SURGOMNI: TOWARDS ALL-IN-ONE SURGICAL VIDEO RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Laparoscopic videos are often affected by diverse degradations arising at different stages of the imaging process, making all-in-one surgical restoration necessary for real-world deployment. However, advancing all-in-one surgical restoration remains constrained by three fundamental limitations: (1) existing training datasets either lack exact pixel-level degraded-clean correspondence or rely on simplified synthesis that inadequately models the formation of complex surgical degradations; (2) existing real-world testing datasets often remain image-based and cover only a limited range of degradation types and surgical scenarios, restricting comprehensive evaluation of model performance; and (3) the relationship between all-in-one training degradation composition and real-world testing generalization remains poorly understood. To address these limitations, we introduce SurgOmni, a unified dataset and benchmark suite for all-in-one surgical video restoration. Specifically, we first propose OmniSyn, a 44,000-frame paired training dataset comprising 14 degradation configurations and constructed through a formation-aware synthesis pipeline, where scene- and lens-space degradations are rendered via a physics-based 3D engine with temporal dynamics, while exposure- and illumination-related degradations are modeled through exposure integration and content-aware spatial photometric modeling, respectively. Meanwhile, we further propose OmniBench, a 10,000-frame real-world video testing dataset spanning 6 surgical types, 10 target organs, 14 degradation configurations, and 27 surgical scenes. Extensive training-dataset comparisons first show that models trained on OmniSyn consistently achieve the strongest generalization to real surgical data. Comprehensive benchmarking across diverse restoration methods on OmniBench then reveals distinct strengths and limitations in restoration quality and temporal motion deviation that are obscured by limited real-world evaluation. Beyond performance evaluation, we further investigate how training-data degradation composition affects real-world testing generalization, revealing the relative effectiveness of different composition strategies. These findings provide practical guidance for training-data design and restoration method development for all-in-one surgical restoration. The paper website is available at https://anonymous.4open.science/w/SurgOmni_31348/.

1 INTRODUCTION

Clear visualization is essential for reliable intraoperative perception in laparoscopic surgery, yet surgical videos are frequently degraded by electrosurgical smoke, lens condensation, fluid splashes, motion blur, and non-uniform illumination arising at different stages of the imaging process (Kubo et al., 2023; Pei et al., 2026). These degradations differ in their formation mechanisms and may occur individually or jointly, motivating all-in-one surgical restoration for real-world deployment.

All-in-one restoration has advanced rapidly in natural imagery, where recent methods increasingly address diverse degradations within unified restoration frameworks (Li et al., 2022; Potlapalli et al., 2023; Cui et al., 2025b; Tian et al., 2025; Lin et al., 2026). In contrast, surgical restoration remains largely degradation-specific, with most methods targeting individual degradations such as surgical smoke (Li et al., 2024; Wu et al., 2024b; Liang et al., 2026). Although recent studies have begun to extend surgical restoration toward multiple degradation settings (Chen et al., 2026b; Pei et al., 2026), all-in-one surgical video restoration remains largely underexplored, which remains constrained by three fundamental limitations.

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.

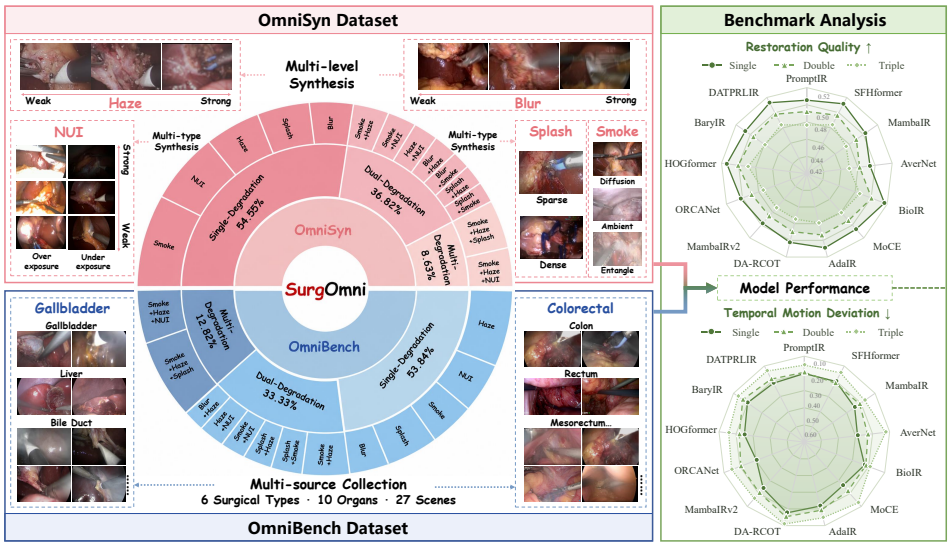


Figure 1: Overview of SurgOmni, comprising formation-aware paired training and real-world testing data. Benchmark analysis reports restoration quality as the normalized mean of TOPIQ, MUSIQ, and Q-Align, together with temporal motion deviation across single, double, and triple degradation.

From the training-data perspective, existing datasets struggle to jointly provide accurate pixel-level degraded-clean correspondence and realistic degradation modeling. Real surgical data preserve authentic degradation appearance but cannot provide exact pixel-level correspondence, leading existing datasets to rely on unpaired samples (Pan et al., 2022; Wu et al., 2024b) or temporally adjacent pseudo-targets with correspondence errors (Xia et al., 2024; Pei et al., 2026). Synthetic datasets provide exact alignment, yet their synthesis pipelines often simplify the formation of complex surgical degradations (Chen et al., 2020; Khan et al., 2020; Zhang et al., 2023). **From the real-world testing-data perspective**, existing datasets often remain image-based and cover only a limited range of degradation types and surgical scenarios, restricting comprehensive evaluation across diverse surgical conditions. **From the training-to-testing generalization perspective**, how training-data degradation composition affects real-world generalization remains poorly understood, limiting the formulation of training composition strategies for all-in-one surgical restoration.

To address these limitations, we introduce **SurgOmni**, a unified dataset and benchmark suite for all-in-one surgical video restoration. As shown in Figure 1, we first propose **OmniSyn**, a 44,000-frame paired training dataset constructed through a formation-aware synthesis pipeline. Scene- and lens-space degradations are rendered with a physics-based 3D engine and temporal dynamics, while exposure- and illumination-related degradations are modeled through exposure integration and content-aware spatial photometric modeling. To explore the training effect of multi-degradation samples, we further construct a dedicated subset by synthesizing multiple degradations on the same clean sequence while retaining the spatial and temporal characteristics of their respective formation processes, yielding pixel-level correspondence across 5 single-degradation types and 9 multi-degradation combinations. Meanwhile, we further propose **OmniBench**, a 10,000-frame real-world video testing dataset spanning 6 surgical types, 10 target organs, 14 degradations, and 27 surgical scenes. Extensive training-dataset comparisons first show that models trained on OmniSyn consistently achieve the strongest generalization to real surgical data. Comprehensive benchmarking across diverse restoration methods on OmniBench then reveals distinct strengths and limitations in restoration quality and temporal motion deviation that are obscured by limited image-based real-world evaluation. Beyond performance evaluation, we finally investigate how training-data degradation composition affects real-world testing generalization, revealing the relative effectiveness of single- and multi-degradation samples under different training compositions.

Our contributions are threefold.

- We propose **OmniSyn**, a 44,000-frame paired video training dataset covering 14 degradation configurations with a formation-aware synthesis pipeline. Extensive training-dataset comparisons suggest that all-in-one surgical restoration training dataset design should

jointly provide accurate degraded-clean correspondence, faithful degradation modeling, and diverse coverage of surgical scenes and degradation severities.

- We propose **OmniBench**, a 10,000-frame real-world video testing dataset spanning 6 surgical types, 10 target organs, 14 degradation configurations, and 27 surgical scenes. Comprehensive benchmarking suggests that all-in-one surgical restoration methods may benefit from preserving fine-grained structures and limiting restoration-induced motion deviations rather than focusing on stronger single-frame restoration or additional temporal inputs.
- We investigate how training degradation composition affects real-world testing generalization. Systematic analyses reveal the limitations of current multi-degradation training and suggest that all-in-one surgical restoration training composition design should prioritize sufficient single-degradation supervision to establish reliable restoration anchors under a limited training budget.

2 Related Work

Surgical Restoration Dataset. Paired surgical restoration data are difficult to acquire because tissue state, viewpoint, instruments, and illumination continuously change during surgery. Early real-world datasets therefore relied on unpaired degraded-clean samples (Venkatesh et al., 2020; Ali et al., 2021; Pan et al., 2022) or pseudo-pairs constructed from temporally adjacent frames with motion compensation (Wu et al., 2024b; Xia et al., 2024), where imperfect alignment introduces residual correspondence errors. More recently, SurgClean (Pei et al., 2026) extends real-world degradation coverage to smoke, haze, and splash, while still treating them as separate image-based restoration tasks and relying on optical-flow-based pseudo-pairs. On the other hand, synthetic datasets resolve the correspondence problem by providing paired synthesized supervision, yet most existing surgical datasets remain limited to individual degradations such as smoke or low light (Chen et al., 2020; Zhang et al., 2023; Liu et al., 2024; Hong et al., 2023; Chang et al., 2024; Chen et al., 2024b; Liang et al., 2026). The ICIP LVQ challenge (Khan et al., 2020) extends synthetic data to multiple degradations, but relies on simplified frame-level degradation modeling developed for quality assessment and recognition rather than surgical restoration. Therefore, existing datasets remain limited in jointly providing exact correspondence, realistic degradation modeling, and broad multiple degradation coverage.

All-in-One Restoration. All-in-one restoration has been extensively studied in natural imagery. Early methods mainly condition a shared restoration model on degradation representations or learned prompts (Li et al., 2022; Potlapalli et al., 2023), while subsequent approaches introduce degradation-adaptive feature modulation and expert routing (Cui et al., 2025b; Tian et al., 2025; Lin et al., 2026). More recent methods further incorporate vision-language and generative priors (Luo et al., 2024; Chen et al., 2025; Rajagopalan et al., 2025) and extend the problem from image restoration to composite (Guo et al., 2024b) and time-varying video degradations (Zhao et al., 2024; Li et al., 2026). In contrast, surgical restoration remains largely degradation-specific, where most methods targeting surgical smoke across GAN, CNN, Transformer, and diffusion-based architectures (Chen et al., 2020; Venkatesh et al., 2020; Pan et al., 2022; Zhang et al., 2023; Li et al., 2024; Wu et al., 2024b; Liang et al., 2026). EndoIR (Chen et al., 2026b) represents a recent step toward all-in-one surgical image restoration by handling multiple degradations within a unified diffusion framework. However, all-in-one surgical video restoration still remains largely underexplored, where progress is constrained by the lack of reliable paired training data and real-world testing data for comprehensive evaluation.

3 SurgOmni

3.1 Degradation Taxonomy

Based on prior surgical imaging and restoration studies (Khan et al., 2020; Kubo et al., 2023; Pei et al., 2026), we define five representative surgical degradation types spanning distinct formation stages of the imaging process: smoke, haze, splash, motion blur, and non-uniform illumination (NUI). Specifically, smoke is modeled as scene-space aerosol, haze as lens-adjacent condensation, splash as lens-adherent fluid, motion blur through exposure integration, and NUI as spatially varying over- and under-illumination. The same taxonomy is used for formation-aware synthesis in OmniSyn and annotation of real degradations in OmniBench. Table 1 compares SurgOmni with existing representative surgical restoration datasets.

Table 1: Comparison of surgical restoration datasets, where the upper and lower blocks summarize synthetic training datasets and real-world testing datasets, respectively.

Dataset	Mode	Resolution	Smoke	Haze	Splash	Blur	NUI	Multi.	Scale	Synthesis Paradigm / Surgical Anatomy
PSv2rs	Image	256×256	✓	×	×	×	×	×	54,420	Frame-wise Smoke Synthesis
MARS-GAN	Image	256×256	✓	×	×	×	×	×	18,000	Frame-wise Smoke Synthesis
PFAN	Image	480×480	✓	×	×	×	×	×	660	Frame-wise Smoke Synthesis
STSVD	Video	1080×720	✓	×	×	×	×	×	12,000	Temporal Smoke Synthesis
ICIP LVQ	Video	512×288	✓	×	×	✓	✓	✓	20,000	Frame-wise Multiple Degradation Synthesis
OmniSyn	**Video**	**1080×720**	✓	✓	✓	✓	✓	✓	**44,000**	**Temporal Multiple Degradation Synthesis**
Vivo	Image	700×350	✓	×	×	×	×	×	961	Prostate
SurgClean	Image	1280×720	✓	✓	✓	×	×	×	3,113	Gall, Bile, Liver, Pan, Spl; Lung, Med, Eso
LSVD	Video	1080×1920	✓	×	×	×	×	×	2,106	N/R
STSVD-R	Video	1080×720	✓	×	×	×	×	×	2,400	Gall
OmniBench	**Video**	**1080×720**	✓	✓	✓	✓	✓	✓	**10,000**	**Gall, Bile, Liver; Cec, AC, TC, DC, Sig, Rect, Meso**

Multi. indicates whether synthetic training datasets or real-world testing datasets include multi-degradation samples. Size is reported in frames for video datasets and images for image datasets. N/R: not reported. Gall: Gallbladder, Bile: Bile Duct, Pan: Pancreas, Spl: Spleen, Med: Mediastinum, Eso: Esophagus, Cec: Cecum, AC: Ascending Colon, TC: Transverse Colon, DC: Descending Colon, Sig: Sigmoid Colon, Rect: Rectum, Meso: Mesorectum.

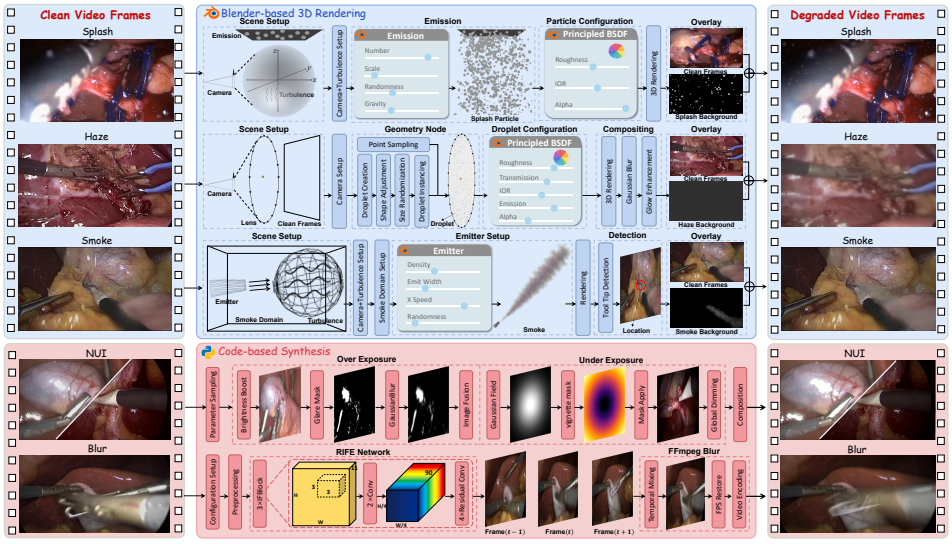


Figure 2: Overview of the OmniSyn degradation synthesis pipeline. Five degradations are modeled according to their formation mechanisms across scene, lens, exposure, and illumination processes, with temporal dynamics explicitly preserved to produce physically consistent degradation evolution.

3.2 OmniSyn: Formation-Aware Synthetic Video Training Dataset

Clean-video curation. We collect temporally continuous clean video clips from CholecTrack20 (Nwoye et al., 2025) and M2CAI16-TOOL (Twinanda et al., 2016), while preserving natural camera, tissue, and instrument motion. **Degradation synthesis.** As illustrated in Figure 2, OmniSyn adopts two formation-aware synthesis branches, where scene- and lens-space degradations are generated through physics-based 3D rendering with temporal trajectory dynamics, while exposure- and illumination-related degradations are modeled through code-based synthesis using exposure integration and content-aware spatial photometric modeling. Specifically, smoke is represented as scene-space aerosol with diffusion, ambient, and entangled patterns (Liang et al., 2026), while haze and splash are represented as lens condensation and lens-adherent fluid using droplet geometry, scattering, refraction, and absorption. Motion blur is synthesized through temporal densification with RIFE (Huang et al., 2022) followed by exposure integration, whereas NUI is synthesized through content-aware spatial photometric modeling with highlight expansion and spatially varying low-light attenuation. To explore the training effect of multi-degradation samples, we further synthesized a multi-degradation training subset on the same clean video clip in 7 dual and 2 triple combinations. Rather than exhaustively enumerating higher-order combinations, we focus on clinically plausible co-occurrences consistent with real surgical data (Ali et al., 2019; Chen et al., 2026b). Across all degradation types, we sample multiple severity levels from mild to severe, with the severity ranges informed by prior surgical studies (Wang et al., 2019; Khan et al., 2020; Wang et al., 2024). **Quality control.** All synthesized videos undergo automatic integrity checks and multi-stage visual inspec-

tion. Finally, OmniSyn contains 44,000 paired high-resolution frames in 440 videos, organized into a single-degradation training subset covering 5 degradation types and a multi-degradation training subset covering 9 degradation combinations. The two subsets contain 240 single-degradation videos and 200 multi-degradation videos, including 162 dual- and 38 triple-degradation videos. Detailed dataset construction processes and distributions are provided in Appendix A.

3.3 OMNIBENCH: REAL-WORLD VIDEO TESTING DATASET

OmniBench is designed for real-world evaluation under authentic surgical degradations. Its construction follows four stages. (1) **Video collection.** We combine public gallbladder recordings (Twinanda et al., 2016; Nwoye et al., 2025) with a newly curated clinical colorectal cohort from a hospital to form the Gallbladder and Colon testing subsets, respectively, broadening the coverage of existing testing datasets across procedures, anatomy, and degradations. For public recordings, OmniSyn and OmniBench are separated at the source-video level, ensuring that no original recording contributes to both datasets and preventing direct content leakage or near-duplicate evaluation. (2) **Clip selection.** Candidate recordings are screened to retain temporally continuous clips with clearly visible degradations and sufficient anatomical or instrumental information for meaningful evaluation. (3) **Degradation annotation.** We apply multi-label annotation at the clip level using the same five single-degradation and nine multi-degradation taxonomies as OmniSyn, allowing naturally co-occurring degradations to be retained in each clip. (4) **Quality control.** Each selected clip is independently verified for degradation visibility, annotation reliability, and restoration feasibility, with disagreements resolved through full-clip review and clinician adjudication when necessary. After verification, OmniBench contains 10,000 high-resolution frames in 100 videos spanning 6 surgical types, 10 target organs, 14 degradation configurations, and 27 surgical scenes, including 43 single-, 45 dual-, and 12 triple-degradation videos. OmniBench is reserved exclusively for evaluation and is not used for training or model selection. Detailed dataset collection processes and distributions are provided in Appendix B.

4 EXPERIMENTS AND ANALYSIS

4.1 EXPERIMENTAL SETUP

Datasets and Implementation Details. For the training-dataset comparison, SurgClean (Pei et al., 2026) and LVQ (Khan et al., 2020) are used as alternative training sources and compared with matched-budget subsets of OmniSyn. For the benchmark, all models are trained on OmniSyn, which is partitioned into training, validation, and test sets through configuration-stratified allocation targeting an 8:1:1 ratio. Real-world performance is evaluated on OmniBench, comprising the degradation-complexity-stratified Gallbladder and Colon subsets at a 6:4 ratio. Following the standard all-in-one training protocol (Potlapalli et al., 2023; Cui et al., 2025b; Tian et al., 2025; Wu et al., 2026), both the training-dataset comparison and the benchmark use single-degradation samples for training. All 15 baselines are implemented in PyTorch and trained on NVIDIA L40S GPUs following their respective official optimization protocols, covering both image and video all-in-one restoration methods spanning CNN, Transformer, state-space, Fourier-based, diffusion, and flow matching architectures.

