On the Robustness of Temporal Vision-Language Models for Surgical Endoscopy Videos

Darakshan Rashid[*1,2], Raza Imam[*,1], Ufaq Khan[1], Muhammad Bilal[3], Shazad Ashraf[4], Dwarikanath Mahapatra[5], Mohammad Yaqub[1], Muhammad Haris Khan[1], Imran Razzak[1], Brejesh Lall[2], Lena Maier-Hein[1,6], and Yutong Xie[1]

[1] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
yutong.xie678@gmail.com
[2] Indian Institute of Technology Delhi, India
[3] Birmingham City University, UK
[4] University Hospitals Birmingham, UK
[5] Khalifa University, UAE
[6] German Cancer Research Center (DKFZ), Germany

Abstract. Temporal vision-language models (TVLMs) offer a reusable, prompt-based interface for surgical video understanding, yet, their robustness under clinically realistic acquisition artifacts in endoscopy remains insufficiently characterized. In practice, degradations such as defocus, haze, motion blur, noise, cautery smoke, and packet loss introduce structured distribution shifts which may compromise video–text alignment. We study the robustness of temporal VLMs under such shifts caused by corruptions in clip frames. We introduce **Endo-C6**, a compact corruption benchmark of six endoscopy-realistic perturbations evaluated at a fixed high severity, and apply it to public Gastrointestinal (GI) endoscopy and laparoscopic cholecystectomy videos. Under a standardized prompt protocol, we benchmark 3 recent surgical TVLM baselines and analyze robustness in both mean and worst-case settings, spanning 294 dataset-level evaluations. Finally, we present **RobustEndoCLIP**, obtained by *few-shot* parameter-efficient tuning with *VeRA*, outperforming existing TVLM baselines. Our findings show that off-the-shelf TVLMs can exhibit severe worst-case collapse under endoscopy-specific corruptions, whereas lightweight few-shot adaptation can substantially improve corrupted performance and robustness without changing the prompt-based interface. We expect Endo-C6 to support standardized robustness reporting and promote more reliable clinical vision-language systems.

Keywords: Robustness · Surgical video · Endoscopy · Vision-Language

1 Introduction

Temporal vision-language models (TVLMs) increasingly act as a *software layer* for surgical video understanding [24,23,22,19,6]: rather than training a bespoke

Dataset and Code is available at: Github
Corresponding Author: Yutong Xie
Equal Contribution *

classifier for each label set, a pretrained video-text representation can be queried with prompts and reused across tasks (phases, tools, events, findings). This is attractive in surgical endoscopy[7], where annotation is costly, label ontologies drift across centers [12,17], and much of the clinical value comes from detecting safety-critical events reliably [21,14]. However, this reuse implicitly assumes video-text alignment remains stable under realistic operating conditions.

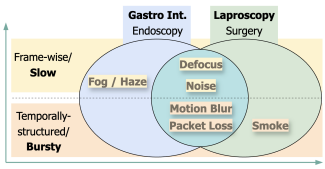
Endoscopy is intrinsically noisy: Endoscopy violates clean-benchmark assumptions [1,2,10]. Signals are shaped by wet optics like lens fog, specular surfaces (glare), non-stationary illumination like overexposure, scope-cast shadows, and abrupt camera motion like motion blur and transient defocus, then further distorted by compression and sometimes live transmission like packet loss. These artifacts are routine: models are relied upon during cautery smoke [15], rapid repositioning, or low-light [3] inspection, when defocus, lens fog [13], blur, and packet loss [18] are most likely. Thus, operational deployment requires robustness, meaning consistent behavior under label-preserving perturbations [9,4].

Why robustness is hard? Robustness benchmarks exist for natural vision [5] and are emerging for medical VLMs [8], but endoscopy needs robustness precisely under adverse conditions (smoke, blur/defocus, low light, compressed video) where rare failures can be high-stakes. We study contrastively trained temporal video-text *dual-encoders* for prompt-based fixed-vocabulary clip classification, yet robustness is hard to compare: cross-dataset transfer confounds artifacts with shifts in procedure mix/labels, and video evaluations often underspecify clip length, stride, and temporal aggregation, which interact with corruptions and alter severity (Fig. 3(c)). Thus, a compact, reproducible protocol that isolates artifact-driven failures while preserving temporal structure remains missing.

