selectively retrieves trustworthy historical features and suppresses unreliable temporal cues caused by severe degradation or unstable motion. Second, Physics-Inspired Degradation Conditioning explicitly characterizes endoscopy-specific visual artifacts and adapts the restoration process to spatially heterogeneous degradation patterns. By encoding interpretable physics-inspired degradation cues at both type and spatial levels, PIDC connects local degradation distributions with region-specific restoration behaviors. Experimental results show that EndoTME improves restoration quality and temporal consistency over representative image and video restoration baselines.

AI USE STATEMENT

During the preparation of this manuscript, the authors used AI-assisted tools for language refinement, code assistance, and figure-preparation support. All AI-assisted outputs were reviewed and verified by the authors, who remain fully responsible for the scientific content, methodology, experimental results, and conclusions presented in this work.

ETHICS STATEMENT

This work uses both publicly available endoscopic datasets and an internally collected real-world dataset, EndoReal. The collection and use of EndoReal followed the applicable institutional ethical and data-governance requirements. Only de-identified patient and procedure identifiers were used for patient-disjoint split construction; no direct patient identifiers were available to the modeling pipeline. The public datasets were used in accordance with their respective licenses and usage policies.

REPRODUCIBILITY STATEMENT

We provide descriptions of the model architecture, RTMR and PIDC components, degradation-cue construction, training objectives, data splits, evaluation protocols, and implementation settings in the main paper and supplementary material. The supplementary material further documents the training and adaptation protocol, internal ablation configurations, dataset splits, and runtime-measurement procedure used in the reported experiments.

REFERENCES

Yuang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, and Ran He. Multimodal prompt perceiver: Empower adaptiveness, generalizability and fidelity for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25432–25444, 2024.

Sabri Bolkar, Congcong Wang, Faouzi Alaya Cheikh, and Sule Yildirim. Deep smoke removal from minimally invasive surgery videos. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 3403–3407, 2018. doi: 10.1109/ICIP.2018.8451815.

Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *ACM Transactions on Graphics*, 34(6):196:1–196:9, 2015. doi: 10.1145/2816795.2818107.

Kelvin C. K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4947–4956, 2021.

Kelvin C. K. 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 (CVPR)*, pp. 5972–5981, 2022a.

Kelvin C. K. 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 (CVPR)*, pp. 5962–5971, 2022b.

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 *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15006 of *Lecture Notes in Computer Science*, pp. 369–379. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72089-5_35.

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.

Proceedings of the AAAI Conference on Artificial Intelligence, 40(4):3047–3055, 2026. doi: 10.1609/aaai.v40i4.37297.

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*, 39(4):75:1–75:13, 2020. doi: 10.1145/3386569.3392457.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011. doi: 10.1109/TPAMI.2010.168.

Xiao He, Zhijun Tu, Kun Cheng, Mingrui Zhu, Jie Hu, Nannan Wang, and Xinbo Gao. Mixture of ranks with degradation-aware routing for one-step real-world image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4681–4689, 2026. doi: 10.1609/aaai.v40i6.42469.

Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8008–8017, 2020.

Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. AOD-Net: All-in-one dehazing network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4770–4778, 2017. doi: 10.1109/ICCV.2017.511.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17452–17462, 2022.

Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. In *Advances in Neural Information Processing Systems*, volume 35, pp. 378–393, 2022.

Qifan Liang, Junlin Li, Zhen Han, Xihao Wang, Zhongyuan Wang, and Bin Mei. Rethinking surgical smoke: A smoke-type-aware laparoscopic video desmoking method and dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(9):6844–6852, 2026. doi: 10.1609/aaai.v40i9.37617.

Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5687–5696, 2022.

Xinyu Liu, Guolei Sun, Cheng Wang, Yixuan Yuan, and Ender Konukoglu. MedVSR: Medical video super-resolution with cross state-space propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11697–11707, 2025.

Wenyang Luo, Haina Qin, Zewen Chen, Libin Wang, Dandan Zheng, Yuming Li, Yufan Liu, Bing Li, and Weiming Hu. Visual-instructed degradation diffusion for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12764–12777, 2025.

Anish Mittal, Anush Krishna Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

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. doi: 10.1007/s11548-022-02595-2.

Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5815–5824, 2023.

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 (CVPR)*, pp. 37413–37422, 2026.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H. Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71275–71293, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. TransWeather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2353–2363, 2022. doi: 10.1109/CVPR52688.2022.00239.

Cong Wang, Jinshan Pan, Wei Wang, Jiangxin Dong, Mengzhu Wang, Yakun Ju, and Junyang Chen. PromptRestorer: A prompting image restoration method with degradation perception. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8898–8912, 2023a.

Congcong Wang, Meng Zhao, Chengguang Zhou, Nanqing Dong, Zohaib Amjad Khan, Xintong Zhao, Faouzi Alaya Cheikh, Azeddine Beghdadi, and Shengyong Chen. Smoke veil prior regularized surgical field desmoking without paired in-vivo data. *Computers in Biology and Medicine*, 168:107761, 2024. doi: 10.1016/j.compbiomed.2023.107761.

Jianyi Wang, Kelvin C. K. 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, 2023b. doi: 10.1609/aaai.v37i2.25353.

Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1954–1963, 2019. doi: 10.1109/CVPRW.2019.00247.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Renlong Wu, Zhilu Zhang, Shuohao Zhang, Longfei Gou, Haobin Chen, Lei Zhang, Hao Chen, and Wangmeng Zuo. Self-supervised video desmoking for laparoscopic surgery. In *Computer Vision – ECCV 2024*, volume 15130 of *Lecture Notes in Computer Science*, pp. 307–324. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73220-1_18.