Evaluation Metrics. Following common evaluation for real-world restoration datasets without clean references, OmniBench adopts three complementary learned no-reference metrics from multi-scale, local, and global perceptual perspectives. Specifically, MUSIQ (Ke et al., 2021) evaluates image quality through multi-scale representations across different spatial resolutions, TOPIQ (Chen et al., 2024a) uses high-level semantic guidance to assess content-dependent local distortions, and Q-Align (Wu et al., 2024a) provides human-aligned global perceptual assessment through a multi-modal large language model. Meanwhile, temporal motion deviation is evaluated using tOF (Chu et al., 2020), defined as the mean endpoint discrepancy between Farnebäck optical flows of consecutive restored and reference frames (Farnebäck, 2003). On OmniBench, the degraded sequence serves as the motion reference in the absence of a clean counterpart, so tOF measures restoration-induced deviation from the observed inter-frame motion. To further assess downstream task-relevant structural preservation beyond restoration performance, we further measure prediction consistency for scene segmentation (Kamtam et al., 2025), instrument segmentation (Ma et al., 2024), and monocular depth (Tian et al., 2024) under frozen surgical perception models, using Dice (Dice, 1945), IoU (Jaccard, 1901), δ_1, AbsRel, SqRel, and RMSE (Eigen et al., 2014). Frame-level scores on Om-

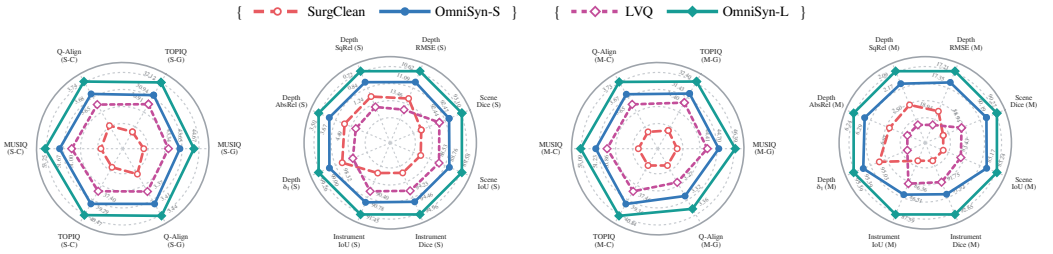


Figure 3: Method-averaged training dataset comparison under matched-budget settings across all restoration methods. From left to right, the radar charts report restoration and downstream results under single degradations (S), followed by multiple degradations (M). Restoration results are further stratified by the Gallbladder (G) and Colon (C) subsets of OmniBench, while downstream prediction consistency is evaluated on OmniSyn, where paired clean frames provide the required references.

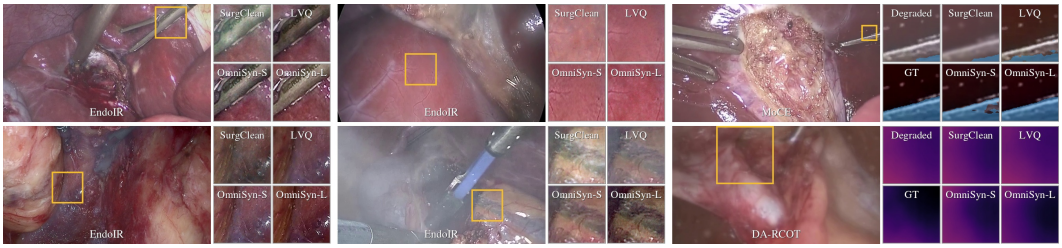


Figure 4: Qualitative comparison across the four training datasets. The left two columns show restoration results, while the rightmost column shows segmentation and monocular depth estimation in the top and bottom rows, respectively.

niBench are first averaged within each clip and then across clips to avoid over-weighting temporally correlated frames. All restoration and downstream results are averaged over three independent runs. Further details on restoration and downstream evaluation protocols are provided in Appendix C.

4.2 Training Dataset Comparison

Comparison Analysis. To validate the effectiveness of OmniSyn as a training dataset, we compare it with the SurgClean image dataset and LVQ video dataset under two matched-budget settings with aligned scale and degradation types. The SurgClean comparison includes 13 image restoration methods, while the LVQ comparison includes the same image restoration methods and 2 video restoration methods. **(1) Matched-scale restoration comparisons show that training datasets combining accurate spatial pairing and formation-aware degradation modeling achieve stronger real-world restoration generalization.** As shown in Figure 3, OmniSyn achieves higher mean performance across all 28 aggregate dimensions in both the 2,488-pair {SurgClean-OmniSyn-S} and the 10,000-pair {LVQ-OmniSyn-L} comparisons. SurgClean relies on optical-flow-aligned pseudo-pairs, where imperfect flow estimation can introduce residual spatial misalignment into the reconstruction targets. LVQ provides exact pairs, but its simplified frame-wise degradation synthesis does not explicitly model the complex formation processes of real surgical degradations. In contrast, OmniSyn preserves exact degraded-clean spatial correspondence while modeling different degradation types according to their respective scene, lens, exposure, or illumination formation stages, achieving stronger transfer to real surgical scenarios. Meanwhile, **(2) Cross-scale restoration comparison suggests that broader coverage of surgical content and degradation severities may partially offset the disadvantage of a reduced training-set size.** Despite using far fewer training pairs, OmniSyn-S remains stronger than LVQ across most aggregate dimensions when the comparison is restricted to the same 13 image restoration methods, with the qualitative comparison in Figure 4 showing the same overall trend. LVQ expands its training data mainly by repeatedly synthesizing a limited set of surgical scenes at a few discrete severity levels, whereas OmniSyn-S covers broader surgical content and samples each degradation over a wider severity range from mild to severe. This broader variation is consistent with the stronger real-world generalization observed for OmniSyn-S under the common-method comparison despite its substantially smaller training set. In addition, **(3) Downstream task comparisons show that the comparative training datasets exhibit systematic task-specific biases, while OmniSyn preserves structural information more consistently across different tasks.** Across both matched comparisons, OmniSyn achieves stronger overall consistency in segmentation and depth estimation, showing that the advantages observed in restoration quality also extend to task-relevant structural preservation. Notably, SurgClean shows a relative bias toward depth estimation, which may reflect the authentic degradation-scene interactions and pho-

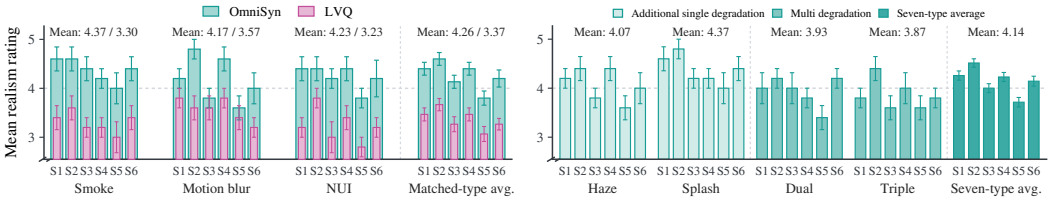


Figure 5: Blinded surgeon assessment of synthetic degradation realism (higher is better). Left: matched comparison of OmniSyn and LVQ on the three shared degradation types. Right: OmniSyn ratings on additional single- and multi-degradation settings. S1-S6 denote the six attending surgeons, with error bars and annotations showing within-surgeon standard and across-surgeon mean.

Table 2: Quantitative comparison of 15 restoration methods on OmniBench across single, double, and triple degradation testing subsets. Green, blue, and red cells denote low, medium, and high-performance tiers from one-dimensional k-means applied separately to each metric and degradation complexity. ‡ denotes models whose parameter count is dominated by a frozen pretrained foundation model. Latency is measured end-to-end on a single NVIDIA L40S GPU at 1080×720, while MACs are computed at 256×256.

Model	Venue	Single Degradation				Double Degradation				Triple Degradation				Efficiency		
		TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓	TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓	TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓	Params↓	MACs↓	Latency↓
PromptIR	NeurIPS'23	0.364	49.870	3.614	0.201	0.357	48.880	3.523	0.194	0.338	48.106	3.441	0.152	35.6 M	158 G	0.85 s
SFHformer	ECCV'24	0.367	50.354	3.658	0.202	0.358	49.024	3.566	0.212	0.342	48.878	3.487	0.145	3.0 M	18 G	0.17 s
MambaIR	ECCV'24	0.362	49.509	3.605	0.244	0.354	48.832	3.510	0.209	0.335	48.298	3.436	0.173	15.8 M	1166 G	7.65 s
AverNet	NeurIPS'24	0.361	48.012	3.564	0.288	0.351	47.520	3.478	0.228	0.329	47.352	3.428	0.124	41.3 M	75 G	0.06 s
BioIR	NeurIPS'25	0.370	51.170	3.672	0.225	0.360	49.579	3.583	0.217	0.342	49.182	3.495	0.192	15.8 M	125 G	0.94 s
MoCE	CVPR'25	0.369	49.888	3.606	0.262	0.359	48.765	3.529	0.219	0.341	48.569	3.461	0.128	11.5 M	41 G	0.56 s
AdaIR	ICLR'25	0.364	50.209	3.642	0.217	0.356	48.808	3.546	0.192	0.338	48.221	3.468	0.147	28.8 M	147 G	1.09 s
DA-RCOT	TPAMI'25	0.363	48.899	3.621	0.176	0.355	48.080	3.532	0.153	0.334	47.751	3.449	0.109	52.3 M	333 G	1.88 s
MambaIRv2	CVPR'25	0.356	48.505	3.604	0.240	0.349	47.981	3.515	0.206	0.333	47.661	3.438	0.161	9.4 M	613 G	7.94 s
ORCANet	ArXiv'26	0.365	48.831	3.612	0.302	0.358	47.880	3.516	0.205	0.335	47.242	3.461	0.148	70.0 M	49 G	0.15 s
EndoIR	AAAI'26	0.344	46.830	3.551	0.764	0.337	46.630	3.461	0.642	0.316	45.480	3.298	0.633	24.8 M	250 G	0.10 s
RAR	CVPR'26	0.222	34.498	2.882	1.632	0.221	34.072	2.845	1.221	0.222	34.244	2.913	1.030	9411.3 M‡	40708 G	8.20 s
HOGformer	AAAI'26	0.374	50.567	3.646	0.249	0.363	49.427	3.576	0.219	0.346	49.362	3.506	0.157	16.6 M	92 G	1.79 s
BaryIR	TPAMI'26	0.369	49.610	3.623	0.198	0.361	48.844	3.548	0.165	0.341	48.047	3.463	0.132	53.8 M	197 G	1.06 s
DATPRLIR	ICLR'26	0.372	50.665	3.643	0.213	0.363	49.512	3.544	0.186	0.343	49.094	3.473	0.133	42.6 M	170 G	1.19 s

tometric variations retained in its real degraded inputs, both of which are important for continuous geometric estimation. LVQ instead shows a relative advantage in segmentation, consistent with its exact degraded-clean correspondence preserving semantic boundaries and region assignments. By contrast, OmniSyn combines these properties by using physics-based 3D rendering to model degradation-scene interactions while maintaining exact degraded-clean correspondence, which is consistent with its more balanced structural preservation across downstream tasks.

Surgeon Assessment. As shown in Figure 5, six attending surgeons (S1-S6) with over 10 years of laparoscopic experience independently evaluated synthetic degradation realism in a blinded five-point rating study. Across the three degradation types shared with LVQ, OmniSyn received higher ratings in all 18 surgeon-type comparisons and achieved a matched-type average of 4.26, compared with 3.37 for LVQ. OmniSyn also maintained high ratings across the additional single- and multi-degradation settings, showing that its synthetic realism remains strong beyond the shared degradation types. Full protocol and results are provided in Appendices C.6 and D.3.

4.3 Benchmark Analysis

Table 2 compares restoration quality and temporal motion deviation across single, double, and triple degradation testing subsets. **(1) The low-tier (green) restoration results show that generative restoration methods lacking explicit measurement-consistency constraints can suffer from reduced restoration fidelity and increased inter-frame motion deviation, while highly compressed latent representations may further aggravate these limitations.** EndoIR (Chen et al., 2026b) conditions pixel-space diffusion on degradation information, while RAR (Chen et al., 2026a) maps degraded latents to clean latents through flow matching with iterative quality assessment. Compared with discriminative restoration methods, these generative mappings have greater freedom in estimating clean content and lack explicit measurement-consistency mechanisms, which may allow tissue boundaries and local textures to deviate from the observed input under real-world distribution shifts. In addition, RAR further performs restoration in the 8× spatially downsampled Stable Diffusion 3.5 latent space, which strongly compresses high-resolution details and may further limit the restoration and temporal performance. **(2) The middle and high-tier (blue and red)**

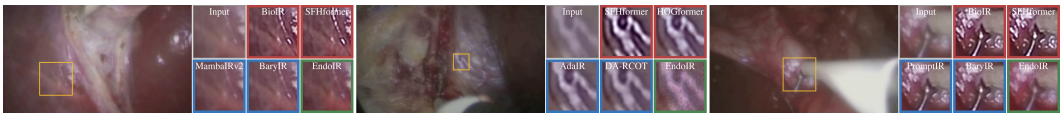


Figure 6: Qualitative comparison of representative restoration methods on OmniBench. Green, blue, and red boxes indicate low, medium, and high restoration-performance tiers, respectively.

cross-method quality results show that methods with explicit mechanisms for preserving fine-grained structures achieve stronger real-world restoration generalization. HOGformer (Wu et al., 2026), SFHformer (Jiang et al., 2024), BioIR (Cui et al., 2025a), and DATPRLIR (Dong et al., 2026) consistently occupy the stronger perceptual tiers and share an explicit focus on fine-grained structural modeling. By explicitly propagating or constraining fine-grained structures such as tissue boundaries and instrument contours during training, these methods achieve stronger restoration generalization in real surgical scenarios, as also illustrated qualitatively in Figure 6. **(3) The middle- and high-tier (blue and red) cross-method temporal results show that methods with more constrained restoration mappings are associated with lower restoration-induced inter-frame motion deviation.** DA-RCOT (Tang et al., 2025) constrains the restoration mapping through residual-conditioned optimal transport, while BaryIR (Tang et al., 2026) maps degraded representations toward a Wasserstein barycenter space. These constrained mappings guide restoration along restricted distributional directions and may reduce additional motion deviations caused by changes to degradation-irrelevant content. Meanwhile, AverNet (Zhao et al., 2024) and ORCANet (Li et al., 2026) do not consistently reach the lowest tOF tier despite using temporal inputs, suggesting that temporal feature aggregation alone does not guarantee lower restoration-induced motion deviations. **(4) The degradation-complexity results show that increasing degradation complexity reduces overall restoration quality while largely preserving the cross-method patterns in restoration quality and temporal behavior.** As the number of co-occurring degradations increases from single to double and triple, most methods exhibit a gradual decline in perceptual restoration quality, the relative performance tier remains largely consistent, indicating that the preceding cross-method findings are robust across different degradation complexities. More benchmark results are further provided in Appendix D.

Table 3: Training-composition comparison across 15 restoration methods on OmniBench. **Bold** and underline represent the best and second-best configuration within each method and metric.

Model	Venue	Single				Multi				Mixed			
		TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓	TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓	TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓
PromptIR	NeurIPS'23	**0.358**	**49.213**	**3.552**	**0.192**	0.340	46.749	3.473	0.637	0.357	48.587	3.549	0.339
SFHformer	ECCV'24	**0.360**	**49.578**	**3.596**	**0.200**	0.327	45.470	3.364	0.634	0.350	48.256	3.528	0.330
MambaIR	ECCV'24	**0.355**	**49.059**	**3.542**	**0.219**	0.338	45.663	3.405	0.759	0.348	48.375	3.540	0.317
AverNet	NeurIPS'24	**0.353**	**47.711**	**3.509**	**0.241**	0.340	45.858	3.345	0.489	0.349	46.281	3.469	0.561
BioIR	NeurIPS'25	**0.362**	**50.215**	**3.611**	**0.217**	0.334	47.183	3.438	0.632	0.353	49.408	3.604	0.299
MoCE	CVPR'25	**0.361**	**49.224**	**3.554**	**0.227**	0.320	44.032	3.264	0.815	0.355	48.517	3.521	0.290
AdaIR	ICLR'25	**0.357**	**49.340**	**3.578**	**0.197**	0.333	44.825	3.331	0.802	0.345	48.261	3.542	0.391
DA-RCOT	TPAMI'25	**0.356**	**48.393**	**3.560**	**0.158**	0.353	47.126	3.513	0.493	0.353	47.718	3.529	0.246
MambaIRv2	CVPR'25	**0.350**	**48.168**	**3.544**	**0.215**	0.337	45.348	3.445	0.588	0.345	47.686	3.519	0.294
ORCANet	ArXiv'26	**0.358**	**48.212**	**3.551**	**0.240**	0.346	44.600	3.413	0.796	0.352	47.435	3.541	0.316
EndoIR	AAAI'26	**0.337**	**46.578**	**3.480**	**0.693**	0.314	45.339	3.296	0.940	0.329	46.157	3.401	0.777
RAR	CVPR'26	0.222	34.276	2.869	1.375	**0.254**	**38.622**	**3.006**	**1.078**	0.228	34.342	2.890	1.304
HOGformer	AAAI'26	**0.366**	**49.909**	**3.598**	**0.225**	0.337	44.793	3.249	0.829	0.359	49.146	3.572	0.281
BaryIR	TPAMI'26	**0.362**	**49.078**	**3.570**	**0.175**	0.348	46.717	3.460	0.621	0.357	48.773	3.558	0.254
DATPRLIR	ICLR'26	**0.364**	**49.958**	**3.578**	**0.191**	0.324	45.011	3.248	0.740	0.364	49.740	3.556	0.339

4.4 Training Composition Analysis

Table 3 compares three equal-scale training compositions for all-in-one surgical restoration across 15 restoration methods. Specifically, "Single" uses 160 single-degradation clips, "Multi" uses 160 multi-degradation clips, and "Mixed" contains 80 clips from each subset, with stratified sampling preserving the full degradation and severity ranges of both source subsets. **The equal-scale training composition results reveal a pronounced synthetic-to-real generalization gap for multi-degradation training.** Across the three compositions, "Single" achieves the strongest overall restoration performance, "Mixed" remains intermediate, and "Multi" performs worst on the real test sets. RAR is the only consistent exception, possibly because multi-degradation samples provide more complete supervision for its Restore-Assess-Repeat mechanism to identify and remove residual degradations over successive restoration steps. The subset-wise results in Figure 7 further

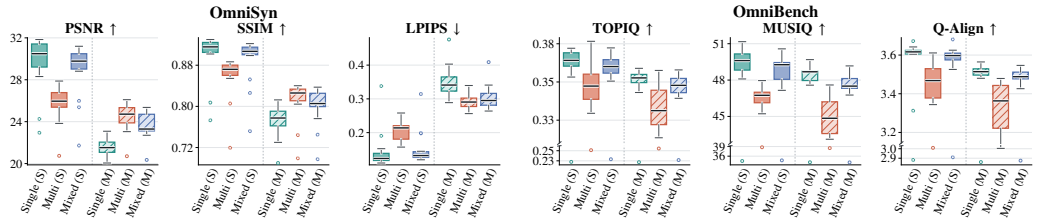


Figure 7: Effect of training degradation composition on single-degradation (S) and multi-degradation (M) test subsets. The left and right panels report method-averaged scores across 15 restoration methods on the OmniSyn test set and OmniBench, respectively, under "Single", "Multi", and "Mixed" three training compositions.

show that "Multi" performs best on the synthetic multi-degradation subset on OmniSyn but worst on the real multi-degradation subset on OmniBench. This reversal provides evidence of a domain gap, as real surgical degradations involve complex interactions among formation processes, degradation severities, and temporal co-occurrence patterns that are not fully represented by current synthesis pipelines. By contrast, **single-degradation training achieves stronger real-world generalization under the same training budget and appears to provide more transferable restoration anchors.** Because individual degradation mechanisms and severity variations can be modeled separately, "Single" provides more direct degradation-specific supervision, while its strong performance on real multi-degradation cases indicates that these anchors remain useful under co-occurring degradations. Under a limited training budget, our results therefore suggest prioritizing sufficient single-degradation training to establish reliable and transferable restoration anchors. More composition results and analysis are provided in Appendix D.4.