Existing gap and direction: Prior surgical TVLMs transfer well under clean/curated settings [24,23,22], and contrastive dual-encoder pretraining (e.g., CLIP) offers a general video-text alignment template [16]. Robustness work shows medical VLMs can fail under perturbations and that parameter-efficient fine-tuning (PEFT) can help [8,7,11], but existing endoscopy robustness studies target depth estimation on SCARED dataset and omit temporal prompt-based VLMs, GI-laparoscopy artifact diversity, and limited-supervision regimes [20]. We fill this gap by stress-testing TVLMs with endoscopy-realistic corruptions, reporting Mean/Worst-case performance, and comparing PEFT under a fixed label budget. Endo-C6 evaluates prompt-based clip classification on CholecT50 surgical phases, Kvasir GI landmark/pathology classes, and TEMSET-24K surgical actions. To this end, we formulate the following research questions:

Q1. Which corruption types are most responsible for failure in zero-shot TVLMs?
Q2. Are corruption effects consistent across Gastrointestinal (GI) endoscopy vs. laparoscopic surgery, or domain-specific?
Q3. Across models, does robustness track clean accuracy, or reflect distinct sensitivity (mean performance vs. worst-case collapse)?

[7] In the remainder of the paper, "endoscopy" refers to surgical endoscopy.



(a) Our endoscopic taxonomy

Dataset ↓	Domain ↓	Train/Test #Clips	Train/Test #Frames	Frames per Clip
CholecT50	Laparoscopic	35/ 5	∼70k/ ∼10k	∼2000
Kvasir	GI Endoscopy	8,000/ 1,200	8,000/ 1,200	1
TEMSET-24K	GI Colorectal TEMS	19,400/ 4,851	∼582k/ ∼146k	30

(b) Dataset statistics and splits

Fig. 1: **Endo-C6 benchmark overview.** (a) Artifact taxonomy across endoscopic domains, organized by domain salience and temporal structure. (b) Dataset statistics across endoscopic domains covered in Endo-C6. Reported train-set counts follow the original dataset splits; unless otherwise stated, we train using only 16% of the listed training data. All evaluations are conducted on the full test splits reported in (b).

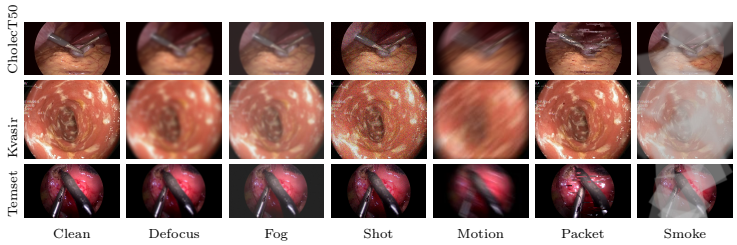


Fig. 2: **Endo-C6 corrupted examples.** Clean and corrupted frames (severity 5) for the six Endo-C6 corruptions on *endoscopy clips* of 3 datasets used in our benchmark.

Q4. With a small label budget, does efficient few-shot tuning improve mean/worst-case robustness?

Contributions: We answer these questions with an in-depth analysis of existing TVLMs across endoscopic domains and highlight how to achieve robustness with our proposed approach. Specifically, we present our contributions:

1. **Diverse Corruption Benchmark:** We define Endo-C6, a diverse *endoscopy-video corruption benchmark*, comprising six endoscopy-realistic corruption types across GI endoscopy and laparoscopic surgery scenarios (See Fig. 1,2).
2. **Benchmarking Study on TVLMs:** We propose a reproducible prompt-based evaluation protocol for endoscopy video clip classification. Under this protocol, we benchmark representative surgical TVLM baselines: SurgVLP [24], HecVLP [23], and PeskaVLP [22], on public endoscopy datasets under clean and corrupted conditions, *analyzing* their fragility and robustness.
3. **Robust Temporal VLM:** We develop RobustEndoCLIP, an endoscopy-specialized temporal CLIP-style classifier obtained by few-shot parameter-efficient tuning with VeRA adapters [11], and provide detailed analysis to conclude *what drives robustness gains and where brittleness remains* in TVLMs.