Handing Xu, Zhenguo Nie, Tairan Peng, Huimin Pan, and Xin-Jun Liu. Real-time endoscopic video enhancement via degradation representation estimation and propagation. *Journal of Imaging*, 12 (3):134, 2026. doi: 10.3390/jimaging12030134.

Jiaqi Xu, Xiaowei Hu, Lei Zhu, Qi Dou, Jifeng Dai, Yu Qiao, and Pheng-Ann Heng. Video dehazing via a multi-range temporal alignment network with physical prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18053–18062, 2023.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12753–12763, 2025.

Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, and Feng Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5825–5835, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25445–25455, 2024.

SUPPLEMENTARY MATERIAL

A SUPPLEMENTARY OVERVIEW

This supplementary material provides additional methodological and experimental details for EndoTME. Section B describes Reliable Trajectory Memory Retrieval (RTMR), including trajectory construction, the four reliability factors, Top-K retrieval, and temporal aggregation. Section C details Physics-Inspired Degradation Conditioning (PIDC), including the six analytical degradation cues, dual-granularity representation, residual cross-attention, and spatial expert routing. Sections D–F describe the training and adaptation objectives, datasets, ethical considerations, preprocessing, baseline settings, metric implementations, and runtime measurement. Sections G–I provide ablation clarifications, additional qualitative results, and failure-case analysis.

B COMPLETE DETAILS OF RTMR

B.1 TRAJECTORY CONSTRUCTION

Given a degraded clip $\mathbf{I}_{LQ}^{t-T:t} = \{\mathbf{I}_{LQ}^{t-T}, \ldots, \mathbf{I}_{LQ}^{t}\}$, the shared encoder extracts $\{\mathbf{F}^{t-T}, \ldots, \mathbf{F}^{t}\}$, where $\mathbf{F}^{\tau} \in \mathrm{R}^{C \times H_f \times W_f}$. Optical flow is estimated between adjacent degraded frames and is resized to the feature resolution. Both temporal directions are available to the propagation framework, while RTMR retrieves only preceding observations for the current time step.

Let $\mathbf{G}_0$ be the regular coordinate grid at feature resolution. The sampling position in a historical frame is obtained by recursively following adjacent motion:

$$\mathbf{G}_j^t = \mathcal{W}_G\left(\mathbf{G}_{j-1}^t, \downarrow \mathbf{O}_j^t\right) + \mathbf{G}_0, \qquad j = 1, \ldots, M, \tag{15}$$

where $\downarrow \mathbf{O}_j^t$ is the flow field resized to $H_f \times W_f$ and $\mathcal{W}_G(\cdot)$ denotes grid warping. The trajectory-guided sampling position $\pi_{p,j}^t$ in Eq. 1 of the main paper is the coordinate read from $\mathbf{G}_j^t$ at location p.

Historical features and local descriptors are sampled with the same trajectory coordinates. Differentiable bilinear interpolation is used for feature sampling, while boundary coordinates are constrained to valid border values. This shared sampling rule ensures that the appearance feature, quality descriptor, flow descriptor, and degradation descriptor of one candidate refer to the same historical observation.

The current frame is not included in the historical candidate set. For early positions in a clip, the number of candidates is the number of available preceding frames. Consequently, RTMR does not create duplicated memories merely to maintain a fixed candidate count. Let $N = H_f W_f$. After flattening the spatial dimensions, the current feature contains N query tokens and each historical frame contributes one trajectory-aligned candidate per query position.

B.2 RELIABILITY DESCRIPTOR DEFINITIONS

RTMR separates geometric correspondence from information reliability. Each candidate is evaluated using appearance similarity, local quality, flow reliability, and degradation risk.

Appearance similarity. For the current query $\mathbf{q}_p^t$ and historical candidate $\mathbf{m}_{p,j}^t$, RTMR uses normalized cosine similarity:

$$s_{p,j}^{\mathrm{sim}} = \left\langle \frac{\mathbf{q}_p^t}{\|\mathbf{q}_p^t\|_2 + \epsilon}, \frac{\mathbf{m}_{p,j}^t}{\|\mathbf{m}_{p,j}^t\|_2 + \epsilon} \right\rangle. \tag{16}$$

This term verifies content correspondence but does not by itself indicate whether the historical region is sufficiently clear.

Local quality. The local quality descriptor is formed from interpretable statistics associated with endoscopic visibility:

$$\mathbf{u}_{p,j}^{\mathrm{qua}} = \left[q_{p,j}^{\mathrm{sharp}}, q_{p,j}^{\mathrm{cont}}, q_{p,j}^{\mathrm{red}}, q_{p,j}^{\mathrm{spec}}\right], \tag{17}$$

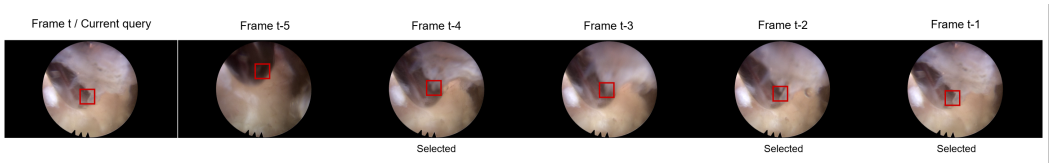


Figure 6: Visualization of Reliable Trajectory Memory Retrieval. For each current query region, trajectory-aligned historical candidates are shown. Historical candidates selected by Top-K retrieval are marked below the corresponding frames. RTMR rejects candidates that are geometrically aligned but severely degraded or motion-unstable.

where the four entries describe local sharpness, local contrast, excessive redness, and specular saturation. All statistics are computed on the degraded RGB frames and sampled at the trajectory-aligned historical position. The bounded mapping $\phi_{\mathrm{qua}}(\cdot)$ rewards sharp and contrasted candidates and suppresses candidates dominated by redness or reflection:

$$s_{p,j}^{\mathrm{qua}} = \phi_{\mathrm{qua}}\left(\mathbf{u}_{p,j}^{\mathrm{qua}}\right). \tag{18}$$

Thus, a content-consistent candidate can still receive a low quality score when its observable evidence is degraded.