4.5 Discussion: Path towards All-in-One Surgical Restoration

Our experiments highlight four key observations. **(1) All-in-one surgical restoration training datasets should aim to jointly provide accurate degraded-clean correspondence, faithful physical degradation modeling, and diverse coverage of surgical scenes and degradation severities.** Accurate correspondence and realistic degradation formation are associated with more consistent task-relevant preservation, while broader scene and severity coverage can provide more transferable supervision and may partially compensate for a smaller training-set scale. **(2) All-in-one surgical restoration training composition strategies should prioritize sufficient single-degradation supervision to establish stable restoration anchors under a limited training budget.** Individual degradation mechanisms and severity variations are easier to model and cover through synthesis, enabling more reliable restoration capabilities for each degradation type. In contrast, multi-degradation supervision remains constrained by the domain gap between synthetic and real formation mechanisms, motivating more realistic multi-degradation synthesis for interaction modeling. **(3) All-in-one surgical restoration methods may benefit from jointly preserving fine-grained structures and limiting restoration-induced motion deviations rather than focusing only on stronger single-frame restoration or additional temporal inputs.** Fine-grained structural modeling is associated with better preservation of tissue boundaries and instrument details, while more constrained restoration mappings may reduce unnecessary inter-frame motion deviations. **(4) All-in-one surgical restoration methods using generative priors may benefit from incorporating measurement-consistency constraints and avoiding excessive latent-space compression.** Without measurement-consistency constraints, generative priors may have greater freedom to re-estimate tissue boundaries and local textures beyond the observed evidence, while strong latent compression may further discard high-resolution texture and motion-relevant information.

5 Conclusion

We introduce SurgOmni, a unified dataset and benchmark suite for all-in-one surgical video restoration, comprising the formation-aware paired training dataset and the real-world video testing dataset. Extensive experiments show that reliable real-world generalization benefits from accurate degraded-clean correspondence, faithful degradation modeling, and effective training composition, while robust restoration requires consideration of both surgical structural preservation and restoration-induced motion deviation. These findings provide a practical foundation for training-data construction, training composition strategy design, and restoration method development in future all-in-one surgical restoration.

AI USE STATEMENT

Generative AI tools were used only to improve the language, clarity, and readability of the manuscript. They were not used to formulate the research methodology, construct the datasets, generate or analyze experimental results, or draw scientific conclusions. All content was reviewed and verified by the authors.

ETHICS STATEMENT

Public surgical videos are used in accordance with their respective dataset terms. The clinical colorectal cohort was collected under institutional ethics approval with informed consent, de-identified before use, and made available through an application-based access process. The clinical data are used exclusively for evaluation, and the reported restoration and prediction-consistency results should not be interpreted as evidence of clinical benefit.

REPRODUCIBILITY STATEMENT

Upon publication, we will release the degradation synthesis code, dataset manifests and annotations, model configurations, training hyperparameters, and evaluation protocols. The clinical evaluation data will be available through an application-based access process. Appendix A and B detail dataset construction, Appendix C specifies the training and evaluation protocols, and Appendix D provides the complete results.

REFERENCES

Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James E. East, Georges Wagnières, Victor Loschenov, Enrico Grisan, Walter Blondel, and Jens Rittscher. Endoscopy artifact detection (EAD 2019) challenge dataset. *CoRR*, abs/1905.03209, 2019.

Sharib Ali, Felix Zhou, Adam Bailey, Barbara Braden, James E. East, Xin Lu, and Jens Rittscher. A deep learning framework for quality assessment and restoration in video endoscopy. *Medical Image Anal.*, 68:101900, 2021. doi: 10.1016/J.MEDIA.2020.101900.

Wenhui Chang, Yufeng Li, Zebang Zhu, and Yuchen Yang. LSD3K: A benchmark for smoke removal from laparoscopic surgery images. In *3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology, AIoTC 2024, Wuhan, China, September 13-15, 2024*, pp. 1–5. IEEE, 2024. doi: 10.1109/AIOTC63215.2024.10748251.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Trans. Image Process.*, 33:2404–2418, 2024a. doi: 10.1109/TIP.2024.3378466.

I-Hsiang Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, and Ming-Hsuan Yang. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 17969–17979. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01674.

I-Hsiang Chen, Isma Hadji, Enrique Sanchez, Adrian Bulat, Sy-Yen Kuo, Radu Timofte, Georgios Tzimiropoulos, and Brais Martínez. Restore, assess, repeat: A unified framework for iterative image restoration. *CoRR*, abs/2603.26385, 2026a. doi: 10.48550/ARXIV.2603.26385.

Long Chen, Wen Tang, Nigel W. John, Tao Ruan Wan, and Jian-Jun Zhang. De-smokegcn: Generative cooperative networks for joint surgical smoke detection and removal. *IEEE Trans. Medical Imaging*, 39(5):1615–1625, 2020. doi: 10.1109/TMI.2019.2953717.

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 Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part VI*, volume 15006 of *Lecture Notes in Computer Science*, pp. 369–379. Springer, 2024b. 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. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 3047–3055. AAAI Press, 2026b. 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 Trans. Graph.*, 39(4):75, 2020. doi: 10.1145/3386569.3392457.

Yuning Cui, Wenqi Ren, and Alois Knoll. Bio-inspired image restoration. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025a.

Yuning Cui, Syed Waqas Zamir, Salman H. Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.

Guanglu Dong, Chunlei Li, Chao Ren, Jingliang Hu, Yilei Shi, Xiao Xiang Zhu, and Lichao Mou. Learning domain-aware task prompt representations for multi-domain all-in-one image restoration. *CoRR*, abs/2603.01725, 2026. doi: 10.48550/ARXIV.2603.01725.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2366–2374, 2014.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigün and Tomas Gustavsson (eds.), *Image Analysis, 13th Scandinavian Conference, SCIA 2003, Halmstad, Sweden, June 29 - July 2, 2003, Proceedings*, volume 2749 of *Lecture Notes in Computer Science*, pp. 363–370. Springer, 2003. doi: 10.1007/3-540-45103-X_50.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVIII*, volume 15076 of *Lecture Notes in Computer Science*, pp. 222–241. Springer, 2024a. doi: 10.1007/978-3-031-72649-1_13.

Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 28124–28133. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02619.

Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. Onerestore: A universal restoration framework for composite degradation. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XIX*, volume 15077 of *Lecture Notes in Computer Science*, pp. 255–272. Springer, 2024b. doi: 10.1007/978-3-031-72655-2_15.

Paul Haeberli and Kurt Akeley. The accumulation buffer: hardware support for high-quality rendering. In Forest Baskett (ed.), *Proceedings of the 17th Annual Conference on Computer Graphics*

and Interactive Techniques, SIGGRAPH 1990, Dallas, TX, USA, August 6-10, 1990, pp. 309–318. ACM, 1990. doi: 10.1145/97879.97913.

Tingxuan Hong, Pu Huang, Xiangyu Zhai, Changming Gu, Baolong Tian, Bin Jin, and Dengwang Li. MARS-GAN: multilevel-feature-learning attention-aware based generative adversarial network for removing surgical smoke. *IEEE Trans. Medical Imaging*, 42(8):2299–2312, 2023. doi: 10.1109/TMI.2023.3245298.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, volume 13674 of *Lecture Notes in Computer Science*, pp. 624–642. Springer, 2022. doi: 10.1007/978-3-031-19781-9_36.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

Xingyu Jiang, Xiuhui Zhang, Ning Gao, and Yue Deng. When fast fourier transform meets transformer for image restoration. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLV*, volume 15103 of *Lecture Notes in Computer Science*, pp. 381–402. Springer, 2024. doi: 10.1007/978-3-031-72995-9_22.

Devanish N Kamtam, Joseph B Shrager, Satya Deepya Malla, Xiaohan Wang, Nicole Lin, Juan J Cardona, Serena Yeung-Levy, and Clarence Hu. A fine-tuned foundational model surgisam2 for surgical video anatomy segmentation and detection. *Scientific Reports*, 15(1):35961, 2025.

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 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 5128–5137. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00510.

Zohaib Amjad Khan, Azeddine Beghdadi, Faouzi Alaya Cheikh, Mounir Kaaniche, Egidijus Pelanis, Rafael Palomar, Åsmund Avdem Fretland, Bjørn Edwin, and Ole Jakob Elle. Towards a video quality assessment based framework for enhancement of laparoscopic videos. In Frank W. Samuelson and Sian Taylor-Phillips (eds.), *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, Houston, TX, USA, February 15-20, 2020*, volume 11316 of *SPIE Proceedings*, pp. 113160P. SPIE, 2020. doi: 10.1117/12.2549266.

Yuto Kubo, Takayuki Yasui, Yuko Matsuda, Yasutomo Takahashi, Kotaro Yamashita, Takuro Saito, Koji Tanaka, Tomoki Makino, Tsuyoshi Takahashi, Yukinori Kurokawa, et al. Mechanism underlying lens fogging and its countermeasure in laparoscopic surgery. *Minimally Invasive Therapy & Allied Technologies*, 32(2):56–61, 2023.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 17431–17441. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01693.

Hao Li, Xiangyu Zhai, Jie Xue, Changming Gu, Baolong Tian, Tingxuan Hong, Bin Jin, Dengwang Li, and Pu Huang. Multi-frequency and smoke attention-aware learning based diffusion model for removing surgical smoke. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part I*, volume 15001 of *Lecture Notes in Computer Science*, pp. 47–56. Springer, 2024. doi: 10.1007/978-3-031-72378-0_5.

Wenrui Li, Hongtao Chen, Yao Xiao, Wangmeng Zuo, Jiantao Zhou, Yonghong Tian, and Xiaopeng Fan. All-in-one video restoration under smoothly evolving unknown weather degradations. *CoRR*, abs/2601.00533, 2026. doi: 10.48550/ARXIV.2601.00533.

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. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 6844–6852. AAAI Press, 2026. doi: 10.1609/AAAI.V40I9.37617.

Jingbo Lin, Zhilu Zhang, Wenbo Li, Renjing Pei, Hang Xu, Hongzhi Zhang, and Wangmeng Zuo. Unirestorer: Universal image restoration via adaptively estimating image degradation at proper granularity. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust (eds.), *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026.

Ziteng Liu, Wenpeng Gao, Jiahua Zhu, Bainan Liu, and Yili Fu. Smoke attention based laparoscopic image desmoking network with hybrid guided embedding. In *IEEE International Conference on Robotics and Biomimetics, ROBIO 2024, Bangkok, Thailand, December 10-14, 2024*, pp. 1018–1023. IEEE, 2024. doi: 10.1109/ROBIO64047.2024.10907529.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Controlling vision-language models for multi-task image restoration. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

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

Venkatanath N., Praneeth D., Maruthi Chandrasekhar Bh., Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. In *Twenty First National Conference on Communications, NCC 2015, Mumbai, India, February 27 - March 1, 2015*, pp. 1–6. IEEE, 2015. doi: 10.1109/NCC.2015.7084843.

Chinedu Innocent Nwoye, Kareem Elgohary, Anvita Srinivas, Fauzan Zaid, Joël L. Lavanchy, and Nicolas Padoy. Cholectrack20: A multi-perspective tracking dataset for surgical tools. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 8942–8952. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00836.

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. *Int. J. Comput. Assist. Radiol. Surg.*, 17(5):885–893, 2022. doi: 10.1007/S11548-022-02595-2.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 37413–37422. Computer Vision Foundation, 2026.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H. Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Sudarshan Rajagopalan, Kartik Narayan, and Vishal M. Patel. Restorevar: Visual autoregressive generation for all-in-one image restoration. *CoRR*, abs/2505.18047, 2025. doi: 10.48550/ARXIV.2505.18047.

Xiaole Tang, Xiang Gu, Xiaoyi He, Xin Hu, and Jian Sun. Degradation-aware residual-conditioned optimal transport for unified image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(8):6764–6779, 2025. doi: 10.1109/TPAMI.2025.3562211.

Xiaole Tang, Xiaoyi He, Jiayi Xu, Xiang Gu, and Jian Sun. Learning continuous wasserstein barycenter space for generalized all-in-one image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(7):8395–8410, 2026. doi: 10.1109/TPAMI.2026.3669121.

Qingyao Tian, Zhen Chen, Huai Liao, Xinyan Huang, Lujie Li, Sébastien Ourselin, and Hongbin Liu. Endoomni: Zero-shot cross-dataset depth estimation in endoscopy by robust self-learning from noisy labels. *CoRR*, abs/2409.05442, 2024. doi: 10.48550/ARXIV.2409.05442.

Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 28165–28175. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02623.

Andru Putra Twinanda, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Single- and multi-task architectures for tool presence detection challenge at M2CAI 2016. *CoRR*, abs/1610.08851, 2016.

Vishal Venkatesh, Neeraj Sharma, Vivek Srivastava, and Munendra Singh. Unsupervised smoke to desmoked laparoscopic surgery images using contrast driven cyclic-desmokegan. *Comput. Biol. Medicine*, 123:103873, 2020. doi: 10.1016/J.COMPBIOMED.2020.103873.

An Wang, Haochen Yin, Beilei Cui, Mengya Xu, and Hongliang Ren. Benchmarking robustness of endoscopic depth estimation with synthetically corrupted data. In Virginia Fernandez, Jelmer M. Wolterink, David Wiesner, Samuel Remedios, Lianrui Zuo, and Adrià Casamitjana (eds.), *Simulation and Synthesis in Medical Imaging - 9th International Workshop, SASHIMI 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 10, 2024, Proceedings*, volume 15187 of *Lecture Notes in Computer Science*, pp. 45–57. Springer, 2024. doi: 10.1007/978-3-031-73281-2_5.

Congcong Wang, Ahmed Kedir Mohammed, Faouzi Alaya Cheikh, Azeddine Beghdadi, and Ole Jacob Elle. Multiscale deep desmoking for laparoscopic surgery. In Elsa D. Angelini and Bennett A. Landman (eds.), *Medical Imaging 2019: Image Processing, San Diego, California, United States, 16-21 February 2019*, volume 10949 of *SPIE Proceedings*, pp. 109491Y. SPIE, 2019. doi: 10.1117/12.2507822.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029. PMLR / OpenReview.net, 2024a.

Jiawei Wu, Zhifei Yang, Zhe Wang, and Zhi Jin. Gradient as conditions: Rethinking HOG for all-in-one image restoration. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 10682–10690. AAAI Press, 2026. doi: 10.1609/AAAI.V40I13.38042.

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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXII*, volume 15130 of *Lecture Notes in Computer Science*, pp. 307–324. Springer, 2024b. doi: 10.1007/978-3-031-73220-1_18.

Wenyao Xia, Victoria Fan, Terry M. Peters, and Elvis C. S. Chen. A new benchmark in vivo paired dataset for laparoscopic image de-smoking. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part I*, volume 15001 of *Lecture Notes in Computer Science*, pp. 3–13. Springer, 2024. doi: 10.1007/978-3-031-72378-0_1.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 12753–12763. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01190.

Jiale Zhang, Wenfeng Huang, Xiangyun Liao, and Qiong Wang. Progressive frequency-aware network for laparoscopic image desmoking. In Qingshan Liu, Hanzi Wang, Zhanyu Ma, Weishi Zheng, Hongbin Zha, Xilin Chen, Liang Wang, and Rongrong Ji (eds.), *Pattern Recognition and Computer Vision - 6th Chinese Conference, PRCV 2023, Xiamen, China, October 13-15, 2023, Proceedings, Part II*, volume 14426 of *Lecture Notes in Computer Science*, pp. 479–492. Springer, 2023. doi: 10.1007/978-981-99-8432-9_38.

Haiyu Zhao, Lei Tian, Xinyan Xiao, Peng Hu, Yuanbiao Gou, and Xi Peng. Avernet: All-in-one video restoration for time-varying unknown degradations. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

SurgOmni: Towards All-in-One Surgical Video Restoration

Appendices

A OmniSyn Construction and Statistics . . . 17

A.1 Clean Video Curation and Preprocessing . . . 17

A.2 Formation-Aware Degradation Synthesis . . . 17

A.2.1 Single-Degradation Synthesis . . . 17

A.2.2 Multi-Degradation Synthesis . . . 18

A.3 Quality Control . . . 18

A.4 Dataset Distribution . . . 19

B OmniBench Curation and Statistics . . . 19

B.1 Video Collection and Clip Selection . . . 19

B.2 Degradation Annotation . . . 20

B.3 Dataset Distribution . . . 20

C Experimental Protocols . . . 20

C.1 Restoration Model Implementations . . . 20

C.2 Training Dataset Comparison Setup . . . 22

C.3 Evaluation Metrics and Statistical Analysis . . . 22

C.4 Efficiency Evaluation Setup . . . 23

C.5 Downstream Evaluation Setup . . . 23

C.6 Surgeon Assessment Protocol . . . 24

D Quantitative Analysis . . . 24

D.1 Training Dataset Comparison . . . 24

D.1.1 Restoration Results . . . 24

D.1.2 Downstream Prediction-Consistency Results . . . 26

D.2 Metric Agreement Analysis . . . 26

D.3 Surgeon Assessment Results . . . 28

D.4 Training Composition Analysis . . . 28

D.4.1 OmniBench Results . . . 28

D.4.2 OmniSyn Results . . . 29

E Qualitative Analysis . . . 31

E.1 Training Dataset Comparisons . . . 31

E.2 Benchmark Restoration Comparisons . . . 33

F Scope, Ethics, and Reproducibility . . . 33

F.1 Benchmark Scope and Future Directions . . . 33

F.2 Ethics and Data Governance . . . 33

F.3 Reproducibility and Release . . . 34

A OmniSyn Construction and Statistics

A.1 Clean Video Curation and Preprocessing

OmniSyn is built from temporally continuous clean sequences collected from CholecTrack20 (Nwoye et al., 2025) and M2CAI16-TOOL (Twinanda et al., 2016). Candidate clips are screened over four-second windows to exclude visible target degradations, severe camera shake, large-area obstruction, abnormal exposure, and persistent loss of anatomical visibility. Natural viewpoint changes, instrument motion, tissue deformation, and mild specular reflections are retained as long as the operative field remains clearly interpretable, preserving genuine surgical dynamics in the clean references. FFmpeg is used to standardize the frame rate and video format. Static invalid peripheral regions are identified once for each clip and removed with a fixed crop. The remaining valid region is then proportionally resized and center-cropped to 1080×720. The same spatial transform is applied throughout each clip and reused for its degraded counterpart, preserving temporal geometry and exact degraded-clean correspondence.