2 Methodology

2.1 Endo-C6: Robustness Benchmark for Endoscopy Clips

A. Benchmark taxonomy: Endo-C6 follows a two-axis taxonomy capturing *where* artifacts arise and *how* they evolve in time (Fig. 1(a)). **Domain salience** separates GI-skewed artifacts (e.g., haze/fluids), laparoscopy-skewed artifacts (e.g., cautery smoke), and domain-common effects (e.g., blur, compression/streaming). **Temporal structure** separates quasi-static degradations (defocus, shot noise, haze) from bursty processes (motion blur spikes, packet-loss bursts, smoke plumes). This makes failure modes easier to attribute to domain conditions and to temporal dynamics.
B. Benchmarking protocol: We treat clean test clips as In-Distribution (ID) and Endo-C6 variants as controlled Out-Of-Distribution (OOD) shifts. For each test clip x, we evaluate $\{x\} \cup \{c(x)\}_{c\in\text{Endo-C6}}$ on **TEMSET-24K**, **Kvasir** and **CholecT50** independently. Corruptions are used only at test time; training and few-shot tuning (for RobustEndoCLIP) use the corresponding clean train split exclusively. Statistics are shown in Fig. 1(b). We report **294 dataset-level evaluations** (7 adaptation *settings* $\times$ 3 datasets $\times$ 7 conditions [Clean+6] $\times$ 2 stride coverages [50%, 100%]). This corresponds to **593,488 clip-level evaluations** over the Endo-C6 test splits (6,056 test clips $\times$ 7 conditions $\times$ 7 *settings* $\times$ 2 coverages). Here "7 adaptation *settings*" refer to 3 baseline models + 4 possible RobustEndoCLIP's configurations as shown in Fig. 5(c).

2.2 RobustEndoCLIP for Temporal Endoscopy Classification

A. Zero-shot inference: We follow the CLIP interface for endoscopic videos: a temporal video encoder $f_v(\cdot)$ maps a clip x to an embedding, and a text encoder $f_t(\cdot)$ maps a class prompt to the same space, as shown in Fig. 3(a). We compute normalized embeddings $z_v(x) = \frac{f_v(x)}{\|f_v(x)\|}$, $z_y = \frac{f_t(t_y)}{\|f_t(t_y)\|}$, and predict with cosine similarity and temperature τ, where $\mathcal{Y}$ indexes the class set:

$$p(y \mid x) = \frac{\exp(\tau\, z_v(x)^\top z_y)}{\sum_{y'\in\mathcal{Y}} \exp(\tau\, z_v(x)^\top z_{y'})}. \tag{1}$$

All baselines (SurgVLP, HecVLP, PeskaVLP), including RobustEndoCLIP, are inferred in this manner. We use a single canonical template $t_y = \pi(y)$ that inserts the class name into a fixed phrase (e.g., "an endoscopy video frame of {cls}").
B. RobustEndoCLIP – Few-shot adaptation with VeRA: RobustEndoCLIP adapts a pretrained temporal CLIP to endoscopy with a small labeled subset of *clean* training clips. The motivation is pragmatic: zero-shot TVLMs often mis-calibrate to endoscopic appearance and label semantics, so light supervision can "anchor" the embedding space without retraining the backbone.

Few-shot head tuning: Given a labeled endoscopy dataset $\mathcal{D} \supset \{(x_i, y_i)\}_{i=1}^{N}$, we fine-tune on a limited annotated subset using a pretrained TVLM backbone. We initialize from SurgVLP [24] (ResNet-50 visual encoder), freeze the encoders,