Flow reliability. The flow descriptor summarizes the displacement and local stability of the trajectory:

$$\mathbf{u}_{p,j}^{\mathrm{flow}} = \left[\|\mathbf{O}_{p,j}^{t}\|_2, v_{p,j}^{\mathrm{traj}}\right], \tag{19}$$

where $\|\mathbf{O}_{p,j}^{t}\|_2$ denotes local flow magnitude and $v_{p,j}^{\mathrm{traj}}$ denotes the variation of the trajectory in a local neighborhood. The bounded mapping $\phi_{\mathrm{flow}}(\cdot)$ assigns lower confidence to unstable or extreme motion:

$$s_{p,j}^{\mathrm{flow}} = \phi_{\mathrm{flow}}\left(\mathbf{u}_{p,j}^{\mathrm{flow}}\right). \tag{20}$$

Degradation risk. The six cue maps introduced in Section C are computed before PIDC tokenization. At each current position and its trajectory-aligned historical position, we collect

$$\mathbf{d}_{p}^{t} = \left[\mathbf{D}_{1}^{t}(p), \ldots, \mathbf{D}_{6}^{t}(p)\right], \tag{21}$$

and

$$\mathbf{d}_{p,j}^{\mathrm{hist}} = \left[\mathbf{D}_{1}^{t-j}(\pi_{p,j}^{t}), \ldots, \mathbf{D}_{6}^{t-j}(\pi_{p,j}^{t})\right]. \tag{22}$$

The pairwise descriptor contains the current state, historical state, and their absolute difference:

$$\mathbf{u}_{p,j}^{\mathrm{deg}} = \mathrm{Concat}\left[\mathbf{d}_{p}^{t}, \mathbf{d}_{p,j}^{\mathrm{hist}}, \left|\mathbf{d}_{p}^{t} - \mathbf{d}_{p,j}^{\mathrm{hist}}\right|\right]. \tag{23}$$

A lightweight MLP followed by a sigmoid predicts $s_{p,j}^{\mathrm{deg}} \in [0,1]$. A high value indicates that the historical candidate is likely to propagate severe or mismatched degradation. The cue maps are used here before PIDC encoding; therefore, RTMR does not depend on the conditioned feature produced by PIDC.

Combined reliability. The final score follows Eq. 3 of the main paper:

$$r_{p,j}^{t} = \alpha_s s_{p,j}^{\mathrm{sim}} + \alpha_q s_{p,j}^{\mathrm{qua}} + \alpha_f s_{p,j}^{\mathrm{flow}} - \alpha_d s_{p,j}^{\mathrm{deg}}. \tag{24}$$

The four terms are normalized to comparable ranges before fusion. The same coefficients are used for all datasets and remain fixed after selection on the paired validation split. This design avoids dataset-specific reliability rules and preserves the interpretation that a candidate should be content-consistent, locally informative, motion-reliable, and minimally degraded.

B.3 Top-K Retrieval and Temporal Aggregation

RTMR performs Top-K retrieval independently at each spatial query:

$$\Omega_p^t = \mathrm{TopK}\left(\{r_{p,j}^t\}_{j=1}^M, K\right). \tag{25}$$

The selected scores are normalized only within Ω_p^t:

$$a_{p,j}^t = \frac{\exp(r_{p,j}^t)}{\sum_{l\in\Omega_p^t}\exp(r_{p,l}^t)}, \qquad j \in \Omega_p^t. \tag{26}$$

The default is $K = 3$, following the sensitivity analysis in Fig. 5(a) of the main paper.

The current query is always retained. The selected historical features are weighted by $a_{p,j}^t$ and arranged in temporal order before aggregation. The aggregation module is a shifted temporal state-space block that jointly models the current token and the retrieved memories:

$$\mathbf{h}_p^t = \mathcal{A}_{\mathrm{traj}}\left(\left[\mathbf{q}_p^t; \{a_{p,j}^t \mathbf{m}_{p,j}^t\}_{j\in\Omega_p^t}\right]\right). \tag{27}$$

A learnable residual gate preserves the current representation while injecting complementary temporal evidence:

$$\mathbf{f}_{\mathrm{traj},p}^t = \mathbf{q}_p^t + \gamma \mathbf{h}_p^t, \tag{28}$$

where γ is optimized jointly with the restoration network. Layer normalization is used inside the temporal aggregation block. The resulting tokens are folded into $\mathbf{F}_{\mathrm{traj}}^t$ and projected to the spatial queries used by PIDC.

The hard Top-K operation is identical during training and inference. Gradients are propagated through the selected candidate features, their normalized weights, and the aggregation module. Candidates that are not selected at a particular query do not contribute to that query's update. The ablation in Table 4 of the main paper demonstrates that the four reliability factors are complementary despite metric-specific trade-offs.

B.4 RTMR Retrieval Visualization

Figure 6 visualizes the internal retrieval process of RTMR. For each current query region, historical candidates are sampled from their trajectory-aligned positions and evaluated using appearance similarity, local quality, flow reliability, and degradation risk. The final reliability score is obtained by combining these complementary factors, after which only the Top-K candidates are retained for temporal aggregation.