A.2 Formation-Aware Degradation Synthesis

A.2.1 Single-Degradation Synthesis

OmniSyn synthesizes five representative surgical degradations according to their formation stages in the laparoscopic imaging process. Smoke, haze, and splash are generated through the Blender 3D rendering engine, where smoke is modeled as a scene-space aerosol, haze as near-lens condensation, and splash as lens-adherent fluid. Motion blur and non-uniform illumination (NUI) are generated through code-based synthesis. Motion blur is formed by temporally densifying the clean sequence with RIFE (Huang et al., 2022) and integrating consecutive frames over a virtual exposure interval, while NUI is modeled through spatial photometric transformations that reproduce local over-illumination and spatially varying under-illumination.

Haze. Haze is modeled as condensation attached to the laparoscope lens rather than as volumetric scattering within the surgical scene, consistent with the formation mechanism of laparoscopic lens fogging (Kubo et al., 2023). We construct a near-lens condensation layer in Blender using Geometry Nodes, which distributes droplets of varying scales over a disk covering the valid field of view. The resulting droplets are rendered with semi-transparent materials whose optical properties control the local attenuation and whitening caused by condensation. The condensation pattern remains fixed in camera coordinates throughout each clip, consistent with fog attached to the laparoscope lens. Gaussian blurring and glow-based compositing then soften the individual droplet responses into a continuous haze layer.

Smoke. Smoke synthesis adapts the smoke-type-aware rendering framework of STSVD (Liang et al., 2026) to the curated clean sequences in OmniSyn. We retain its diffusion, ambient, and entangled smoke modes and rerender them within the Blender-based pipeline to maintain exact correspondence with the clean clips. Diffusion smoke is generated from a localized emitter placed near the detected working end of a visible surgical instrument, with its emission and turbulence parameters governing plume evolution. Ambient smoke represents accumulated aerosol with broader coverage and weaker directionality, while the entangled mode combines a newly generated local plume with an existing ambient field to model their coexistence during repeated cauterization. The smoke field evolves continuously throughout each clip before being composited with the corresponding clean sequence, preserving its transport and deformation across adjacent frames.

Splash. Splash is modeled as lens-adherent fluid using a Blender particle system. Particle sampling determines the spatial distribution of the contamination, while simulated dynamics govern droplet motion. The particles are converted into continuous fluid surfaces using Metaball geometry, allowing neighboring droplets to merge into irregular adherent structures rather than remaining as isolated image-space overlays. A transparent refractive material is applied to the resulting surface, causing local displacement and attenuation of the underlying surgical content according to the droplet geometry. The simulation evolves continuously throughout each clip before compositing with the corresponding clean sequence, preserving coherent changes in splash shape and position over time.

Motion Blur. Motion blur is synthesized through temporal exposure integration rather than by applying a fixed spatial blur kernel independently to each frame. For an output frame at time t, the exposure process can be expressed as

$$B_t(p) = \frac{1}{t_e - t_s} \int_{t_s}^{t_e} I(p, \tau)\, d\tau, \tag{1}$$

where p denotes a pixel location, $[t_s, t_e]$ is the exposure interval, $I(p, \tau)$ is the instantaneous RGB observation at time τ, and $B_t(p)$ is the resulting motion-blurred value. Because the source videos

contain only discrete temporal samples, we first densify each sequence to $16\times$ its original frame rate using RIFE (Huang et al., 2022). Consecutive interpolated frames are then averaged with FFmpeg to approximate the exposure integral (Haeberli & Akeley, 1990). After temporal integration, the sequence is returned to the original frame rate while preserving frame-wise correspondence with the clean video. Since the integration follows the motion present in the source sequence, fast-moving structures exhibit stronger motion smearing, while nearly stationary regions remain comparatively sharp.

Non-uniform Illumination. Non-uniform illumination (NUI) is synthesized using separate over- and under-illumination models. For over-illumination, bright regions are identified from the luminance of each frame and spatially expanded with Gaussian filtering before being composited back into the image. The resulting saturation and glare therefore follow the visible tissue and instrument content rather than a predefined spatial mask. Under-illumination is modeled using a spatial attenuation field,

$$V(p) = (1 - \lambda) + \lambda G(p), \tag{2}$$

$$I_{\mathrm{low},t}(p) = \mathrm{clip}(a\, x_t(p) V(p), 0, 1)\,, \tag{3}$$

where p denotes a pixel location, $x_t(p)$ is the clean RGB value at frame t, and $G(p) \in [0, 1]$ is a normalized Gaussian field centered on the image. $V(p)$ is the resulting illumination field, with $\lambda \in [0, 1]$ controlling peripheral attenuation and a controlling the global brightness level. $I_{\mathrm{low},t}(p)$ denotes the synthesized under-illuminated output, and $\mathrm{clip}(\cdot, 0, 1)$ constrains its intensity to the valid range. The photometric parameters remain fixed throughout each clip, while the over-illumination mask is recomputed from each frame so that its spatial support follows changes in the observed scene.

Severity Control. Since the five degradations arise from different formation processes, their severity is adjusted through mechanism-specific parameters rather than a common image-space intensity scale. The parameter ranges are chosen to cover mild-to-severe conditions, informed by the multi-level degradation settings and distributions reported in prior laparoscopic imaging studies (Wang et al., 2019; Khan et al., 2020; Wang et al., 2024). Haze is strengthened by increasing the extent of lens condensation and reducing its effective transmission. Smoke severity is varied through the density and spatial coverage of the rendered aerosol, whereas splash is controlled by the extent of lens-adherent fluid. Motion blur is adjusted through the temporal integration window, with longer exposure intervals producing stronger motion smearing. For NUI, over-illumination is strengthened by increasing localized saturation, while under-illumination is intensified through stronger spatial attenuation. The sampled ranges are visually checked to ensure that different settings remain distinguishable while retaining sufficient anatomical visibility for restoration.

A.2.2 Multi-Degradation Synthesis

To explore the effect of multi-degradation supervision, we construct a dedicated subset by applying multiple formation-aware degradation processes to the same clean sequence, which serves as their shared reference. During composition, each active factor retains the spatial and temporal characteristics defined in the single-degradation pipeline. Smoke continues to evolve in scene space, haze and splash remain anchored to the camera, NUI follows the corresponding illumination process, and motion blur is generated through temporal exposure integration. The subset contains seven dual-degradation configurations, {Smoke+Haze}, {Smoke+NUI}, {Haze+NUI}, {Blur+Haze}, {Blur+Smoke}, {Splash+Haze}, and {Splash+Smoke}, together with two triple-degradation configurations, {Smoke+Haze+Splash} and {Smoke+Haze+NUI}. We do not exhaustively enumerate all possible combinations, but instead select co-occurrences that are plausible in laparoscopic imaging and consistent with degradation patterns reported in real surgical data (Ali et al., 2019; Chen et al., 2026b). This keeps the subset focused on representative degradation interactions relevant to the training-composition study.

A.3 Quality Control

Quality control is applied to both clean-video curation and degradation synthesis. Candidate clean clips are reviewed in full by two reviewers with experience in surgical restoration and retained only when the target degradations are absent while normal surgical dynamics remain clearly interpretable. Disagreements are re-examined jointly, and unresolved cases are referred to a clinician for final adjudication. After synthesis, every generated clip first undergoes integrity checks for valid decoding and exact degraded-clean correspondence, followed by cross-checking by three reviewers with experience in surgical restoration. They verify that the rendered degradation is consistent with its intended formation mechanism and remains temporally coherent throughout the sequence. Clips with implausible artifacts or excessive obstruction are rejected.

A.4 Dataset Distribution

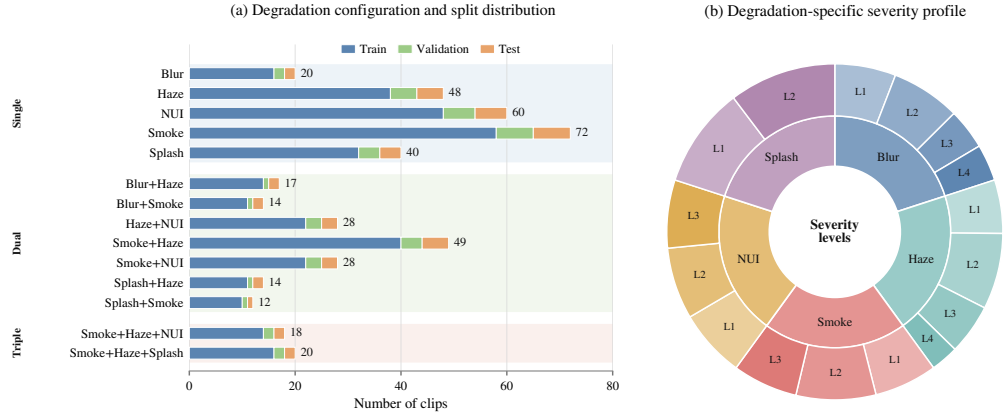


Figure 8: Distribution of OmniSyn degradation configurations and severity levels. (a) Clip counts across the training, validation, and test splits. (b) Degradation-specific severity distributions, where arc length indicates the proportion of samples at each factor-specific severity level.

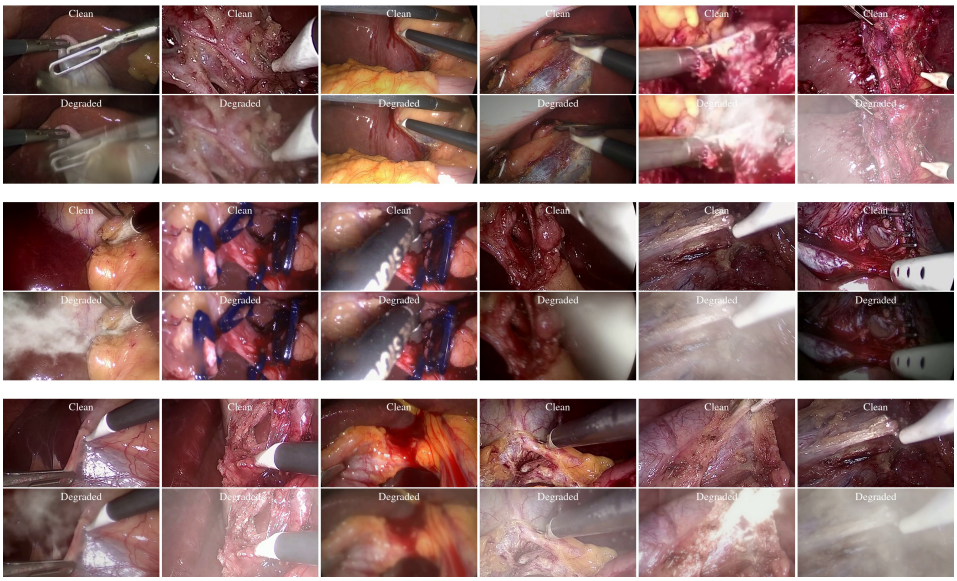


Figure 9: Representative clean-degraded pairs from OmniSyn covering all five single-degradation types, mode-specific variants of NUI, smoke, and splash, and all seven dual- and two triple-degradation configurations.

OmniSyn contains 44,000 paired frames in 440 videos, comprising 240 single-degradation clips and 200 multi-degradation clips, including 162 dual- and 38 triple-degradation clips. Partition assignments are determined for the curated clean clips before degradation synthesis, and every synthetic variant derived from a clean clip inherits the same assignment. Configuration-wise stratified allocation under the target 8:1:1 ratio yields 192 training, 24 validation, and 24 test clips for single degradations, together with 160 training, 18 validation, and 22 test clips for multiple degradations after integer allocation within each configuration. The resulting partitions contain 352 training, 42 validation, and 46 test clips overall. All training-composition subsets are subsequently constructed from the 352-clip training partition, after the validation and test clips have been excluded. Figure 8 reports the configuration-wise split distribution and the degradation-specific severity profiles. Because severity is controlled by different formation parameters for each degradation, these levels are interpreted only within the same factor rather than as a shared scale across degradation types.

B OmniBench Curation and Statistics

B.1 Video Collection and Clip Selection

OmniBench combines public gallbladder recordings from CholecTrack20 (Nwoye et al., 2025) and M2CAI16-TOOL (Twinanda et al., 2016) with a clinical colorectal cohort collected at a hospital.

For public recordings, OmniSyn and OmniBench are separated at the source-video level, ensuring that no original recording contributes to both datasets and preventing direct content leakage or near-duplicate evaluation. Candidate recordings are reviewed in temporal context to identify intervals containing clearly visible degradations. Clips are retained only when the degradation remains identifiable and sufficient anatomical or instrumental structure is visible for meaningful restoration assessment. We exclude segments with severe defocus, prolonged loss of visibility, non-operative content, or ambiguous degradation. To limit redundancy, selected clips are temporally separated within the same recording and sampled across different source videos whenever possible. In addition, both public datasets contain laparoscopic cholecystectomy and therefore contribute one procedure type. The clinical cohort includes low anterior sphincter-preserving and radical rectal cancer resections, sigmoid colectomy, left hemicolectomy, total colectomy, transverse colectomy, and other rectal procedures. For benchmark reporting, the three rectal labels are grouped under rectal surgery, while the four colectomy procedures remain separate. The resulting six types are laparoscopic cholecystectomy, rectal surgery, sigmoid colectomy, left hemicolectomy, total colectomy, and transverse colectomy.

B.2 Degradation Annotation

OmniBench adopts clip-level multi-label annotation using the same five-factor taxonomy as OmniSyn. All degradation factors that remain consistently identifiable over a coherent temporal interval are retained rather than assigning each clip only a dominant label. Ambiguous cases are resolved by considering spatial appearance together with temporal behavior. Smoke is identified by its scene-space drift and diffusion, whereas haze remains relatively stable in camera coordinates. Splash requires persistent lens-attached fluid structures and is distinguished from isolated specular highlights. Motion blur is confirmed from directional smearing across adjacent frames, while NUI is labeled only when spatially uneven over- or under-illumination causes a clear reduction in visibility. Each clip is independently reviewed over its full duration by three reviewers with experience in surgical restoration. Their annotations are compared for factor presence and label completeness. Disagreements are resolved through joint review, with unresolved cases referred to clinicians for final adjudication.

B.3 Dataset Distribution

Figure 10 summarizes the degradation configurations in the Gallbladder and Colon subsets of OmniBench, which contain 60 and 40 clips, respectively. These subset names denote their primary surgical groupings rather than individual procedures or anatomical targets, and the benchmark as a whole spans 6 surgical types, 10 target organs, and 27 surgical scenes. Across the 100 videos, 43 contain single degradations, 45 contain dual degradations, and 12 contain triple degradations. Both subsets cover all five single-degradation factors, while their naturally occurring multi-degradation configurations overlap in four dual and one triple combinations.

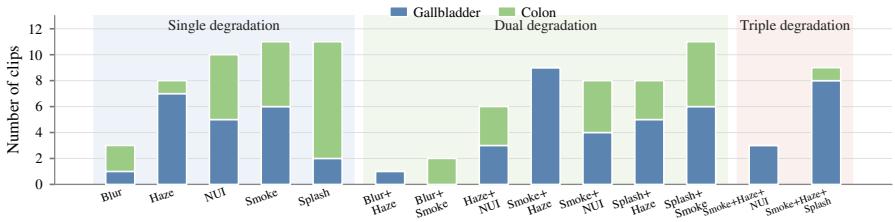


Figure 10: Distribution of degradation configurations across the Gallbladder and Colon subsets of OmniBench.

Figure 11 shows representative real-world samples from the Gallbladder and Colon subsets of OmniBench, covering both single- and multi-degradation cases. The examples illustrate diverse degradation appearances across surgical scenes while retaining sufficient anatomical or instrumental structure for restoration evaluation.

C Experimental Protocols

C.1 Restoration Model Implementations

We evaluate 15 all-in-one image and video restoration methods spanning CNN, Transformer, state-space, Fourier-based, diffusion, and flow-matching architectures. Each implementation is based on its official code and configuration, with adaptations limited to the data interface and distributed execution. The model architecture, training objective, and inference rule remain unchanged, and released pretrained components are retained when required by the method. Paired inputs and targets undergo synchronized random cropping and geometric augmentation. Cross-method benchmark-

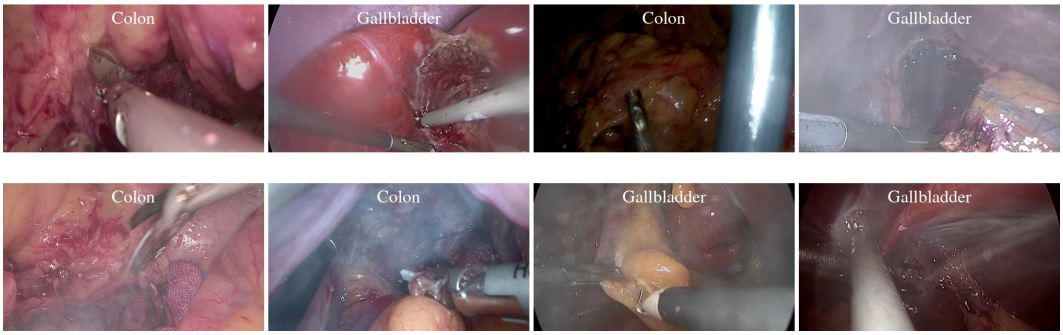


Figure 11: Representative real-world testing samples from OmniBench across single- and multi-degradation configurations in the Gallbladder and Colon subsets.

ing retains the architecture-specific optimization recipe of each method. Within every controlled comparison, the same architecture retains its implementation, initialization, learning objective, base optimization configuration, preprocessing, and inference procedure, leaving the sampled training data as the experimental variable.

Table 4: Core architecture-specific configurations used for the "Single" training condition. The resulting checkpoints are shared by the benchmark in Table 2 and the "Single" entries in Table 3. Input denotes one image or the number of consecutive video frames.

Method	Input	Spatial input	Optimizer / LR	Training objective
PromptIR	Image	128×128	AdamW / 2×10^{-4}	L1
SFHformer	Image	256×256	AdamW / 6×10^{-4}	L1 + frequency loss
MambaIR	Image	128×128	Adam / 2×10^{-4}	Charbonnier
AverNet	12 frames	256×256	Adam / 1×10^{-4}	Charbonnier
BioIR	Image	128×128	AdamW / 2×10^{-4}	L1 + frequency loss
MoCE	Image	128×128	AdamW / 2×10^{-4}	L1 + frequency and routing-balance losses
AdaIR	Image	128×128	AdamW / 2×10^{-4}	L1
DA-RCOT	Image	128×128	RMSprop / 1×10^{-4}	L1 + residual OT and adversarial losses
MambaIRv2	Image	128×128	Adam / 2×10^{-4}	Charbonnier
ORCANet	12 frames	256×256	Adam / 1×10^{-4}	Charbonnier + degradation-aware auxiliary losses
EndoIR	Image	256×256	Adam / 2×10^{-4}	SmoothL1 for signal and noise branches
RAR	Image	256×256	CAME / 3.54×10^{-5}	Recursive flow matching
HOGformer	Image	128×128	AdamW / 2×10^{-4}	L1 + HOG and correlation losses
BaryIR	Image	128×128	RMSprop / 2×10^{-4}	L1 + barycenter regularization
DATPRLIR	Image	128×128	AdamW / 2×10^{-4}	L1 + frequency loss

CNN. BioIR (Cui et al., 2025a) adopts a convolution-dominant U-shaped architecture and exchanges large-field context with fine-grained details through peripheral-to-foveal and foveal-to-peripheral interactions. AverNet (Zhao et al., 2024) extends convolutional restoration to video using bidirectional feature propagation and prompt-guided deformable alignment for time-varying degradations. ORCANet (Li et al., 2026) adopts recurrent video restoration with flow-guided propagation and static and dynamic prompts that model longer-term degradation identity and frame-level variation. Both video methods retain their native temporal input protocols.