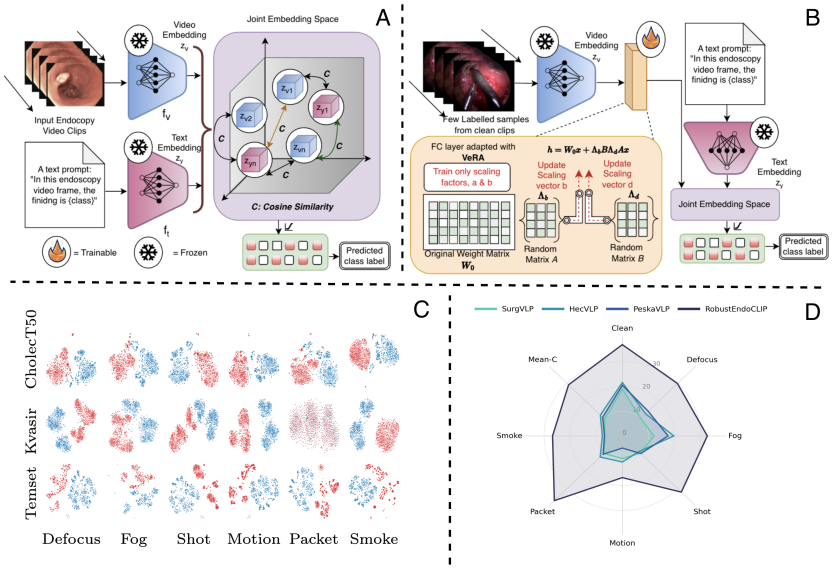


Fig. 3: **A)** Zero-shot inference maps an endoscopy image/frame and text class prompts into a shared embedding space. Prediction is the class whose prompt has the highest cosine similarity. **B)** Few-shot adaptation uses a small clean labeled subset from three datasets: CholecT50 (surgical phase labels), Kvasir (GI landmark/pathology class labels) and Temset24k (surgical action labels). The FC projection layer is adapted with Vector-based Random Matrix Adaptation (VeRA). Please also refer to Sec.2.2. **C)** t-SNE projections of video embeddings showing how each Endo-C6 corruption shifts the feature space relative to clean clips. **D)** On average across three corruption datasets of Endo-C6, RobustEndoCLIP outperforms other TVLM baselines by a large margin.

and update only the final FC layer using a cross-entropy objective over prompt-induced class scores:

$$\mathcal{L}_{\text{cls}} = -\sum_{(x_i,y_i)\in\mathcal{D}_s} \log \frac{\exp(\tau\, z_v(x_i)^\top z_{y_i})}{\sum_{y'\in\mathcal{Y}} \exp(\tau\, z_v(x_i)^\top z_{y'})}. \tag{2}$$

This preserves the prompt-driven deployment interface while improving alignment to the target domain. Overall workflow is described in Fig. 3.

Vector-based Random Matrix Adaptation (VeRA): LoRA adapts a weight matrix $W_0 \in \mathbb{R}^{m\times n}$ using a trainable low-rank update $\Delta W = BA$ [7]. VeRA instead fixes random matrices A, B and learns only scaling vectors (Fig. 3(b)), yielding (following [11])

$$h = W_0 x + \Delta W x = W_0 x + \Lambda_b\, B\, \Lambda_d\, A\, x, \tag{3}$$

where Λ_b and Λ_d are diagonal matrices formed from trainable vectors b and d. Intuitively, VeRA constrains the update direction and reduces the number of free parameters, which is desirable in low-shot settings to limit overfitting and

Table 1: **Robustness Analysis of Temporal VLMs on Endo-C6.** Accuracy (%) on clean and corrupted test clips across three datasets. **Bold** denotes Best, while Underline denotes Second-best.

(a) CholecT50 / Laparoscopy (downstream)

Temporal MVLM ↓	Clean	Defocus	Fog	Shot	Motion	Packet	Smoke	Mean-C	Worst-C	Δ
SurgVLP	40.21	16.15	21.86	11.39	7.38	11.58	7.08	12.57	7.08	0.00
HecVLP	44.41	18.96	32.54	17.09	12.78	17.44	7.30	17.69	7.30	5.12
PeskaVLP	39.23	24.32	37.59	14.89	7.74	13.30	6.91	17.46	6.91	4.89
RobustEndoCLIP	**47.06**	**42.09**	**42.92**	**42.66**	**16.83**	**46.30**	**38.91**	**38.29**	**16.83**	**25.72**

(b) Kvasir / Gastrointestinal (downstream)

Temporal MVLM ↓	Clean	Defocus	Fog	Shot	Motion	Packet	Smoke	Mean-C	Worst-C	Δ
SurgVLP	13.75	9.50	12.25	12.80	**14.75**	14.08	10.91	12.38	9.50	0.00
HecVLP	16.20	15.91	20.16	8.25	13.75	14.91	**12.16**	14.19	8.25	1.81
PeskaVLP	16.20	14.83	12.08	12.25	6.50	**15.08**	11.08	11.97	6.50	-0.41
RobustEndoCLIP	**36.50**	**18.83**	**26.17**	**26.08**	13.50	**37.17**	11.42	**22.20**	**11.42**	**9.82**