The visualization highlights that geometric correspondence alone does not guarantee useful historical evidence. Some candidates remain well aligned with the current region but contain severe veiling, specular saturation, blur, or weak structural information. Conversely, a visually clear candidate may still be rejected when its trajectory is motion-unstable. RTMR therefore favors candidates that are both content-consistent and sufficiently reliable, reducing the propagation of corrupted historical information.

C Complete Details of PIDC

C.1 Definitions of the Six Degradation Cue Maps

PIDC extracts six analytical maps from the degraded frame and its motion information. Each cue is an interpretable visual indicator rather than a parameter of a strict endoscopic image-formation model. For a spatial map $\mathbf{X}$, min–max normalization is

$$\mathcal{N}(\mathbf{X}) = \frac{\mathbf{X} - \min(\mathbf{X})}{\max(\mathbf{X}) - \min(\mathbf{X}) + \epsilon}, \qquad \epsilon = 10^{-6}. \tag{29}$$

Red-veil contamination. Let R, G, and B be the RGB channels, and let $\sigma_7(\mathbf{Y})$ denote the local standard deviation of the grayscale image in a 7×7 neighborhood. The red-veil cue is

$$\mathbf{D}_{red} = \mathcal{N}\left(\frac{R}{G+B+\epsilon}\left[1-\mathcal{N}\left(\sigma_7(\mathbf{Y})\right)\right]\right). \tag{30}$$

The response increases in reddish regions with weak local contrast, which commonly occur under blood-related veiling.

Specular reflection. Let $\mathbf{V} = \max(R, G, B)$ and let $\mathbf{S}$ denote the channel saturation. The reflection cue is

$$\mathbf{D}_{spec} = \mathcal{N}\left(\mathbf{1}[\mathbf{V} > 0.9]\,\mathbf{1}[\mathbf{S} < 0.3]\,\mathbf{V}\right). \tag{31}$$

The hard threshold provides a conservative analytical response to bright, weakly saturated highlights. Since the indicator operation is non-differentiable, gradients are not propagated through this thresholding step.

Local blur. A 3×3 Laplacian operator measures local high-frequency response:

$$\mathbf{D}_{blur} = 1 - \mathcal{N}\left(|\nabla^2 \mathbf{Y}|\right). \tag{32}$$

A larger value represents a weaker local Laplacian response and hence a stronger blur indication.

Weak contrast. The low-contrast cue is defined by

$$\mathbf{D}_{cont} = 1 - \mathcal{N}\left(\sigma_7(\mathbf{Y})\right). \tag{33}$$

Structural-response cue. Using 3×3 Sobel filters, the observable structural response is

$$\mathbf{D}_{str} = \mathcal{N}\left(\sqrt{(\nabla_x \mathbf{Y})^2 + (\nabla_y \mathbf{Y})^2}\right). \tag{34}$$

A high value indicates strong observable boundary evidence, whereas a low response indicates stronger local structural attenuation.

Flow-magnitude cue. The motion cue uses normalized local flow magnitude:

$$\mathbf{D}^{\tau}_{flow} = \mathcal{N}\left(\|\mathbf{O}^{\tau-1\to\tau}\|_2\right). \tag{35}$$

The first frame has no preceding flow and is assigned a zero map. This cue does not estimate optical-flow uncertainty directly. Instead, it provides an explicit motion-intensity signal to the conditioning and routing modules, allowing the restoration process to adapt to regions with relatively large displacement.

All maps are resized to $H_f \times W_f$ by bilinear interpolation, clamped to $[0, 1]$, and stacked as $\mathbf{D}^t_{map} \in \mathrm{R}^{6\times H_f \times W_f}$. The continuous cue constructions are differentiable with respect to their inputs, whereas the thresholded specular cue is treated as a non-differentiable analytical indicator.

C.2 Dual-Granularity Degradation Representation

PIDC represents the six maps at type and spatial levels. For the m-th cue, global average pooling gives

$$g^t_m = \mathrm{GAP}\left(\mathbf{D}^t_m\right). \tag{36}$$

Each cue has an independent learnable type embedding $\mathbf{e}_m \in \mathrm{R}^d$. Its token is

$$\mathbf{z}^t_m = g^t_m \mathbf{e}_m, \qquad m = 1, \ldots, 6. \tag{37}$$

Stacking the six tokens yields $\mathbf{Z}^t_d \in \mathrm{R}^{6\times d}$. This implementation corresponds to the cue-specific projection $\mathcal{P}_m(\cdot)$ in Eq. 6 of the main paper.

The dense representation is

$$\mathbf{E}^t_d = \mathrm{Conv}_{1\times 1}\left(\mathrm{Concat}[\mathbf{D}^t_1, \ldots, \mathbf{D}^t_6]\right). \tag{38}$$

The two representations are complementary: $\mathbf{Z}^t_d$ preserves which degradation types are present, whereas $\mathbf{E}^t_d$ preserves where they occur. The ablation in Table 5 of the main paper confirms that replacing degradation-derived tokens or removing either representation reduces overall restoration performance.