Transformer. PromptIR (Potlapalli et al., 2023) conditions a shared Transformer backbone on learned degradation prompts. MoCE (Zamfir et al., 2025) introduces experts with different receptive fields and computational complexity for degradation-adaptive routing. HOGformer (Wu et al., 2026) incorporates learnable HOG representations into self-attention to strengthen gradient-aware structural modeling. DA-RCOT (Tang et al., 2025) and BaryIR (Tang et al., 2026) use Transformer restoration backbones with residual-conditioned optimal transport and Wasserstein-barycenter learning, respectively, to constrain the restoration mapping. DATPRLIR (Dong et al., 2026) instead introduces degradation-aware prompt representations that adapt the shared backbone across restoration domains.

State-space. MambaIR (Guo et al., 2024a) adapts selective state-space modeling to image restoration while supplementing long-range dependency modeling with local enhancement and channel attention. MambaIRv2 (Guo et al., 2025) further introduces attentive state-space modeling to relax causal scanning and strengthen interactions between distant but semantically related pixels.

Fourier-based. SFHformer (Jiang et al., 2024) combines spatial Transformer processing with Fourier-domain modeling to jointly capture local structure and global frequency information.

AdaIR (Cui et al., 2025b) performs degradation-adaptive frequency mining and modulation, allowing frequency components to be adjusted according to the observed degradation.

Generative. EndoIR (Chen et al., 2026b) performs all-in-one endoscopic restoration through degradation-aware conditional diffusion with dual-domain prompting and noise-aware routing. RAR (Chen et al., 2026a) operates in a pretrained latent space and combines noise-free flow matching with iterative degradation assessment, refining the restoration according to the estimated residual degradation.

C.2 Training Dataset Comparison Setup

SurgClean. SurgClean (Pei et al., 2026) is a real-world surgical image restoration dataset covering smoke, haze, and splash. It preserves authentic degradation-scene interactions and provides temporally adjacent degraded and clean frames. Following the official protocol, we align these pairs once before training. The clean frame is first resized to the degraded input resolution, after which a fixed RAFT-Large model estimates bidirectional optical flow and warps the clean frame to the input geometry. Forward-backward consistency is recorded only as an alignment-quality diagnostic and does not weight the training loss. The resulting aligned targets are then processed by the same architecture-specific training pipeline used for OmniSyn-S, without online warping or an additional masking loss. This design keeps the restoration model, optimization budget, and inference procedure unchanged within each architecture while preserving the pairing procedure required by the two training sources. Accordingly, the comparison isolates the supervision provided by each training source under a controlled restoration learner rather than comparing source-specific training algorithms.

LVQ. The ICIP 2020 LVQ dataset (Khan et al., 2020) is a synthetic laparoscopic video dataset. It contains five distortion types, of which smoke, motion blur, and uneven illumination overlap with the SurgOmni taxonomy. AWGN and defocus blur were introduced in LVQ as distortion categories for laparoscopic video quality assessment and distortion classification, which fall outside the degradation taxonomy considered in the restoration setting. We therefore exclude them to maintain aligned degradation types in the comparison. LVQ provides exact degraded-clean correspondence, but its degradations are synthesized in image space and cover a relatively limited set of surgical scenes and discrete severity levels.

For the matched-budget comparisons, we construct OmniSyn-S and OmniSyn-L from single-degradation clips in the official training partition to match the scale and degradation type of SurgClean and the LVQ dataset, respectively. OmniSyn-S contains 2,488 pairs covering smoke, haze, and splash. OmniSyn-L contains 10,000 pairs, including smoke, NUI, and motion-blur samples. Within each degradation type, samples are distributed as evenly as possible across the available severity settings to avoid bias toward particular intensity ranges. Validation and test clips are excluded before either subset is sampled. For each comparison, models with the same architecture use identical initialization, preprocessing, optimization, and inference settings.

Training Composition Comparisons. The composition study is constructed exclusively from the official OmniSyn training partition. Its single-degradation pool contains 192 clips. "Single" selects 160 of these clips through stratified sampling, with the samples distributed as evenly as possible across the five degradation types and their available severity levels. The multi-degradation training pool contains 160 clips, all of which are used by "Multi". "Mixed" contains 80 single-degradation and 80 multi-degradation clips sampled from the same training-only pools. Each condition therefore contains 160 clips and is trained under the same data budget. The three sample lists are fixed across all 15 methods, and no validation or test clip is used to construct any training condition. Within each architecture, the model implementation, initialization, learning objective, base optimization configuration, preprocessing, and inference procedure remain unchanged across the three compositions. Evaluation uses the fixed held-out single- and multi-degradation test subsets defined by the official split. The "Single" condition also supplies the checkpoints used for the benchmark in Table 2. Its degradation-stratified scores are reported there, while the "Single" block of Table 3 aggregates predictions from the same checkpoints over the full OmniBench test set. These results therefore do not arise from a second benchmark training run.

C.3 Evaluation Metrics and Statistical Analysis

Restoration Quality. We use three learned no-reference metrics that assess complementary aspects of perceptual quality. MUSIQ (Ke et al., 2021) represents an image at multiple spatial scales and jointly models quality cues across different resolutions. TOPIQ (Chen et al., 2024a) uses cross-scale attention to propagate high-level semantic information toward lower-level distortion features, making its assessment sensitive to content-dependent local degradations. Q-Align (Wu et al., 2024a) aligns a multimodal large language model with discrete text-defined quality levels and converts the

resulting predictions into fine-grained perceptual scores. The three metrics are reported independently to preserve these complementary assessment perspectives. NIQE (Mittal et al., 2013) and PIQE (N. et al., 2015) are not used as primary metrics because their hand-crafted assessment mechanisms are less stable across diverse surgical scenarios. NIQE measures deviations from a predefined statistical model, which can confound restoration-induced changes with variations caused by tissue structure, specular reflections, and instrument appearance. PIQE relies on fixed activity and distortion thresholds to identify locally distorted regions, making its responses sensitive to the substantial contrast and illumination variations commonly observed in laparoscopic imagery.

Temporal Behavior. Temporal behavior is evaluated using tOF (Chu et al., 2020). In our setting, tOF is computed as the mean endpoint discrepancy between optical-flow fields estimated from consecutive restored and reference frames using Farnebäck optical flow (Farnebäck, 2003):

$$\mathrm{tOF}(\hat{x}, r) = \frac{1}{(T-1)|\Omega|} \sum_{t=1}^{T-1} \sum_{p \in \Omega} \left\| \mathcal{F}(\hat{x}_t, \hat{x}_{t+1})(p) - \mathcal{F}(r_t, r_{t+1})(p) \right\|_2 , \tag{4}$$

where T denotes the sequence length, Ω the image domain, p a pixel location, and $\mathcal{F}$ the optical-flow estimator. On OmniSyn, r is the paired clean sequence, so tOF measures deviation from clean-reference motion. On OmniBench, where no clean counterpart is available, r is the degraded input sequence and tOF measures how much restoration changes the motion observed in the input. These two quantities have different interpretations. In particular, an identity mapping has zero input-referenced tOF by construction, while removing moving smoke or shifting splash can legitimately alter the estimated flow. We therefore do not treat input-referenced tOF as an estimate of clean-motion fidelity or as a standalone measure of restoration quality. It is used only to characterize restoration-induced motion deviation alongside TOPIQ, MUSIQ, and Q-Align, and its absolute values are not compared with clean-reference tOF across datasets.

Downstream Prediction Consistency. Scene and instrument segmentation are evaluated using Dice (Dice, 1945) and IoU (Jaccard, 1901), which measure agreement between restored-frame predictions and their clean-frame references through region overlap. For monocular depth, δ_1 reports the fraction of predictions within the standard 1.25 multiplicative threshold, while AbsRel and SqRel quantify relative depth error with linear and squared weighting. RMSE measures the absolute magnitude of the prediction error (Eigen et al., 2014).

Statistical Analysis. Video clips are used as the aggregation units. Frame-level measurements are first averaged within each clip so that longer clips do not receive greater weight. Gallbladder and Colon results are reported separately, while the overall OmniBench result is obtained by averaging the eligible clip-level scores across both subsets. Each result is computed independently for the three training runs before the run-level values are averaged. The same aggregation is used for the no-reference control on the 22 held-out OmniSyn multi-degradation clips reported in Table 19. The rank-agreement analysis in Table 13 uses these same clips, "Single" checkpoints, and restored sequences. Spearman's ρ and Kendall's τ_b are computed across the 15 unrounded method-level means.

C.4 Efficiency Evaluation Setup

Efficiency is evaluated using parameter count, MACs, and end-to-end inference latency on a single NVIDIA L40S GPU. Latency is measured at 1080×720 after model warm-up and covers each method's complete native inference path, including internal resizing or iterative processing when required. CUDA synchronization is applied immediately before and after timing to avoid bias from asynchronous execution. MACs are computed at 256×256 to provide a consistent measure of computational complexity across methods. Parameter counts include every module required at inference time, regardless of whether its weights are trainable or frozen. Accordingly, the parameter count reported for RAR includes its frozen pretrained foundation model and is marked with ‡ in the efficiency table.

C.5 Downstream Evaluation Setup

Prediction consistency is evaluated on the paired OmniSyn test set using frozen surgical perception models, with the clean-frame prediction serving as the reference. Model weights, valid frames, prompts, and task-specific processing remain fixed across restoration settings, so the reported comparisons isolate differences in how consistently each setting preserves the predictions obtained from clean frames.

Scene Segmentation. SurgiSAM2 (Kamtam et al., 2025) is used for scene segmentation. Candidate regions are obtained from the clean frame, and one interior point from each valid region is reused

at the same coordinate for the degraded and restored counterparts. Dice and IoU are then computed against the clean-frame prediction under identical spatial prompts.

Instrument Segmentation. MedSAM (Ma et al., 2024) is used for prompt-conditioned instrument segmentation. Bounding boxes derived from the clean frame are reused unchanged for the corresponding degraded and restored frames. The clean-frame prediction provides the reference, and Dice and IoU are computed over the same set of valid prompts for all restoration methods.

Monocular Depth. EndoOmni (Tian et al., 2024) provides monocular depth predictions. To account for the global scale ambiguity of monocular depth, each restored prediction is median-aligned to its paired clean prediction before evaluation:

$$\tilde{z}(p) = \hat{z}(p)\frac{\mathrm{median}_{p\in\Omega}\, z(p)}{\mathrm{median}_{p\in\Omega}\, \hat{z}(p)}, \tag{5}$$

where p denotes a valid pixel location, Ω is the valid image domain, z is the clean-frame prediction, and $\hat{z}$ is the restored prediction. Agreement is measured using δ_1, AbsRel, SqRel, and RMSE.

For all three tasks, the same valid frames and task-specific references are retained across restoration methods. Segmentation scores are averaged within each frame before frame-to-clip aggregation, while depth follows the same clip-level evaluation hierarchy.

C.6 Surgeon Assessment Protocol

Readers and Assessment Task. Six attending surgeons independently completed the assessment, each with more than 10 years of laparoscopic experience. Three primarily practiced hepatobiliary surgery and three colorectal surgery. Surgeons evaluated synthetic degradation realism based on visual appearance, spatial placement, and temporal evolution.

Calibration and Rating Scale. Before the assessment, each surgeon viewed 15 real OmniBench clips covering smoke, motion blur, and non-uniform illumination, with five clips per degradation type. These real clips served as score-5 calibration references and were excluded from the test analysis. Synthetic clips were then rated on a five-point scale, where 1 indicated a clearly implausible artifact, 2 a recognizable but artificial degradation, 3 a partially convincing degradation, 4 a realistic degradation with only minor synthetic cues, and 5 a degradation indistinguishable from a real example of the same type.

Stimuli and Presentation. The assessment included 50 synthetic videos. The matched comparison contained five OmniSyn and five LVQ (Khan et al., 2020) videos for each of the three shared degradation types, namely smoke, motion blur, and non-uniform illumination. The additional comparison contained five OmniSyn videos for each of haze, splash, dual degradation, and triple degradation. Surgeons were instructed to evaluate degradation formation and temporal behavior rather than image resolution. Dataset identity was not disclosed, and both condition order and clip order were randomized independently for each surgeon.

D Quantitative Analysis

D.1 Training Dataset Comparison

D.1.1 Restoration Results

Tables 5 and 6 report matched-budget {SurgClean-OmniSyn-S} comparisons on the single- and multi-degradation test subsets. Tables 7 and 8 provide the corresponding {LVQ-OmniSyn-L} comparison. Each table reports the Gallbladder and Colon test subset on OmniBench separately. SurgClean omits AverNet and ORCANet because it lacks the required temporal sequences, whereas LVQ covers all 15 benchmarked methods.

Table 5: Model-wise SurgClean-OmniSyn-S comparison on the single-degradation (S) subsets of OmniBench. Entries follow the SurgClean / OmniSyn-S order. Colored boldface marks the higher unrounded value using the corresponding source color in Figure 3.

Model	Venue	Gallbladder (S)			Colon (S)		
		TOPIQ↑	MUSIQ↑	Q-Align↑	TOPIQ↑	MUSIQ↑	Q-Align↑
PromptIR	NeurIPS'23	0.248/**0.317**	40.510/**44.299**	3.325/**3.477**	0.301/**0.395**	44.272/**52.309**	3.460/**3.749**
SFHformer	ECCV'24	0.235/**0.321**	39.269/**45.070**	3.346/**3.482**	0.268/**0.408**	42.050/**54.450**	3.378/**3.785**
MambaIR	ECCV'24	0.278/**0.314**	42.381/**43.621**	3.333/**3.437**	0.321/**0.408**	46.999/**53.199**	3.557/**3.745**
BioIR	NeurIPS'25	0.251/**0.325**	41.222/**45.797**	3.505/**3.520**	0.292/**0.407**	46.658/**54.506**	3.590/**3.799**
MoCE	CVPR'25	0.226/**0.320**	37.686/**44.903**	3.199/**3.474**	0.269/**0.396**	41.097/**53.202**	3.328/**3.763**
AdaIR	ICLR'25	0.247/**0.316**	38.882/**44.660**	3.307/**3.465**	0.307/**0.401**	44.166/**52.939**	3.492/**3.747**
DA-RCOT	TPAMI'25	0.281/**0.318**	43.070/**44.182**	**3.460**/3.425	0.350/**0.409**	48.841/**53.167**	3.682/**3.739**
MambaIRv2	CVPR'25	0.275/**0.312**	42.007/**43.731**	3.369/**3.424**	0.315/**0.413**	46.271/**53.510**	3.533/**3.737**
EndoIR	AAAI'26	0.252/**0.292**	31.755/**39.146**	2.822/**3.133**	0.315/**0.410**	38.123/**47.616**	3.302/**3.607**
RAR	CVPR'26	0.228/**0.232**	**37.075**/37.002	**3.000**/2.997	0.242/**0.263**	36.894/**37.522**	2.877/**2.889**
HOGformer	AAAI'26	0.246/**0.319**	39.515/**45.348**	3.359/**3.507**	0.304/**0.400**	45.389/**53.101**	3.511/**3.771**
BaryIR	TPAMI'26	0.248/**0.317**	40.011/**44.716**	3.330/**3.468**	0.303/**0.400**	44.454/**52.888**	3.490/**3.754**
DATPRLIR	ICLR'26	0.237/**0.318**	39.207/**44.975**	3.274/**3.481**	0.276/**0.398**	42.496/**53.299**	3.340/**3.744**
Mean (13)		0.250/**0.309**	39.430/**43.650**	3.279/**3.407**	0.297/**0.393**	43.670/**51.670**	3.426/**3.679**

Table 6: Model-wise SurgClean-OmniSyn-S comparison on the multi-degradation (M) subsets of OmniBench. Entries follow the SurgClean / OmniSyn-S order. Colored boldface marks the higher unrounded value using the corresponding source color in Figure 3.

Model	Venue	Gallbladder (M)			Colon (M)		
		TOPIQ↑	MUSIQ↑	Q-Align↑	TOPIQ↑	MUSIQ↑	Q-Align↑
PromptIR	NeurIPS'23	0.252/**0.323**	39.589/**45.534**	3.228/**3.383**	0.286/**0.395**	42.446/**51.964**	3.288/**3.747**
SFHformer	ECCV'24	0.244/**0.328**	38.343/**46.274**	3.229/**3.400**	0.244/**0.405**	39.612/**53.953**	3.215/**3.795**
MambaIR	ECCV'24	0.280/**0.321**	42.122/**45.199**	3.254/**3.354**	0.314/**0.410**	46.628/**53.036**	3.507/**3.769**
BioIR	NeurIPS'25	0.255/**0.329**	40.004/**46.609**	3.378/**3.430**	0.271/**0.407**	45.128/**54.128**	3.441/**3.822**
MoCE	CVPR'25	0.229/**0.324**	37.050/**45.691**	3.103/**3.385**	0.251/**0.392**	39.878/**52.509**	3.173/**3.762**
AdaIR	ICLR'25	0.253/**0.321**	38.763/**45.766**	3.222/**3.373**	0.300/**0.402**	43.301/**52.303**	3.326/**3.747**
DA-RCOT	TPAMI'25	0.285/**0.322**	43.240/**45.285**	3.352/**3.362**	0.351/**0.411**	48.282/**52.647**	3.634/**3.745**
MambaIRv2	CVPR'25	0.279/**0.321**	42.214/**45.536**	3.303/**3.338**	0.308/**0.416**	46.046/**53.164**	3.498/**3.751**
EndoIR	AAAI'26	0.257/**0.291**	32.510/**40.699**	2.724/**3.022**	0.321/**0.413**	37.681/**46.327**	3.121/**3.427**
RAR	CVPR'26	0.229/**0.239**	35.836/**36.384**	2.917/**2.943**	0.245/**0.264**	**37.875**/37.550	2.880/**2.880**
HOGformer	AAAI'26	0.252/**0.323**	39.254/**46.210**	3.292/**3.417**	0.295/**0.403**	44.058/**52.895**	3.378/**3.789**
BaryIR	TPAMI'26	0.251/**0.323**	39.439/**45.970**	3.235/**3.386**	0.294/**0.401**	43.201/**52.596**	3.356/**3.772**
DATPRLIR	ICLR'26	0.241/**0.321**	38.507/**45.944**	3.195/**3.394**	0.263/**0.398**	41.354/**52.918**	3.238/**3.742**
Mean (13)		0.254/**0.314**	38.990/**44.700**	3.187/**3.322**	0.288/**0.394**	42.730/**51.230**	3.312/**3.673**

As shown in Tables 5 and 6, OmniSyn-S achieves the higher score in 152 of 156 model-metric comparisons, corresponding to a 97.4% win rate. The advantage remains consistent across degradation complexity, with 75 of 78 comparisons favoring OmniSyn-S on the single-degradation subsets and 77 of 78 on the multi-degradation subsets. On Gallbladder, the mean gains in TOPIQ, MUSIQ, and Q-Align are 0.059, 4.220, and 0.128 for single degradations and 0.060, 5.710, and 0.135 for multiple degradations. The corresponding Colon gains are larger at 0.096, 8.000, and 0.253 under single degradations and 0.106, 8.500, and 0.361 under multiple degradations.

As shown in Tables 7 and 8, OmniSyn-L consistently outperforms LVQ across both single- and multi-degradation tests, achieving higher scores in 157 of 180 method-metric comparisons. Since both datasets provide exact degraded-clean pairs and use matched training scales, this comparison controls for pairing accuracy and data quantity, making degradation modeling and data coverage the main differences between the two training sources. LVQ relies on frame-wise image-space synthesis over a relatively limited set of surgical scenes and discrete severity levels, whereas OmniSyn-L models degradations according to their formation stages while covering broader surgical content and severity ranges. The advantage of OmniSyn-L is particularly evident on the Colon subset for TOPIQ and MUSIQ, while Q-Align remains consistently higher across both anatomical subsets. Although a small number of Gallbladder cases favor LVQ, the overall trend remains unchanged. These results further support that accurate correspondence alone is insufficient for robust real-world transfer, and that formation-aware degradation modeling together with broader content and severity coverage provides more transferable supervision.