(c) TEMSET-24K (downstream)

Temporal MVLM ↓	Clean	Defocus	Fog	Shot	Motion	Packet	Smoke	Mean-C	Worst-C	Δ
SurgVLP	2.82	4.24	2.23	5.17	5.10	4.35	2.07	3.86	2.07	0.00
HecVLP	3.92	2.98	6.59	1.49	4.92	3.67	3.51	3.86	1.49	0.00
PeskaVLP	5.78	2.98	2.87	2.88	0.40	2.86	2.56	2.43	0.40	-1.43
RobustEndoCLIP	**26.67**	**28.67**	**28.59**	**27.14**	**19.98**	**27.05**	**29.92**	**26.89**	**19.98**	**23.03**

stabilize tuning. We apply VeRA to the final FC layer of the vision encoder only while keeping the backbone frozen. VeRA (being 704× lighter) directly outperforms LoRA under matched capacity as discussed in Tab. 2 and Fig. 5.

3 Experimental Setup

A. Implementation Setup: In our experiments, few-shot tuning uses a ResNet-50 visual backbone (SurgVLP instantiation). We use *few-shot* for clean-label budgets up to 16% (0/4/8/16% in Fig. 4), and keep all remaining training data unused to preserve the ID/OOD interpretation of Endo-C6. This 16% budget is applied per dataset (TEMSET-24K, Kvasir, and CholecT50) using each dataset's own train split. Optimization uses Adam for 50 epochs with learning rate 10^{-3}, batch size 16, and 224×224 resolution. Endo-C6 corruptions (severity 5) are applied exclusively at test time to maintain a clean-to-corrupted distribution shift.

We report Top-1 accuracy on clean clips, Mean-C (average accuracy across the six corruptions), Worst-C (lowest accuracy among the corruptions), and the robustness gap $\Delta = \text{Mean-C} - \text{Mean-C}_{\text{SurgVLP}}$ denotes the *Mean-C gain* over SurgVLP (dataset-wise).

B. Comparative TVLMs: We compare against representative temporal vision-language models used in surgical video understanding, including SurgVLP, HecVLP, and PeskaVLP. These models differ in their pretraining supervision and the surgical video-text sources used for alignment, but none are trained with explicit exposure to the Endo-C6 corruption operators. Unless explicitly specified as a tuned variant, models are evaluated in their off-the-shelf configuration, using the same clip sampling policy and the same class-name prompt template, so that

Table 2: **PEFT ablation.** Comparison of LoRA vs VeRA (and zero-shot) under a fixed label budget, reported across datasets. For each dataset we report Clean and Mean-C, and we also report the Average across datasets. **Best**. Second-best.

Adaptation	# Params	Label Budget	CholecT50		Kvasir		TEMSET-24K		**Average**	
			Clean	Mean	Clean	Mean	Clean	Mean	Clean	Mean
Zero-shot	0	0	40.21	12.57	13.75	12.38	2.82	3.86	18.93	9.60
LoRA	720896	16%	**54.18**	18.31	15.92	14.29	**29.48**	18.17	33.19	16.92
VeRA	**1024**	16%	47.06	**38.29**	**36.50**	**22.20**	26.67	**26.89**	**36.74**	**29.13**

observed degradation under Endo-C6 reflects distribution shifts rather than differences in evaluation protocol.

4 Results and Discussion

4.1 Main Results

Tab. 1 summarizes clean and Endo-C6 performance across three datasets.

A. Corruption-wise breakdown (Q1): Tab. 1 shows corruption-specific failure modes. On CholecT50, **smoke** is the dominant worst-case for baselines (Worst-C $\approx$ 6.9-7.3%), while RobustEndoCLIP lifts the tail to 16.83%. On Kvasir, the weakest corruption is model-dependent (motion for PeskaVLP at 6.50, shot for HecVLP at 8.25, and defocus for SurgVLP at 9.50), indicating non-uniform sensitivity across TVLMs. On TEMSET-24K (downstream), baselines remain unstable (Worst-C as low as 0.40), whereas RobustEndoCLIP improves accuracy across all six corruptions and raises Worst-C to 19.98%. Overall, the largest gains align with deployment-relevant artifacts, notably **smoke** in laparoscopy and particular improvements under **defocus**/**fog**/**shot** in GI.