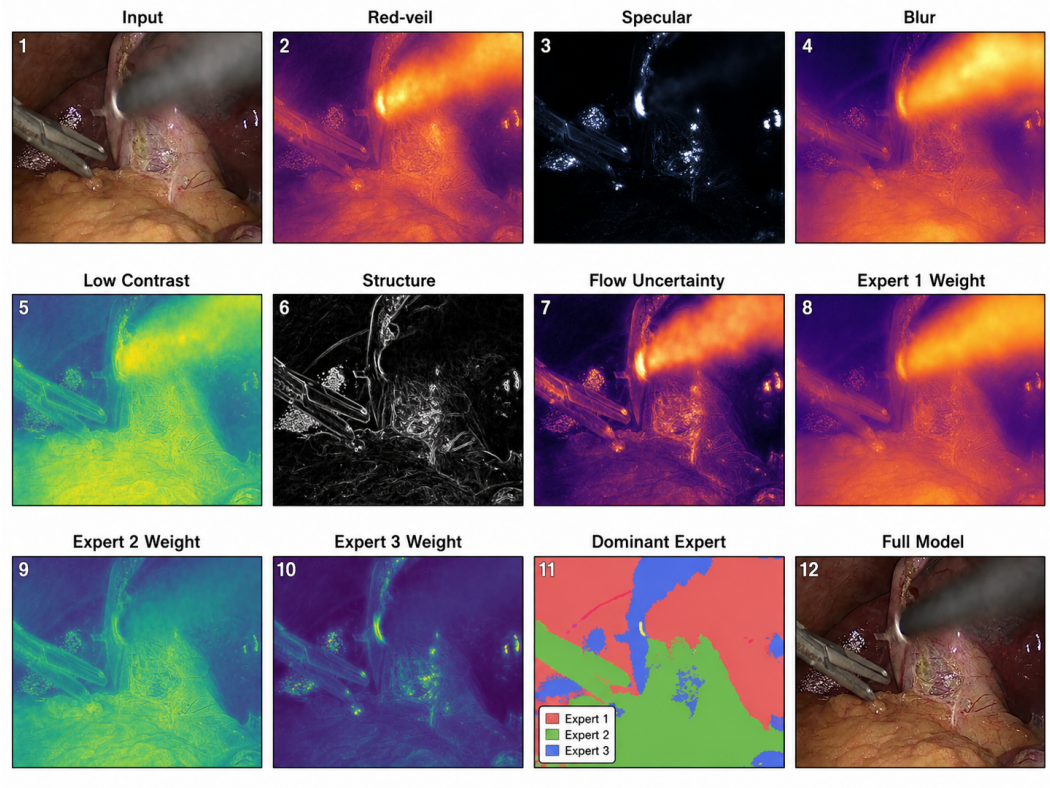


Figure 7: Visualization of Physics-Inspired Degradation Conditioning. The degradation cue maps describe complementary spatial patterns associated with red veiling, specular reflection, blur, low contrast, structural response, and flow uncertainty. The three expert-weight maps and dominant-expert assignment show that PIDC applies spatially varying restoration behaviors rather than a uniform strategy over the entire frame.

C.3 Residual Cross-Attention

The trajectory feature is flattened and projected to $\mathbf{Q}^t \in \mathrm{R}^{H_f W_f \times d}$. The six type tokens are projected to $\mathbf{K}^t, \mathbf{V}^t \in \mathrm{R}^{6\times d}$. The compact single-head expression in Eq. 8 of the main paper is implemented as four-head cross-attention. For head h,

$$\mathbf{A}_h^t = \mathrm{Softmax}\left(\frac{\mathbf{Q}_h^t(\mathbf{K}_h^t)^\top}{\sqrt{d_h}}\right)\mathbf{V}_h^t. \tag{39}$$

The head outputs are concatenated and projected:

$$\widetilde{\mathbf{Q}}^t = \mathrm{LN}\left(\mathbf{Q}^t + \mathrm{Concat}[\mathbf{A}_1^t, \ldots, \mathbf{A}_4^t]\mathbf{W}_O\right). \tag{40}$$

A two-layer feed-forward network with GELU activation is applied after the attention block. No attention dropout is used. The residual path retains trajectory information, while attention injects the degradation-type composition inferred from the current frame.

C.4 Spatial Router and Restoration Experts

The conditioned feature and dense degradation embedding are first fused to obtain the routing feature:

$$\mathbf{H}_{route}^t = \mathrm{GELU}\left(\mathrm{Conv}_{1\times1}^{(1)}\left(\mathrm{Concat}\left[\mathbf{F}_{cond}^t, \mathbf{E}_d^t\right]\right)\right), \tag{41}$$

where $\mathbf{H}_{route}^t$ denotes the intermediate routing representation. The location-dependent expert weights are then predicted as

$$\left\{\mathbf{W}_e^t\right\}_{e=1}^3 = \mathrm{Softmax}_e\left(\mathrm{Conv}_{1\times1}^{(2)}\left(\mathbf{H}_{route}^t\right)\right). \tag{42}$$

The softmax operation is applied along the expert dimension, ensuring that the three routing weights sum to one at every spatial location.

The experts operate on intermediate features rather than producing independent RGB outputs.

Structure-detail expert. This expert uses an edge-guided sigmoid gate and depthwise detail blocks to emphasize recoverable local boundaries without uniformly sharpening the entire frame.

Veil-reflection suppression expert. This expert combines dilated convolutions with channel-wise gating. The enlarged receptive field supports removal of spatially extended veiling, while the gate moderates responses in saturated regions.

Temporal-stability expert. This expert combines temporal three-dimensional convolution, depthwise spatial processing, and a spatio-temporal gate. It favors temporally stable refinement when local motion or visibility is uncertain.

The three expert outputs are fused densely:

$$\mathbf{F}_{pidc}^{t} = \sum_{e=1}^{3} \mathbf{W}_{e}^{t} \odot \mathcal{E}_{e}\left(\mathbf{F}_{cond}^{t}\right). \tag{43}$$

All experts are evaluated for each frame, and routing remains a soft spatial mixture. This avoids discontinuities associated with hard expert assignment and permits smooth transitions between neighboring degradation regions. The shared reconstruction head maps the fused feature to $\Delta\mathbf{I}^{t}$.