For the cross-scale comparison, all aggregate statistics are recomputed over the 13 image restoration methods shared by OmniSyn-S and LVQ. OmniSyn-S achieves the better result in 23 of 28 dimensions, comprising 10 of 12 restoration-quality dimensions and 13 of 16 downstream prediction-consistency dimensions. Because training scale, surgical content, severity coverage, and

Table 7: Model-wise LVQ-OmniSyn-L comparison on the single-degradation (S) subsets of OmniBench. Entries follow the LVQ / OmniSyn-L order. Colored boldface marks the higher unrounded value using the corresponding source color in Figure 3.

Model	Venue	Gallbladder (S)			Colon (S)		
		TOPIQ↑	MUSIQ↑	Q-Align↑	TOPIQ↑	MUSIQ↑	Q-Align↑
PromptIR	NeurIPS'23	0.312/**0.328**	43.700/**45.525**	3.420/**3.489**	0.384/**0.412**	51.638/**54.277**	3.728/**3.802**
SFHformer	ECCV'24	**0.337**/0.333	43.520/**45.927**	3.385/**3.492**	0.385/**0.417**	52.085/**54.853**	3.717/**3.786**
MambaIR	ECCV'24	**0.339**/0.332	**45.291**/45.140	3.436/**3.494**	0.389/**0.411**	52.466/**54.800**	3.715/**3.844**
AverNet	NeurIPS'24	0.290/**0.318**	42.976/**43.591**	**3.490**/3.402	0.367/**0.416**	50.031/**53.502**	3.723/**3.731**
BioIR	NeurIPS'25	0.307/**0.339**	43.801/**46.926**	3.423/**3.517**	0.389/**0.433**	52.266/**56.843**	3.713/**3.883**
MoCE	CVPR'25	0.335/**0.335**	44.210/**46.243**	3.421/**3.505**	0.385/**0.414**	52.197/**54.771**	3.714/**3.816**
AdaIR	ICLR'25	0.309/**0.330**	43.401/**45.889**	3.423/**3.498**	0.382/**0.421**	51.784/**55.040**	3.701/**3.826**
DA-RCOT	TPAMI'25	0.313/**0.330**	43.805/**45.208**	3.425/**3.476**	0.384/**0.417**	51.831/**54.631**	3.716/**3.823**
MambaIRv2	CVPR'25	0.320/**0.330**	44.334/**45.115**	3.437/**3.488**	0.386/**0.407**	51.688/**53.780**	3.719/**3.777**
ORCANet	ArXiv'26	0.297/**0.324**	43.077/**44.996**	3.371/**3.490**	0.375/**0.424**	51.141/**54.693**	3.645/**3.808**
EndoIR	AAAI'26	**0.313**/0.290	**42.300**/40.501	3.080/**3.261**	**0.391**/0.385	**51.889**/48.984	3.630/**3.668**
RAR	CVPR'26	**0.245**/0.242	**38.937**/37.000	**2.984**/2.977	0.244/**0.258**	**40.438**/37.314	2.879/**2.932**
HOGformer	AAAI'26	0.311/**0.324**	43.915/**45.980**	3.425/**3.507**	0.384/**0.409**	51.973/**54.675**	3.721/**3.804**
BaryIR	TPAMI'26	0.305/**0.332**	43.521/**46.106**	3.399/**3.494**	0.382/**0.417**	51.843/**54.687**	3.698/**3.795**
DATPRLIR	ICLR'26	0.305/**0.331**	43.310/**45.903**	3.429/**3.495**	0.382/**0.428**	51.729/**55.899**	3.703/**3.834**
Mean (15)		0.309/**0.321**	43.340/**44.670**	3.370/**3.439**	0.374/**0.405**	51.000/**53.250**	3.648/**3.742**

Table 8: Model-wise LVQ-OmniSyn-L comparison on the multi-degradation (M) subsets of OmniBench. Entries follow the LVQ / OmniSyn-L order. Colored boldface marks the higher unrounded value using the corresponding source color in Figure 3.

Model	Venue	Gallbladder (M)			Colon (M)		
		TOPIQ↑	MUSIQ↑	Q-Align↑	TOPIQ↑	MUSIQ↑	Q-Align↑
PromptIR	NeurIPS'23	0.315/**0.332**	45.040/**47.297**	3.310/**3.417**	0.387/**0.415**	51.391/**54.183**	3.737/**3.785**
SFHformer	ECCV'24	**0.342**/0.334	45.786/**47.403**	3.310/**3.413**	0.388/**0.419**	52.151/**54.538**	3.743/**3.783**
MambaIR	ECCV'24	**0.340**/0.337	46.432/**47.606**	3.331/**3.437**	0.388/**0.412**	51.815/**54.638**	3.739/**3.851**
AverNet	NeurIPS'24	0.295/**0.323**	44.913/**45.915**	**3.433**/3.323	0.362/**0.424**	50.035/**53.797**	3.737/**3.745**
BioIR	NeurIPS'25	0.315/**0.342**	44.452/**48.711**	3.276/**3.450**	0.389/**0.435**	52.040/**56.363**	3.748/**3.871**
MoCE	CVPR'25	**0.339**/0.338	45.958/**47.727**	3.319/**3.425**	0.387/**0.414**	52.065/**54.595**	3.732/**3.800**
AdaIR	ICLR'25	0.314/**0.333**	44.479/**47.654**	3.301/**3.422**	0.383/**0.427**	51.316/**54.685**	3.721/**3.821**
DA-RCOT	TPAMI'25	0.316/**0.333**	44.709/**47.363**	3.319/**3.420**	0.387/**0.420**	51.620/**54.066**	3.729/**3.806**
MambaIRv2	CVPR'25	0.318/**0.333**	45.064/**47.359**	3.335/**3.396**	0.386/**0.409**	51.410/**53.772**	3.716/**3.784**
ORCANet	ArXiv'26	0.300/**0.330**	44.452/**47.053**	3.219/**3.402**	0.373/**0.431**	50.409/**54.567**	3.635/**3.824**
EndoIR	AAAI'26	**0.315**/0.283	**43.181**/41.916	3.076/**3.138**	0.389/**0.399**	**51.660**/49.459	3.527/**3.651**
RAR	CVPR'26	**0.254**/0.244	**37.945**/36.561	**2.933**/2.916	0.245/**0.256**	**42.199**/37.061	2.851/**2.879**
HOGformer	AAAI'26	0.316/**0.327**	44.789/**47.470**	3.317/**3.439**	0.386/**0.413**	51.998/**54.573**	3.751/**3.815**
BaryIR	TPAMI'26	0.314/**0.334**	44.476/**47.851**	3.285/**3.422**	0.385/**0.422**	51.370/**54.623**	3.702/**3.798**
DATPRLIR	ICLR'26	0.316/**0.337**	44.476/**47.964**	3.314/**3.442**	0.383/**0.431**	51.423/**55.429**	3.711/**3.814**
Mean (15)		0.314/**0.324**	44.410/**46.390**	3.272/**3.364**	0.375/**0.408**	50.860/**53.090**	3.652/**3.735**

degradation-type composition vary jointly between the two sources, this result is treated as descriptive evidence rather than an isolated scale ablation.

D.1.2 Downstream Prediction-Consistency Results

Tables 9 and 10 report the downstream prediction-consistency comparison between SurgClean and OmniSyn-S on the single- and multi-degradation test subsets, respectively. Tables 11 and 12 provide the corresponding comparison between LVQ and OmniSyn-L. Evaluation uses the paired OmniSyn test set, where predictions on clean frames provide fixed references for scene segmentation, instrument segmentation, and monocular depth estimation. The tables report Dice and IoU for the two segmentation tasks together with δ_1, AbsRel, SqRel, and RMSE for depth estimation.

As shown in Tables 9 and 10, OmniSyn-S outperforms SurgClean in 181 of 208 model-metric comparisons, while OmniSyn-L outperforms LVQ in 227 of 240 comparisons. More importantly, every model-averaged comparison favors OmniSyn across scene segmentation, instrument segmentation, and monocular depth estimation under both single- and multi-degradation tests. Since segmentation and depth estimation rely on different task-relevant structural cues, this consistent ordering suggests that the advantage of OmniSyn is not limited to restoration quality, but extends to more consistent preservation of semantic and geometric information across downstream tasks.

D.2 Metric Agreement Analysis

Table 13 audits the rankings from the same 15 "Single" checkpoints and OmniSyn multi-degradation results reported in the corresponding composition tables. All nine associations are positive, with

Table 9: Model-wise downstream prediction consistency for SurgClean and OmniSyn-S on the single-degradation subset of the paired OmniSyn test set, denoted by S. Entries follow the SurgClean / OmniSyn-S order. Colored boldface marks the better unrounded value using the corresponding source color in Figure 3.

Model	Venue	Scene Segmentation (S)		Instrument Segmentation (S)		Monocular Depth (S)			
		Dice↑	IoU↑	Dice↑	IoU↑	δ_1↑	AbsRel↓	SqRel↓	RMSE↓
PromptIR	NeurIPS'23	0.915/**0.930**	0.869/**0.896**	0.935/**0.947**	0.889/**0.913**	0.992/**0.994**	0.040/**0.033**	0.008/**0.006**	0.125/**0.101**
SFHformer	ECCV'24	0.914/**0.928**	0.868/**0.895**	0.923/**0.952**	0.872/**0.919**	0.989/**0.994**	0.042/**0.034**	0.009/**0.006**	0.133/**0.103**
MambaIR	ECCV'24	0.920/**0.931**	0.878/**0.895**	0.942/**0.943**	0.901/**0.905**	**0.995**/0.995	0.037/**0.033**	0.006/**0.006**	0.113/**0.101**
BioIR	NeurIPS'25	0.916/**0.930**	0.871/**0.896**	0.927/**0.953**	0.878/**0.921**	0.990/**0.993**	0.039/**0.034**	0.008/**0.007**	0.120/**0.105**
MoCE	CVPR'25	0.912/**0.932**	0.865/**0.899**	0.925/**0.952**	0.876/**0.918**	0.992/**0.996**	0.040/**0.030**	0.008/**0.005**	0.124/**0.092**
AdaIR	ICLR'25	0.916/**0.932**	0.872/**0.897**	0.935/**0.947**	0.890/**0.912**	**0.995**/0.995	0.037/**0.033**	0.007/**0.006**	0.113/**0.100**
DA-RCOT	TPAMI'25	0.922/**0.933**	0.882/**0.897**	0.945/**0.948**	0.906/**0.913**	**0.997**/0.996	0.035/**0.032**	0.006/**0.005**	0.107/**0.097**
MambaIRv2	CVPR'25	0.919/**0.932**	0.877/**0.896**	**0.943**/0.943	0.903/**0.906**	**0.995**/0.995	0.035/**0.033**	**0.006**/0.006	0.108/**0.102**
EndoIR	AAAI'26	0.812/**0.878**	0.737/**0.821**	0.880/**0.923**	0.807/**0.875**	0.864/**0.939**	0.123/**0.077**	0.075/**0.035**	0.331/**0.223**
RAR	CVPR'26	0.905/**0.906**	0.857/**0.857**	**0.926**/0.924	**0.878**/0.875	**0.990**/0.989	0.043/**0.041**	**0.009**/0.009	0.129/**0.125**
HOGformer	AAAI'26	0.917/**0.932**	0.872/**0.898**	0.934/**0.952**	0.889/**0.918**	**0.995**/0.994	0.037/**0.033**	0.006/**0.006**	0.114/**0.098**
BaryIR	TPAMI'26	0.915/**0.930**	0.870/**0.895**	0.937/**0.949**	0.892/**0.914**	**0.995**/0.994	0.039/**0.034**	0.007/**0.006**	0.121/**0.104**
DATPRLIR	ICLR'26	0.917/**0.928**	0.874/**0.894**	0.925/**0.948**	0.877/**0.913**	0.993/**0.996**	0.036/**0.030**	0.007/**0.005**	0.111/**0.092**
Mean (13)		0.908/**0.925**	0.861/**0.888**	0.929/**0.945**	0.881/**0.908**	0.983/**0.990**	0.045/**0.037**	0.012/**0.008**	0.135/**0.111**

Table 10: Model-wise downstream prediction consistency for SurgClean and OmniSyn-S on the multi-degradation subset of the paired OmniSyn test set, denoted by M. Entries follow the SurgClean / OmniSyn-S order. Colored boldface marks the better unrounded value using the corresponding source color in Figure 3.

Model	Venue	Scene Segmentation (M)		Instrument Segmentation (M)		Monocular Depth (M)			
		Dice↑	IoU↑	Dice↑	IoU↑	δ_1↑	AbsRel↓	SqRel↓	RMSE↓
PromptIR	NeurIPS'23	0.890/**0.911**	0.834/**0.864**	0.909/**0.925**	0.849/**0.873**	0.959/**0.967**	0.064/**0.055**	0.019/**0.016**	0.180/**0.154**
SFHformer	ECCV'24	0.877/**0.901**	0.819/**0.854**	0.907/**0.925**	0.845/**0.872**	0.959/**0.961**	0.063/**0.058**	0.019/**0.019**	0.177/**0.158**
MambaIR	ECCV'24	0.899/**0.900**	0.845/**0.847**	**0.916**/0.915	**0.862**/0.858	0.950/**0.952**	0.068/**0.062**	0.023/**0.021**	0.189/**0.170**
BioIR	NeurIPS'25	0.884/**0.914**	0.826/**0.868**	0.909/**0.927**	0.849/**0.877**	0.957/**0.958**	0.062/**0.059**	0.019/**0.019**	0.170/**0.161**
MoCE	CVPR'25	0.890/**0.910**	0.835/**0.863**	0.911/**0.923**	0.851/**0.871**	**0.964**/0.960	0.060/**0.058**	**0.017**/0.018	0.171/**0.157**
AdaIR	ICLR'25	0.897/**0.913**	0.842/**0.865**	0.913/**0.919**	0.854/**0.865**	0.953/**0.961**	0.064/**0.058**	0.020/**0.018**	0.179/**0.162**
DA-RCOT	TPAMI'25	0.904/**0.910**	0.852/**0.861**	0.916/**0.921**	0.859/**0.868**	0.961/**0.964**	0.061/**0.059**	0.018/**0.018**	0.170/**0.165**
MambaIRv2	CVPR'25	0.903/**0.903**	0.851/**0.851**	**0.918**/0.916	**0.862**/0.860	**0.956**/0.953	0.063/**0.062**	**0.020**/0.020	0.176/**0.171**
EndoIR	AAAI'26	0.747/**0.830**	0.670/**0.765**	0.868/**0.893**	0.789/**0.831**	0.866/**0.903**	0.131/**0.101**	0.090/**0.051**	0.348/**0.276**
RAR	CVPR'26	0.893/**0.896**	0.840/**0.842**	**0.922**/0.914	**0.868**/0.857	**0.942**/0.938	0.072/**0.072**	**0.026**/0.027	0.198/**0.196**
HOGformer	AAAI'26	0.892/**0.915**	0.838/**0.868**	0.913/**0.926**	0.854/**0.875**	**0.964**/0.954	**0.060**/0.061	**0.017**/0.021	0.171/**0.168**
BaryIR	TPAMI'26	0.897/**0.912**	0.842/**0.864**	0.910/**0.923**	0.851/**0.871**	0.959/**0.962**	0.062/**0.058**	0.018/**0.018**	0.172/**0.162**
DATPRLIR	ICLR'26	0.895/**0.909**	0.841/**0.861**	0.911/**0.922**	0.851/**0.868**	0.964/**0.967**	0.058/**0.055**	0.017/**0.016**	0.162/**0.154**
Mean (13)		0.882/**0.902**	0.826/**0.852**	0.909/**0.919**	0.850/**0.865**	0.950/**0.954**	0.068/**0.063**	0.025/**0.022**	0.189/**0.173**

Table 11: Model-wise downstream prediction consistency for LVQ and OmniSyn-L on the single-degradation subset of the paired OmniSyn test set, denoted by S. Entries follow the LVQ / OmniSyn-L order. Colored boldface marks the better unrounded value using the corresponding source color in Figure 3.

Model	Venue	Scene Segmentation (S)		Instrument Segmentation (S)		Monocular Depth (S)			
		Dice↑	IoU↑	Dice↑	IoU↑	δ_1↑	AbsRel↓	SqRel↓	RMSE↓
PromptIR	NeurIPS'23	0.930/**0.939**	0.894/**0.907**	0.944/**0.952**	0.907/**0.919**	0.990/**0.997**	0.043/**0.032**	0.010/**0.006**	0.127/**0.098**
SFHformer	ECCV'24	0.928/**0.935**	0.892/**0.903**	0.945/**0.952**	0.909/**0.919**	**0.990**/0.984	0.042/**0.037**	0.010/**0.010**	0.126/**0.114**
MambaIR	ECCV'24	0.925/**0.935**	0.886/**0.901**	0.948/**0.951**	0.912/**0.917**	0.993/**0.996**	0.041/**0.035**	0.009/**0.006**	0.122/**0.106**
AverNet	NeurIPS'24	0.911/**0.921**	0.866/**0.880**	0.933/**0.948**	0.889/**0.911**	0.932/**0.989**	0.074/**0.039**	0.029/**0.009**	0.210/**0.119**
BioIR	NeurIPS'25	0.930/**0.935**	0.894/**0.902**	0.944/**0.951**	0.906/**0.917**	0.985/**0.996**	0.046/**0.031**	0.011/**0.006**	0.134/**0.095**
MoCE	CVPR'25	0.932/**0.936**	0.896/**0.904**	0.947/**0.954**	0.911/**0.922**	**0.997**/0.997	0.036/**0.031**	0.007/**0.006**	0.111/**0.095**
AdaIR	ICLR'25	0.931/**0.936**	0.895/**0.903**	0.946/**0.950**	0.909/**0.916**	0.987/**0.996**	0.044/**0.031**	0.011/**0.006**	0.130/**0.096**
DA-RCOT	TPAMI'25	0.932/**0.936**	0.896/**0.903**	0.946/**0.954**	0.910/**0.921**	0.992/**0.996**	0.040/**0.031**	0.010/**0.005**	0.120/**0.095**
MambaIRv2	CVPR'25	0.931/**0.940**	0.895/**0.908**	0.946/**0.954**	0.909/**0.921**	0.985/**0.995**	0.047/**0.035**	0.012/**0.007**	0.140/**0.106**
ORCANet	ArXiv'26	0.908/**0.929**	0.862/**0.891**	0.934/**0.952**	0.892/**0.917**	0.975/**0.993**	0.055/**0.034**	0.016/**0.007**	0.164/**0.104**
EndoIR	AAAI'26	0.901/**0.902**	0.852/**0.854**	0.928/**0.940**	0.883/**0.900**	**0.971**/0.967	**0.056**/0.058	**0.016**/0.018	**0.162**/0.169
RAR	CVPR'26	0.901/**0.912**	0.849/**0.866**	**0.927**/0.926	**0.878**/0.877	0.941/**0.992**	0.067/**0.041**	0.028/**0.009**	0.199/**0.126**
HOGformer	AAAI'26	0.934/**0.938**	0.900/**0.905**	0.950/**0.953**	0.915/**0.919**	0.990/**0.997**	0.040/**0.030**	0.009/**0.005**	0.119/**0.091**
BaryIR	TPAMI'26	0.936/**0.938**	0.901/**0.906**	0.950/**0.954**	0.914/**0.921**	0.995/**0.997**	0.038/**0.030**	0.008/**0.005**	0.116/**0.092**
DATPRLIR	ICLR'26	**0.936**/0.935	0.901/**0.903**	0.948/**0.953**	0.914/**0.920**	0.993/**0.997**	0.041/**0.029**	0.009/**0.005**	0.123/**0.088**
Mean (15)		0.924/**0.931**	0.885/**0.896**	0.942/**0.950**	0.904/**0.914**	0.981/**0.993**	0.047/**0.035**	0.013/**0.007**	0.140/**0.106**

Spearman's ρ ranging from 0.500 to 0.957 and Kendall's τ_b from 0.486 to 0.867. MUSIQ and Q-Align show strong agreement across PSNR, SSIM, and direction-reversed LPIPS, while TOPIQ shows moderate positive agreement. This result supports the joint use of the three no-reference metrics on OmniBench without treating any single score as a substitute for reference-based fidelity.