C.5 Cue and Routing Visualization

Figure 7 visualizes the degradation cues and spatial routing behavior of PIDC. The six analytical maps describe complementary aspects of the degraded frame. The red-veil, specular, blur, low-contrast, structural, and flow-uncertainty cues respond to different spatial patterns rather than producing a single uniform degradation representation.

The expert-weight maps further show that PIDC does not assign one restoration strategy to the entire frame. Instead, the spatial router predicts location-dependent weights according to both the degradation-conditioned feature and the dense cue embedding. Regions affected by diffuse veiling, weak structure, localized reflection, or motion uncertainty consequently receive different mixtures of expert responses. The dominant-expert map reports the expert with the largest weight at each spatial position and provides an intuitive summary of this region-dependent routing behavior.

D Training and Adaptation Objectives

D.1 Supervised Training Objective

For paired training, EndoTME uses the same objective as Eqs. 11–14 of the main paper:

$$\mathcal{L}_{rec} = \frac{1}{T+1} \sum_{\tau=t-T}^{t} \|\hat{\mathbf{I}}^{\tau} - \mathbf{I}_{GT}^{\tau}\|_{1}, \tag{44}$$

$$\mathcal{L}_{per} = \frac{1}{T+1} \sum_{\tau=t-T}^{t} \sum_{l} \|\Phi_{l}(\hat{\mathbf{I}}^{\tau}) - \Phi_{l}(\mathbf{I}_{GT}^{\tau})\|_{1}, \tag{45}$$

and

$$\mathcal{L}_{temp} = \frac{1}{T} \sum_{\tau=t-T+1}^{t} \|\hat{\mathbf{I}}^{\tau} - \mathcal{W}_{I}(\hat{\mathbf{I}}^{\tau-1}, \mathbf{O}_{\tau-1\to\tau})\|_{1}. \tag{46}$$

The perceptual extractor Φ is fixed throughout training. The same feature layers are used for every experiment. The temporal loss is computed for adjacent enhanced frames using the flow estimator shared by the temporal pipeline. Invalid warped coordinates are excluded from the loss.

The supervised objective is

$$\mathcal{L}_{sup} = \mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{temp}\mathcal{L}_{temp}. \tag{47}$$

The loss weights are selected on the paired validation split and fixed for all subsequent experiments. All terms are normalized by the number of frames and valid pixels, preventing the loss scale from changing with clip length.

D.2 Reference-Free Adaptation Objective

For LSVD and EndoReal, frame-aligned clean references are unavailable. Starting from the paired-data checkpoint, we perform dataset-specific reference-free adaptation using only each real dataset's adaptation split. The reconstruction and perceptual terms that require clean targets are removed.

The adaptation objective combines temporal consistency, structural preservation, and conservative residual regularization:

$$\mathcal{L}_{adapt} = \lambda^a_{temp}\mathcal{L}^a_{temp} + \lambda_{str}\mathcal{L}_{str} + \lambda_{reg}\mathcal{L}_{reg}. \tag{48}$$

The reference-free temporal term follows the image-domain warping principle used in supervised training:

$$\mathcal{L}^a_{temp} = \frac{1}{T}\sum_{\tau=t-T+1}^{t} \|\hat{\mathbf{I}}^{\tau} - \mathcal{W}_I(\hat{\mathbf{I}}^{\tau-1}, \mathbf{O}_{\tau-1\to\tau})\|_1. \tag{49}$$

Structural preservation constrains the enhanced frame to retain observable input boundaries:

$$\mathcal{L}_{str} = \frac{1}{T+1}\sum_{\tau=t-T}^{t} \|\nabla\hat{\mathbf{I}}^{\tau} - \nabla\mathbf{I}^{\tau}_{LQ}\|_1, \tag{50}$$

where ∇ is the spatial gradient operator. Residual regularization prevents unconstrained enhancement:

$$\mathcal{L}_{reg} = \frac{1}{T+1}\sum_{\tau=t-T}^{t} \|\hat{\mathbf{I}}^{\tau} - \mathbf{I}^{\tau}_{LQ}\|_1. \tag{51}$$

The three terms have complementary roles. Temporal consistency discourages flicker, structural preservation protects visible tissue boundaries, and residual regularization prevents over-enhancement or large color shifts. None of these terms uses a clean real-world reference.

D.3 Checkpoint Flow and Model Selection

The paired-to-real protocol is implemented independently for the two real datasets. First, EndoTME is supervisedly trained on the paired STSVD training split. The selected paired checkpoint initializes both real-world branches. Second, LSVD and EndoReal are adapted separately on their respective adaptation splits. Third, the corresponding validation split is used for checkpoint selection, and the held-out test split is opened only for final evaluation.

This protocol yields one STSVD checkpoint, one LSVD-adapted checkpoint, and one EndoReal-adapted checkpoint. Parameters are not transferred between LSVD and EndoReal, and the two real datasets are not mixed during adaptation. Consequently, the reported result on each dataset corresponds to a model selected without access to its test clips. All frames and clips from the same patient remain in one EndoReal split.

D.4 Implementation Settings

EndoTME is implemented in PyTorch and optimized using Adam with an initial learning rate of 1×10^{-4}. A cosine schedule decays the learning rate during training. Random cropping and random flipping are applied to the paired training clips. The same geometric transform is applied to every frame in a clip to preserve temporal alignment. The optimizer, schedule, temporal window, Top-K setting, number of experts, and loss weights are fixed within each reported experiment. The default RTMR setting is $K = 3$, and PIDC uses three spatial experts and four cross-attention heads. Experiments are conducted on NVIDIA RTX 3090 GPUs.

E Dataset Details and Ethical Considerations

E.1 Dataset Statistics and Splits

STSVD. STSVD contains 120 paired smoky–clean videos, with 100 frames per video (Liang et al., 2026). It is used for supervised learning, full-reference evaluation, ablation studies, and hyperparameter selection. We follow the public dataset partition and keep all frames from one video in the same split.

LSVD. LSVD contains 486 real laparoscopic clips collected from 40 patients, with 20–50 frames per clip (Wu et al., 2024). Following its official partition, 416 clips are available for training or adaptation and 70 clips are held out for testing. Clean frame-aligned references are not used by the reference-free adaptation objective.

EndoReal. EndoReal contains 5,691 real surgical endoscopic clips collected from 320 patients. Most recordings use the two principal resolutions reported in the main paper. The dataset is partitioned into adaptation, validation, and held-out test subsets at the patient level. No patient appears in more than one subset. All clips from the same surgical procedure are assigned together, preventing adjacent or related clips from crossing split boundaries.

E.2 EndoReal Construction

EndoReal is assembled from authentic surgical endoscopic recordings provided by collaborating clinical practitioners. The clips contain naturally occurring degradations, including blood or fluid contamination, smoke-like veiling, specular reflection, motion blur, weak illumination, low contrast, camera motion, instrument occlusion, and combinations of these effects.

Clip extraction preserves temporal order and does not construct paired clean targets. Invalid segments, including unusable transitions and frames without a valid endoscopic field of view, are excluded before training and evaluation. Patient and procedure identifiers are used only for split construction and are removed from the model input. The large number of clips is therefore used to increase the diversity of real surgical conditions rather than to create frame-level overlap between training and testing.

E.3 Preprocessing

Frames are decoded in RGB order and normalized consistently across all methods. During training, a spatial crop is sampled once per clip and applied to all frames. Random flipping follows the same clip-consistent rule. These operations avoid artificial inter-frame displacement.

During evaluation, the complete field of view is preserved. When overlapping tiled inference is required by GPU memory, all methods are evaluated with the same spatial protocol and overlapping predictions are smoothly merged. No test-time augmentation or metric-specific post-processing is used. The temporal order of each clip is preserved for both restoration and temporal evaluation.

E.4 Ethics, Consent, and Privacy

The EndoReal recordings were provided by collaborating clinical practitioners under authorization for research use. Patients were informed of the research use of the surgical videos and provided consent in accordance with the collaborating institution's data-use procedure. The videos were de-identified before research analysis. Patient names, clinical identifiers, and other identifying metadata are not used by the model or included in the experimental outputs.

The data are used only for method development and evaluation. The study does not use EndoTME to make clinical decisions, and the enhanced videos are not treated as replacements for the original clinical recordings.

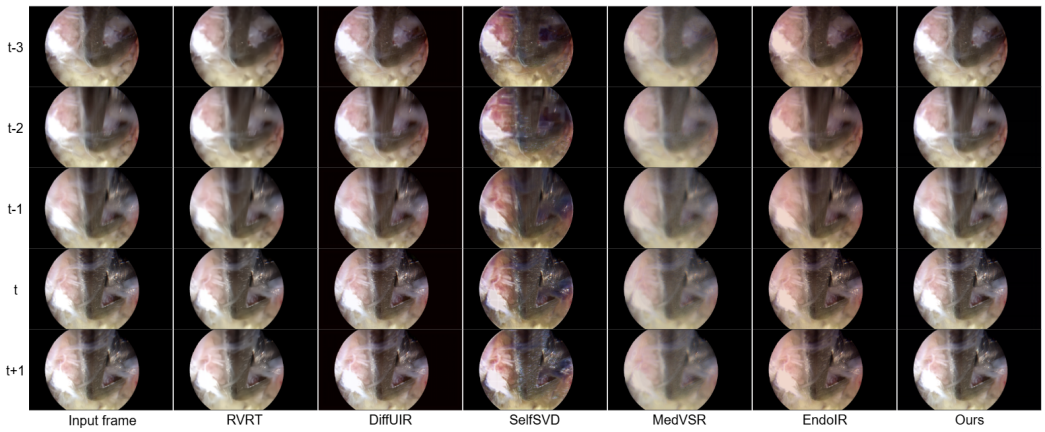


Figure 8: Additional consecutive-frame comparison on real-world endoscopic videos. Each column corresponds to one restoration method, and each row shows the same spatial region at a consecutive time step. All methods use identical source frames and crop coordinates. EndoTME preserves clearer tissue boundaries and more stable local appearance across time.

F Complete Experimental Protocol

F.1 Hardware and Software

All experiments are implemented in Python and PyTorch under Ubuntu. The main environment uses Python 3.10, PyTorch 2.1.2 with CUDA 11.8, and NVIDIA RTX 3090 GPUs with 24 GB memory. The optical-flow network is kept fixed during restoration training and adaptation. Mixed precision was used for training, while metric computation is performed in a fixed evaluation precision for every method.

F.2 Baseline Configuration

We compare with RVRT, DiffUIR, SelfSVD, MedVSR, and EndoIR using their official implementations or released model definitions. The comparison preserves the native input type of each method. Image restoration methods are applied frame by frame, and their outputs are reassembled in the original temporal order. Video restoration methods receive temporally ordered clips.