Table 12: Model-wise downstream prediction consistency for LVQ and OmniSyn-L on the multi-degradation subset of the paired OmniSyn test set, denoted by M. Entries follow the LVQ / OmniSyn-L order. Colored boldface marks the better unrounded value using the corresponding source color in Figure 3.

Model	Venue	Scene Segmentation (M)		Instrument Segmentation (M)		Monocular Depth (M)			
		Dice↑	IoU↑	Dice↑	IoU↑	δ_1↑	AbsRel↓	SqRel↓	RMSE↓
PromptIR	NeurIPS'23	0.899/**0.910**	0.845/**0.861**	0.922/**0.926**	0.869/**0.877**	0.943/**0.966**	0.068/**0.057**	0.026/**0.017**	0.184/**0.161**
SFHformer	ECCV'24	0.890/**0.907**	0.837/**0.859**	0.916/**0.927**	0.863/**0.876**	0.950/**0.957**	0.064/**0.059**	0.021/**0.019**	0.175/**0.163**
MambaIR	ECCV'24	0.897/**0.905**	0.844/**0.856**	0.925/**0.929**	0.874/**0.879**	0.950/**0.954**	0.064/**0.061**	0.022/**0.020**	0.175/**0.167**
AverNet	NeurIPS'24	0.856/**0.900**	0.792/**0.849**	0.905/**0.926**	0.845/**0.875**	0.921/**0.955**	0.089/**0.062**	0.041/**0.020**	0.241/**0.170**
BioIR	NeurIPS'25	0.881/**0.909**	0.823/**0.861**	0.918/**0.929**	0.863/**0.878**	0.939/**0.958**	0.073/**0.059**	0.030/**0.019**	0.195/**0.162**
MoCE	CVPR'25	0.905/**0.911**	0.854/**0.863**	0.925/**0.930**	0.874/**0.881**	0.954/**0.956**	0.062/**0.060**	0.021/**0.019**	0.171/**0.165**
AdaIR	ICLR'25	0.894/**0.910**	0.841/**0.861**	0.918/**0.925**	0.864/**0.875**	0.944/**0.958**	0.068/**0.059**	0.026/**0.019**	0.184/**0.165**
DA-RCOT	TPAMI'25	0.887/**0.904**	0.833/**0.854**	0.918/**0.928**	0.865/**0.879**	0.948/**0.958**	0.065/**0.061**	0.023/**0.019**	0.178/**0.169**
MambaIRv2	CVPR'25	0.902/**0.905**	0.851/**0.854**	0.919/**0.929**	0.868/**0.879**	0.937/**0.949**	0.072/**0.064**	0.029/**0.022**	0.192/**0.174**
ORCANet	ArXiv'26	0.873/**0.904**	0.814/**0.854**	0.911/**0.928**	0.854/**0.878**	0.932/**0.956**	0.078/**0.061**	0.031/**0.020**	0.212/**0.167**
EndoIR	AAAI'26	**0.866**/0.854	**0.808**/0.793	0.901/**0.925**	0.841/**0.872**	0.916/**0.923**	0.090/**0.086**	0.040/**0.039**	0.242/**0.236**
RAR	CVPR'26	0.888/**0.905**	0.831/**0.854**	0.916/**0.922**	0.864/**0.867**	0.934/**0.950**	0.076/**0.067**	0.032/**0.023**	0.206/**0.183**
HOGformer	AAAI'26	0.902/**0.902**	0.849/**0.852**	**0.925**/0.924	**0.873**/0.873	0.941/**0.955**	0.068/**0.060**	0.027/**0.019**	0.183/**0.166**
BaryIR	TPAMI'26	0.904/**0.911**	0.852/**0.862**	0.921/**0.926**	0.867/**0.876**	0.945/**0.958**	0.065/**0.061**	0.025/**0.020**	0.175/**0.169**
DATPRLIR	ICLR'26	0.895/**0.904**	0.842/**0.855**	0.922/**0.925**	0.870/**0.873**	0.948/**0.956**	0.065/**0.060**	0.023/**0.019**	0.177/**0.164**
Mean (15)		0.889/**0.903**	0.834/**0.852**	0.917/**0.927**	0.864/**0.876**	0.940/**0.954**	0.071/**0.062**	0.028/**0.021**	0.193/**0.172**

Table 13: Rank agreement between no-reference and full-reference quality metrics on the paired OmniSyn multi-degradation test subset. The analysis uses the same 15 "Single" checkpoints and restored sequences reported in Tables 18 and 19. Frame-level scores are averaged within each clip and then equally across the 22 clips. Correlations are computed from the unrounded method-level means, with LPIPS direction-reversed so that higher ranks consistently denote better quality.

NR Metric	Reference Metric	Spearman's ρ	Kendall's τ_b
TOPIQ	PSNR	0.500	0.486
	SSIM	0.575	0.619
	LPIPS	0.668	0.562
MUSIQ	PSNR	0.900	0.771
	SSIM	0.957	0.867
	LPIPS	0.914	0.771
Q-Align	PSNR	0.825	0.695
	SSIM	0.896	0.752
	LPIPS	0.879	0.733

D.3 Surgeon Assessment Results

As shown in Figure 5 and Table 14, OmniSyn is consistently rated more realistic than LVQ across the three shared degradation types. For every surgeon and shared degradation type, the five-clip mean of OmniSyn exceeds that of LVQ, yielding 18 of 18 favorable surgeon-type comparisons. Averaged equally across the three shared types, OmniSyn achieves a matched-type average of 4.26 compared with 3.37 for LVQ. In addition, 92.2% of OmniSyn ratings reach at least 4, compared with 40.0% for LVQ. These results show that the realism advantage of OmniSyn is consistent across both surgeons and shared degradation types. Across the additional single- and multi-degradation settings, OmniSyn receives mean ratings from 3.87 to 4.37. Its average across all seven evaluated settings is 4.14, with 88.1% of ratings reaching at least 4. The consistently high ratings show that the synthetic realism of OmniSyn extends beyond the degradation types shared with LVQ to the broader degradation settings covered by the dataset.

D.4 Training Composition Analysis

The results in this section use the three fixed, equal-scale training compositions defined in Appendix C. Each composition contains 160 clips drawn only from the official training partition, and the same sample lists are used for every restoration method. Results on OmniSyn are measured on its held-out official test partition, which remains disjoint from all three training compositions.

D.4.1 OmniBench Results

All subset and aggregate values are computed from unrounded clip-level scores and rounded only for display. Aggregates recomputed from the printed three-decimal values may therefore differ by 0.001.

Tables 15 and 16 decompose the aggregate OmniBench results in Table 3 into single- and multi-degradation test subsets. Across both real-world test subsets, "Single" remains the strongest overall training composition, achieving the best result in 54 of 60 model-metric comparisons on each sub-

Table 14: Surgeon-level ratings of synthetic degradation realism. S1-S6 report the mean rating over five clips, with the within-surgeon standard error in parentheses. The final column reports the mean $\pm$ standard deviation across the six surgeons. Scores range from 1 to 5, with higher values indicating greater realism.

Source	Setting	S1	S2	S3	S4	S5	S6	Mean $\pm$ SD
OmniSyn	Smoke	4.60 (0.24)	4.60 (0.24)	4.40 (0.24)	4.20 (0.20)	4.00 (0.32)	4.40 (0.24)	4.37 $\pm$ 0.23
LVQ	Smoke	3.40 (0.24)	3.60 (0.24)	3.20 (0.20)	3.20 (0.20)	3.00 (0.32)	3.40 (0.24)	3.30 $\pm$ 0.21
OmniSyn	Motion Blur	4.20 (0.20)	4.80 (0.20)	3.80 (0.20)	4.60 (0.24)	3.60 (0.24)	4.00 (0.32)	4.17 $\pm$ 0.46
LVQ	Motion Blur	3.80 (0.20)	3.60 (0.24)	3.60 (0.24)	3.80 (0.20)	3.40 (0.24)	3.20 (0.20)	3.57 $\pm$ 0.23
OmniSyn	NUI	4.40 (0.24)	4.40 (0.24)	4.20 (0.20)	4.40 (0.24)	3.80 (0.20)	4.20 (0.37)	4.23 $\pm$ 0.23
LVQ	NUI	3.20 (0.20)	3.80 (0.20)	3.00 (0.32)	3.40 (0.24)	2.80 (0.20)	3.20 (0.20)	3.23 $\pm$ 0.34
OmniSyn	Haze	4.20 (0.20)	4.40 (0.24)	3.80 (0.20)	4.40 (0.24)	3.60 (0.24)	4.00 (0.32)	4.07 $\pm$ 0.33
OmniSyn	Splash	4.60 (0.24)	4.80 (0.20)	4.20 (0.20)	4.20 (0.20)	4.00 (0.32)	4.40 (0.24)	4.37 $\pm$ 0.29
OmniSyn	Dual	4.00 (0.32)	4.20 (0.20)	4.00 (0.32)	3.80 (0.20)	3.40 (0.24)	4.20 (0.20)	3.93 $\pm$ 0.30
OmniSyn	Triple	3.80 (0.20)	4.40 (0.24)	3.60 (0.24)	4.00 (0.32)	3.60 (0.24)	3.80 (0.20)	3.87 $\pm$ 0.30

Table 15: Training-composition results on the OmniBench single-degradation (S) test subset. Colored boldface and colored underlining mark the best and second-best configurations, respectively, using the split colors of Figure 7.

Model	Venue	Single (S)				Multi (S)				Mixed (S)			
		TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓	TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓	TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓
PromptIR	NeurIPS'23	**0.364**	**49.870**	**3.614**	**0.201**	0.347	47.297	3.509	0.632	0.364	49.375	3.598	0.325
SFHformer	ECCV'24	**0.367**	**50.354**	**3.658**	**0.202**	0.339	46.978	3.455	0.605	0.360	49.351	3.595	0.333
MambaIR	ECCV'24	**0.362**	**49.509**	**3.605**	**0.244**	0.347	46.743	3.468	0.732	0.354	48.818	3.602	0.336
AverNet	NeurIPS'24	**0.361**	**48.012**	**3.564**	**0.288**	0.348	46.280	3.429	0.490	0.358	46.502	3.526	0.571
BioIR	NeurIPS'25	**0.370**	**51.170**	3.672	**0.225**	0.338	46.995	3.384	0.692	0.363	50.576	**3.681**	0.281
MoCE	CVPR'25	**0.369**	**49.888**	**3.606**	**0.262**	0.329	45.245	3.352	0.783	0.364	49.347	3.573	0.286
AdaIR	ICLR'25	**0.364**	**50.209**	**3.642**	**0.217**	0.343	46.485	3.444	0.742	0.353	49.267	3.610	0.367
DA-RCOT	TPAMI'25	**0.363**	**48.899**	**3.621**	**0.176**	0.360	47.958	3.611	0.501	0.359	48.260	3.594	0.274
MambaIRv2	CVPR'25	**0.356**	**48.505**	**3.604**	**0.240**	0.347	45.977	3.521	0.608	0.351	47.995	3.576	0.319
ORCANet	ArXiv'26	**0.365**	**48.831**	**3.612**	**0.302**	0.357	45.218	3.479	0.879	0.360	48.128	3.611	0.314
EndoIR	AAAI'26	**0.344**	**46.830**	**3.551**	**0.764**	0.314	45.000	3.359	1.234	0.330	46.311	3.461	0.984
RAR	CVPR'26	0.222	34.498	2.882	1.632	**0.252**	**38.772**	**3.009**	**1.275**	0.229	34.700	2.910	1.542
HOGformer	AAAI'26	**0.374**	**50.567**	**3.646**	**0.249**	0.348	46.657	3.344	0.766	0.368	49.991	3.625	0.285
BaryIR	TPAMI'26	**0.369**	**49.610**	**3.623**	**0.198**	0.356	47.577	3.539	0.606	0.365	49.518	3.619	0.265
DATPRLIR	ICLR'26	0.372	**50.665**	**3.643**	**0.213**	0.336	46.888	3.372	0.687	**0.372**	50.497	3.609	0.359

set. The same trend holds across TOPIQ, MUSIQ, and Q-Align for the large majority of methods, indicating that the result is not driven by a single perceptual quality metric. Notably, "Single" remains superior when multiple degradations co-occur in real surgical videos, even though "Multi" is trained directly on synthetic multi-degradation samples. Combined with the paired OmniSyn results below, where multi-degradation training performs better on synthetic multi-degradation data, this opposite trend reveals a clear gap between synthetic multi-degradation supervision and real-world generalization. Restoration capabilities learned from isolated degradations therefore transfer more reliably to authentic degradation combinations under the current synthesis setting. "Mixed" generally remains between "Single" and "Multi" and does not surpass "Single" under the same training budget. RAR is the consistent exception, with "Multi" performing best across all four metrics on both subsets. This behavior is consistent with its iterative Restore-Assess-Repeat mechanism, which repeatedly estimates and corrects residual degradations and may therefore benefit more from direct exposure to co-occurring degradations.

D.4.2 OmniSyn Results

Tables 17 and 18 evaluate the three training compositions on the paired single- and multi-degradation subsets of OmniSyn. The clean references enable PSNR, SSIM, LPIPS, and clean-reference tOF to directly evaluate reconstruction fidelity and motion preservation. On OmniSyn, the strongest training composition follows the degradation structure of the test subset. "Single" achieves the best result in 48 of 60 model-metric comparisons on the synthetic single-degradation subset, while "Multi" achieves the best result in 47 of 60 comparisons on the synthetic multi-degradation subset. However, this advantage of "Multi" on synthetic multi-degradation data does not transfer to real surgical videos. On the corresponding multi-degradation subset of OmniBench, "Single" instead achieves the best result in 54 of 60 comparisons. This reversal between synthetic and real multi-degradation testing highlights the transfer gap between current synthetic multi-degradation supervision and authentic degradation combinations. "Mixed" generally remains between the two specialized settings under the same fixed training budget.

Table 16: Training-composition results on the OmniBench multi-degradation (M) test subset. Colored boldface and colored underlining mark the best and second-best configurations, respectively, using the split colors of Figure 7.

Model	Venue	Single (M)				Multi (M)				Mixed (M)			
		TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓	TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓	TOPIQ↑	MUSIQ↑	Q-Align↑	tOF↓
PromptIR	NeurIPS'23	**0.353**	**48.717**	3.506	**0.185**	0.334	46.336	3.446	0.641	0.352	47.993	**3.512**	0.350
SFHformer	ECCV'24	**0.355**	**48.993**	**3.549**	**0.198**	0.318	44.333	3.295	0.656	0.343	47.430	3.477	0.328
MambaIR	ECCV'24	**0.350**	**48.720**	**3.494**	**0.201**	0.331	44.849	3.358	0.780	0.344	48.041	3.493	0.302
AverNet	NeurIPS'24	**0.346**	**47.485**	**3.467**	**0.206**	0.334	45.540	3.282	0.488	0.343	46.114	3.426	0.554
BioIR	NeurIPS'25	**0.356**	**49.495**	**3.564**	**0.211**	0.331	47.325	3.479	0.587	0.346	48.527	3.546	0.313
MoCE	CVPR'25	**0.355**	**48.724**	**3.515**	**0.200**	0.313	43.117	3.197	0.839	0.348	47.891	3.482	0.293
AdaIR	ICLR'25	**0.352**	**48.684**	**3.530**	**0.183**	0.325	43.573	3.245	0.847	0.339	47.502	3.491	0.409
DA-RCOT	TPAMI'25	**0.351**	**48.011**	**3.515**	**0.144**	0.348	46.498	3.439	0.487	0.348	47.309	3.480	0.225
MambaIRv2	CVPR'25	**0.346**	**47.914**	**3.499**	**0.197**	0.329	44.874	3.387	0.573	0.341	47.453	3.476	0.275
ORCANet	ArXiv'26	**0.353**	**47.746**	**3.504**	**0.193**	0.338	44.134	3.363	0.734	0.346	46.912	3.489	0.318
EndoIR	AAAI'26	**0.333**	**46.388**	**3.427**	0.640	0.314	45.595	3.248	0.718	0.328	46.041	3.356	**0.621**
RAR	CVPR'26	0.221	34.108	2.859	1.181	**0.256**	**38.509**	**3.003**	**0.929**	0.227	34.072	2.875	1.125
HOGformer	AAAI'26	**0.359**	**49.413**	**3.561**	**0.206**	0.328	43.387	3.177	0.877	0.352	48.508	3.532	0.278
BaryIR	TPAMI'26	**0.357**	**48.676**	**3.530**	**0.158**	0.342	46.068	3.401	0.632	0.351	48.211	3.512	0.245
DATPRLIR	ICLR'26	**0.359**	**49.424**	**3.529**	**0.175**	0.315	43.595	3.154	0.780	0.358	49.169	3.516	0.324

Table 17: Training-composition results on the OmniSyn single-degradation (S) test subset. tOF uses the clean motion reference. Colored boldface and colored underlining mark the best and second-best configurations, respectively, using the split colors of Figure 7.

Model	Venue	Single (S)				Multi (S)				Mixed (S)			
		PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓
PromptIR	NeurIPS'23	**30.489**	**0.923**	**0.123**	**0.861**	25.909	0.878	0.189	1.223	29.903	0.908	0.132	0.939
SFHformer	ECCV'24	**31.654**	**0.925**	**0.123**	**0.794**	27.877	0.886	0.193	1.156	30.907	0.917	0.129	0.818
MambaIR	ECCV'24	**30.484**	**0.917**	**0.116**	**0.832**	25.306	0.854	0.228	1.485	29.456	0.899	0.145	1.052
AverNet	NeurIPS'24	**28.308**	**0.902**	**0.138**	**0.928**	24.810	0.872	0.173	1.095	25.985	0.852	0.198	1.468
BioIR	NeurIPS'25	**31.515**	**0.928**	**0.119**	**0.790**	26.862	0.883	0.199	1.172	29.954	0.917	0.132	0.846
MoCE	CVPR'25	**31.842**	**0.929**	**0.114**	**0.768**	26.862	0.880	0.215	1.324	31.196	0.922	0.120	0.820
AdaIR	ICLR'25	**31.316**	**0.924**	**0.123**	**0.852**	26.028	0.869	0.219	1.347	30.433	0.915	0.133	0.901
DA-RCOT	TPAMI'25	28.899	0.902	0.139	0.996	26.689	0.880	0.157	1.136	**29.227**	**0.905**	**0.124**	**0.906**
MambaIRv2	CVPR'25	**29.528**	**0.907**	**0.133**	**0.964**	25.956	0.867	0.163	1.139	28.814	0.904	0.133	0.965
ORCANet	ArXiv'26	**30.573**	**0.914**	**0.110**	**0.788**	25.454	0.859	0.214	1.261	29.779	0.902	0.123	0.844
EndoIR	AAAI'26	**22.940**	**0.773**	**0.190**	**1.090**	20.740	0.719	0.221	1.244	21.722	0.752	0.198	1.125
RAR	CVPR'26	24.257	0.807	0.338	1.711	23.832	0.806	**0.258**	**1.235**	**25.378**	**0.826**	0.314	1.471
HOGformer	AAAI'26	**31.801**	**0.930**	**0.110**	**0.787**	25.525	0.861	0.238	1.469	30.703	0.915	0.123	0.863
BaryIR	TPAMI'26	**29.900**	**0.908**	0.136	0.978	26.657	0.887	0.170	1.172	29.758	0.907	**0.131**	**0.930**
DATPRLIR	ICLR'26	**31.158**	**0.916**	0.152	1.098	27.545	0.880	0.221	1.294	30.538	0.909	**0.142**	**1.053**

Table 18: Training-composition results on the OmniSyn multi-degradation (M) test subset. tOF uses the clean motion reference. Colored boldface and colored underlining mark the best and second-best configurations, respectively, using the split colors of Figure 7.