All generated sequences are evaluated by the same scripts, metric implementations, color range, and temporal order. Dataset splits and test clips are identical across methods. Hyperparameter selection and checkpoint selection do not use the held-out test subsets. This common evaluation pipeline prevents differences in metric software or data aggregation from influencing the comparison. STANet (Liang et al., 2026) is closely related to STSVD and is therefore discussed in the main paper. However, it is not included in the quantitative comparison because, at the time of submission, the publicly available STSVD repository provides the dataset and project resources but does not release the STANet model implementation or pretrained checkpoints required for a faithful reproduction.

Real-world adaptation protocol. To make the real-world comparison transparent, we report the initialization, target-domain data access, adaptation strategy, and checkpoint-selection protocol for each baseline and EndoTME. All methods are evaluated on exactly the same held-out test clips, and no test clip is used for hyperparameter or checkpoint selection.

F.3 Metric Implementation

PSNR is computed from mean squared error on paired RGB frames. SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), NIQE (Mittal et al., 2013), BRISQUE (Mittal et al., 2012), and CLIP-IQA (Wang et al., 2023b) are computed with a common image-quality

evaluation pipeline. LPIPS uses the AlexNet feature configuration. Full-reference metrics are reported only for STSVD, where aligned clean targets are available. NIQE, BRISQUE, and CLIP-IQA are used for LSVD and EndoReal because clean references are unavailable.

Temporal warping error measures disagreement between one enhanced frame and the preceding enhanced frame warped to the current time. Temporal optical-flow consistency measures the stability of estimated motion across the restored sequence. The same optical-flow estimator and valid-region handling are used for all methods.

Metrics are first averaged over frames within a clip. Clip-level values are then averaged within each dataset. For EndoReal, patient separation is retained during aggregation so that clips from one patient do not create training–test leakage. Higher values indicate better performance for PSNR, SSIM, and CLIP-IQA, while lower values are preferred for LPIPS, DISTS, NIQE, BRISQUE, tWE, and tOF.

F.4 Runtime Measurement

Runtime is measured with batch size one after GPU warm-up. CUDA synchronization is performed before and after each timed forward pass. Data loading, file decoding, disk writing, and metric computation are excluded. Optical-flow estimation and all modules required by EndoTME's forward pass are included. Per-frame runtime is obtained by dividing the elapsed clip time by the number of restored frames. Parameter counts include all trainable restoration parameters. The same measurement principle is applied to all methods in Fig. 5(b) of the main paper.

G Additional Protocol Clarifications

G.1 Ablation Study

The component ablation in Table 3 of the main paper evaluates RTMR and PIDC at the module level using a factorial design. The RTMR-only configuration disables PIDC, the PIDC-only configuration disables RTMR, and the baseline disables both modules.

For the internal RTMR ablation in Table 4, PIDC is enabled and kept fixed for all configurations, while the reliability components of RTMR are progressively incorporated from trajectory alignment to degradation-risk assessment. Conversely, for the internal PIDC ablation in Table 5, RTMR is enabled and kept fixed while only the degradation representation is modified. Thus, each internal ablation varies only the design of the target module.

Every ablation is trained and evaluated using the same STSVD split, optimization budget, checkpoint-selection rule, and evaluation pipeline. The Top-K experiment changes only K, with all other components fixed to the full EndoTME configuration.

H Additional Qualitative Results

H.1 RTMR Visualization

Additional RTMR cases focus on temporal situations in which motion alignment alone is insufficient. The selected examples include historical frames affected by blood haze, strong reflection, blur, or large displacement. For each query, the visualization shows the candidate regions and the final retrieval decision. Cases in which a visually similar but degraded candidate is rejected are particularly useful because they directly illustrate the distinction between motion correspondence and memory reliability.

H.2 PIDC Cue and Routing Visualization

Additional PIDC cases connect the six analytical maps to the learned routing behavior. The visualization keeps the input, cue maps, expert weights, and enhanced output spatially aligned. Rather than claiming a one-to-one semantic label for every expert, the analysis demonstrates that spatial

routing changes according to the local combination of degradation cues and trajectory-conditioned features.

H.3 Additional Temporal Comparisons

Figure 8 presents additional comparisons over consecutive frames. All methods receive the same input sequence, and the same spatial region is displayed across time. This presentation enables direct inspection of frame-wise appearance variation, tissue-boundary stability, residual degradation, and temporally propagated artifacts.

Frame-based restoration methods can improve individual frames but may produce inconsistent color, contrast, or detail across adjacent time steps. Generic video restoration methods improve temporal interaction, but unreliable historical observations may still introduce persistent or drifting artifacts. By combining selective trajectory-memory retrieval with degradation-adaptive local restoration, EndoTME maintains clearer tissue structures and a more stable local appearance throughout the sequence.

I Failure Cases and Limitations

EndoTME relies on optical flow to establish temporal correspondence. Extremely rapid camera movement, severe non-rigid deformation, or large instrument occlusion can reduce the accuracy of the historical trajectory. RTMR mitigates this problem by lowering the reliability of unstable candidates, but it cannot recover evidence that is absent from all observed frames.

The physics-inspired maps are analytical indicators rather than a complete physical model of endoscopic image formation. Bright instruments can activate the specular cue, and large valid motion can produce a strong flow-magnitude response. PIDC reduces sensitivity to any single cue by combining type-level cross-attention, dense spatial embedding, and learned expert routing.

Finally, no-reference image-quality and temporal metrics provide a practical evaluation for real surgical videos but do not fully measure clinical utility or anatomical fidelity. EndoTME is designed as a visual enhancement method and does not replace the original video or clinical judgment. Multi-center validation and clinician-centered evaluation remain important future directions.