Model	Venue	Single (M)				Multi (M)				Mixed (M)			
		PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓	PSNR↑	SSIM↑	LPIPS↓	tOF↓
PromptIR	NeurIPS'23	21.241	0.782	0.351	2.297	**24.869**	**0.829**	**0.280**	**1.966**	23.257	0.799	0.310	2.178
SFHformer	ECCV'24	23.085	0.813	0.289	1.933	**25.762**	**0.840**	0.275	1.923	25.359	0.837	**0.275**	**1.894**
MambaIR	ECCV'24	21.463	0.775	0.341	2.297	**24.039**	**0.814**	0.311	2.095	23.439	0.805	**0.301**	**2.076**
AverNet	NeurIPS'24	21.523	0.772	0.328	2.188	**24.576**	**0.815**	**0.270**	**1.946**	22.701	0.776	0.322	2.291
BioIR	NeurIPS'25	22.720	0.806	0.311	2.064	**25.638**	**0.840**	**0.274**	**1.916**	24.707	0.833	0.283	1.941
MoCE	CVPR'25	22.482	0.803	0.318	2.080	**25.508**	**0.839**	0.281	1.957	24.969	0.833	**0.279**	**1.933**
AdaIR	ICLR'25	21.918	0.791	0.329	2.215	**24.716**	**0.823**	**0.294**	**2.077**	23.893	0.813	0.304	2.121
DA-RCOT	TPAMI'25	20.714	0.759	0.371	2.435	**23.157**	**0.804**	0.306	2.126	22.991	0.800	**0.295**	**2.100**
MambaIRv2	CVPR'25	21.612	0.777	0.341	2.321	**23.727**	**0.807**	0.292	2.029	23.274	0.805	**0.283**	**1.968**
ORCANet	ArXiv'26	22.174	0.788	0.306	2.101	**26.099**	**0.835**	**0.256**	**1.832**	24.760	0.822	0.264	1.878
EndoIR	AAAI'26	20.054	0.690	0.362	2.236	**20.703**	**0.699**	**0.340**	**2.074**	20.349	0.697	0.341	2.171
RAR	CVPR'26	21.270	0.730	0.476	2.913	**23.049**	0.744	**0.331**	**1.959**	22.817	**0.745**	0.409	2.280
HOGformer	AAAI'26	21.772	0.790	0.336	2.242	**25.172**	**0.831**	0.291	2.103	24.707	0.828	**0.278**	**1.965**
BaryIR	TPAMI'26	20.905	0.765	0.367	2.417	**24.718**	**0.826**	**0.279**	**1.981**	23.238	0.811	0.294	2.087
DATPRLIR	ICLR'26	20.822	0.756	0.403	2.548	**25.031**	**0.832**	**0.299**	**2.128**	23.296	0.802	0.340	2.401

To separate the synthetic-to-real reversal from the choice of evaluation metric, we also evaluate the "Single" and "Multi" checkpoints on the 22 held-out OmniSyn multi-degradation clips using the same three no-reference metrics adopted for OmniBench. Table 19 shows that "Multi" performs better in all 45 method-metric comparisons. Across the 15 methods, the mean TOPIQ, MUSIQ, and

Table 19: Same-metric comparison of the "Single" and "Multi" training compositions on the OmniSyn multi-degradation (M) test subset. Colored boldface marks the higher unrounded value using the "Multi" color from Figure 7.

Model	TOPIQ↑		MUSIQ↑		Q-Align↑		Model	TOPIQ↑		MUSIQ↑		Q-Align↑	
	Single	Multi	Single	Multi	Single	Multi		Single	Multi	Single	Multi	Single	Multi
PromptIR	0.245	**0.268**	33.891	**38.703**	2.663	**3.080**	MambaIRv2	0.241	**0.262**	32.345	**34.365**	2.533	**2.675**
SFHformer	0.271	**0.271**	38.511	**38.988**	2.953	**3.038**	ORCANet	0.248	**0.277**	35.157	**39.023**	2.884	**3.096**
MambaIR	0.248	**0.261**	33.690	**35.881**	2.625	**2.832**	EndoIR	0.285	**0.330**	31.467	**36.135**	2.516	**2.773**
AverNet	0.244	**0.269**	33.905	**37.509**	2.742	**3.021**	RAR	0.173	**0.240**	24.424	**36.139**	2.205	**2.902**
BioIR	0.262	**0.273**	37.944	**40.738**	2.952	**3.202**	HOGformer	0.253	**0.271**	36.199	**39.476**	2.815	**3.113**
MoCE	0.261	**0.273**	36.364	**40.184**	2.786	**3.120**	BaryIR	0.243	**0.270**	32.319	**38.405**	2.598	**3.034**
AdaIR	0.256	**0.266**	36.359	**38.255**	2.819	**3.027**	DATPRLIR	0.242	**0.266**	31.975	**39.284**	2.540	**3.104**
DA-RCOT	0.236	**0.256**	31.266	**34.544**	2.550	**2.812**	Mean	0.247	**0.270**	33.721	**37.842**	2.679	**2.989**

Q-Align scores increase from 0.247 to 0.270, 33.72 to 37.84, and 2.679 to 2.989, respectively. The advantage of multi-degradation training on the matched synthetic test distribution therefore persists when the metric family is held fixed. Its reversal on OmniBench is consequently not explained by the use of full-reference metrics on OmniSyn and no-reference metrics on real videos.

Table 20: Summary of comparison counts across the main quantitative analyses. Training-dataset comparisons aggregate 28 evaluation dimensions, with the cross-scale comparison restricted to the 13 shared image restoration methods. Training-composition comparisons cover the synthetic and real single- and multi-degradation subsets.

Analysis	Setting	Outcome	Count	Rate (%)
Training-dataset comparison	Matched scale	OmniSyn-S over SurgClean	28/28	100.0
	Matched scale	OmniSyn-L over LVQ	28/28	100.0
	Cross-scale	OmniSyn-S over LVQ	23/28	82.1
Training-composition comparison	OmniBench (S)	Single	54/60	90.0
	OmniBench (M)	Single	54/60	90.0
	OmniSyn (S)	Single	48/60	80.0
	OmniSyn (M)	Multi	47/60	78.3

Table 20 summarizes how consistently the main quantitative patterns recur across the reported comparison dimensions. In both matched-scale training-dataset comparisons, OmniSyn achieves the better model-averaged result across all 28 aggregate evaluation dimensions. Under the common 13-method cross-scale comparison, OmniSyn-S achieves the better result in 23 of 28 dimensions despite using substantially fewer training pairs than LVQ. The training-composition results also capture the synthetic-to-real reversal. "Multi" achieves the best result in 47 of 60 model-metric comparisons on the synthetic multi-degradation subset, whereas "Single" achieves the best result in 54 of 60 comparisons on the corresponding real multi-degradation subset. These counts are descriptive summaries across correlated model-metric dimensions and are not treated as independent statistical tests.

E Qualitative Analysis

E.1 Training Dataset Comparisons

Figure 12 compares the four training sources on OmniBench. Each panel shows the degraded input and the outputs obtained from models trained on SurgClean, LVQ, OmniSyn-S, and OmniSyn-L using the same restoration architecture and image region. The examples cover both single- and multi-degradation cases and qualitatively follow the overall source ordering in Figure 3, with OmniSyn-trained models showing stronger recovery of obscured tissue and local structures.

Figure 13 provides the corresponding comparison on paired OmniSyn examples. Each panel shows the degraded input, outputs from the four training-source checkpoints, and the clean target for the same tissue region. Across the selected single- and multi-degradation cases, all four checkpoints improve visibility over the degraded input, while OmniSyn-L generally recovers the clean appearance most closely, followed by OmniSyn-S, LVQ, and SurgClean. The use of different restoration methods across panels shows that this trend is not specific to a single architecture.

Figure 14 compares downstream segmentation predictions under a fixed evaluation protocol. The same spatial prompts are used for the degraded input, restored outputs, and paired clean frame. Restoration reduces local prediction errors caused by degradation, with the examples following the aggregate segmentation trend in Figure 3. OmniSyn-L shows the strongest consistency with the clean-frame prediction, followed by OmniSyn-S, LVQ, and SurgClean.

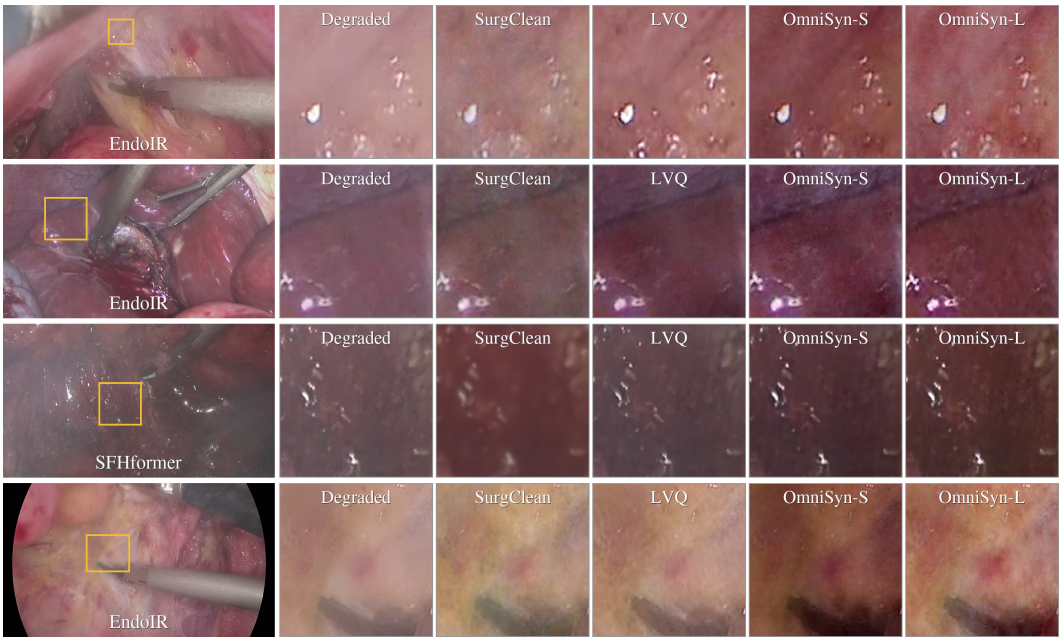


Figure 12: Qualitative comparison of training sources on OmniBench. Enlarged regions compare the degraded input with restoration results from models trained on SurgClean, LVQ, OmniSyn-S, and OmniSyn-L.

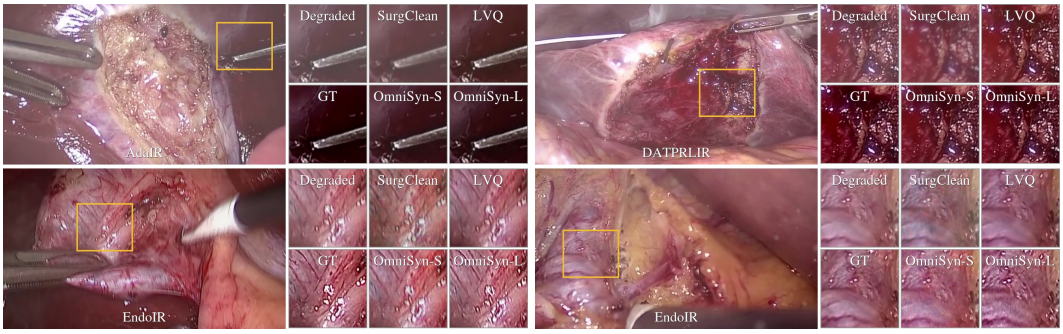


Figure 13: Qualitative comparison of training sources on OmniSyn. Enlarged regions compare the degraded input and GT with restoration results from models trained on SurgClean, LVQ, OmniSyn-S, and OmniSyn-L.

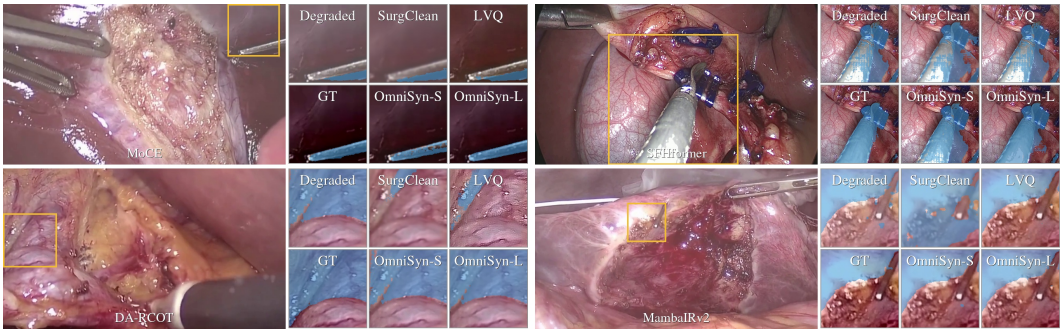


Figure 14: Qualitative comparison of training sources for scene segmentation. The four panels correspond to MoCE, SFHformer, DA-RCOT, and MambaIRv2, respectively, and compare predictions from the degraded input, models trained on SurgClean, LVQ, OmniSyn-S, and OmniSyn-L, and the GT.

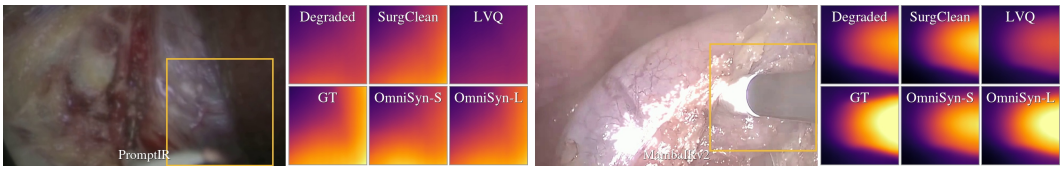


Figure 15: Qualitative comparison of training sources for monocular depth estimation using PromptIR and MambaIRv2. Each panel compares depth predictions from the degraded input, four training-source checkpoints, and the GT.

Figure 15 provides the corresponding comparison for monocular depth estimation. The restored predictions are median-aligned to the clean-frame prediction to remove global scale differences. OmniSyn-L and OmniSyn-S remain closest to the clean reference, followed by SurgClean and LVQ. This ordering is consistent with the task-specific pattern reported in the main paper, where LVQ is relatively stronger in segmentation while SurgClean is relatively stronger in depth. GT in both figures denotes the output of the same frozen predictor on the paired clean frame.

E.2 Benchmark Restoration Comparisons

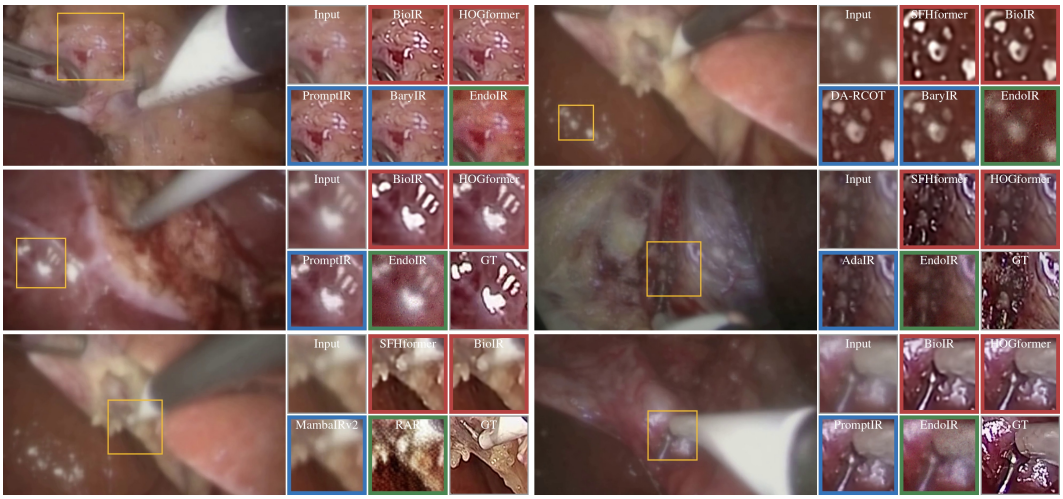


Figure 16: Additional qualitative comparison of representative restoration methods across diverse degradations. From top to bottom, the first row shows OmniBench examples, while the second and third rows show paired OmniSyn examples. Green, blue, and red boxes indicate low, medium, and high restoration-performance tiers, respectively.

F Scope, Ethics, and Reproducibility

F.1 Benchmark Scope and Future Directions

Since the same evolving surgical scene cannot be captured simultaneously as a real degraded observation and its perfectly aligned clean counterpart, SurgOmni combines a paired synthetic video training dataset with a real-world video testing dataset containing naturally occurring surgical degradations, allowing restoration quality, temporal behavior, and downstream prediction consistency to be evaluated under the reference conditions available in each setting. Future extensions can broaden the benchmark to additional institutions, imaging systems, surgical procedures, and longer temporal contexts, and evaluate a wider range of downstream surgical perception tasks.

F.2 Ethics and Data Governance

SurgOmni is intended as a research benchmark, and the reported restoration and prediction-consistency results should not be interpreted as evidence of clinical benefit. Public surgical videos are used in accordance with their respective dataset terms, while the clinical colorectal cohort was collected under institutional ethics approval with informed consent and was de-identified before use. The OmniSyn and OmniBench are distributed under controlled access, with research applicants providing their institutional affiliation and intended use and agreeing to the applicable data-use terms. Approved users evaluate the same fixed clip identifiers, degradation labels, and evaluation partitions

reported in the paper, while the released metadata records the provenance and access condition of each clip without exposing patient information.

F.3 Reproducibility and Release

Upon publication, we will release the degradation synthesis code, dataset manifests and annotations, model configurations, training hyperparameters, and evaluation scripts used in the reported experiments. The reproducibility package will organize the released information into four types of records.

- **Data records** specify the source, frame interval, degradation configuration, severity setting, and dataset split for each clip. OmniBench records additionally include the verified multi-label annotations and access conditions.
- **Model records** specify the implementation version, training configuration, preprocessing, random seed, and checkpoint used for each restoration method.
- **Evaluation records** specify the metric implementation, temporal reference, downstream prediction protocol, and frame-to-clip aggregation used to produce the reported results.
- **Statistical records** retain the clip-level measurements, training-run summaries, and aggregation records underlying the quantitative analyses.

In conclusion, these records distinguish fixed benchmark specifications, including dataset splits, source identifiers, degradation labels, evaluation protocols, and aggregation rules, from method-specific settings. This organization supports reproduction of the reported experiments and consistent evaluation of future methods under the same benchmark protocol.