B. TVLMs across endoscopy domains (Q1, Q2): Failure modes differ across domains. CholecT50 is most sensitive to **smoke** (baseline Worst-C $\approx$ 7%), consistent with cautery-driven visibility loss, whereas GI datasets (Kvasir, TEMSET-24K) exhibit more heterogeneous worst cases spanning **motion**, **shot noise**, and **defocus** depending on the model. RobustEndoCLIP shifts performance upward in both domains; however, the limiting corruption differs by dataset (smoke for CholecT50, motion for Kvasir, and motion for TEMSET-24K), reinforcing that endoscopy robustness is driven by domain- and dataset-specific artifact profiles rather than a single universal failure mode.

C. Robustness of TVLMs (Q3, Q4): Across all three datasets, off-the-shelf TVLMs exhibit a pronounced robustness gap: corrupted accuracy and especially Worst-C can collapse even when clean accuracy is moderate. Among baselines, HecVLP is typically strongest on Mean-C (CholecT50/Kvasir), while SurgVLP and PeskaVLP show sharper worst-case failures depending on the dataset. RobustEndoCLIP achieves the highest Mean-C and Worst-C across datasets while preserving the prompt-based interface, indicating that lightweight few-shot adaptation can improve both average robustness and the tail.

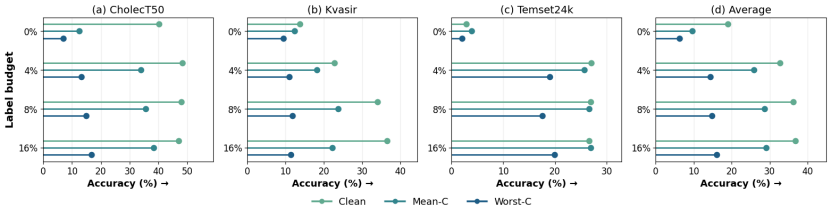


Fig. 4: **Label budget scaling.** We compare the robustness of RobustEndoCLIP on different Endo-C6 datasets when tuning with different labeled fractions of the *clean* training set. We report Clean, Mean-C, and Worst-C across datasets.

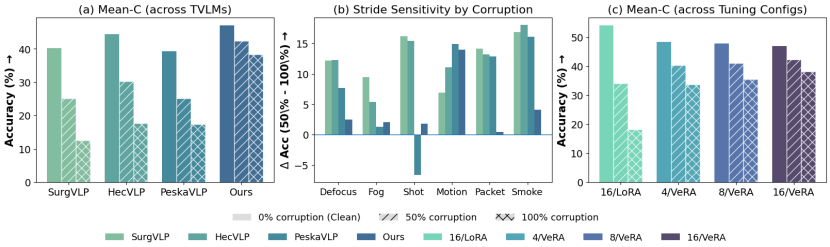


Fig. 5: **Corruption stride analysis.** (a) Mean-C for temporal VLMs under clean clips and random corruption applied to 50% vs. 100% of clip frames. (b) Per-corruption stride sensitivity, ΔAcc = Acc(50%) − Acc(100%), across TVLMs (lower is better). (c) Mean-C across tuning configurations (LoRA/VeRA, varying label budgets) under different stride settings. "Ours" and 16/VeRA are same representing RobustEndoCLIP.

4.2 Discussion and Analyses

Effect of adaptation (VeRA vs. LoRA) (Q4): Tab. 2 compares zero-shot, LoRA, and VeRA under the same 16% label budget with head-only adaptation. VeRA delivers the strongest robustness, achieving the best Mean-C on each dataset and the best average Mean-C (29.13). LoRA improves clean accuracy (notably on CholecT50) but yields smaller Mean-C gains, indicating weaker robustness transfer under Endo-C6.

Label budget scaling (0/4/8/16%) (Q4): Fig. 4 shows that supervision improves robustness, but with dataset-dependent returns. On CholecT50, Mean-C and Worst-C increase steadily with label budget. On Kvasir, clean accuracy continues to rise, while Mean-C saturates earlier. On TEMSET-24K, few-shot tuning quickly lifts both clean and corrupted accuracy, reducing the clean-corrupted gap.

Effect of corruption stride in clips (50% vs. 100%) (Q1, Q3): Fig. 5 shows that increasing corruption coverage from 50% to 100% consistently lowers Mean-C. Stride sensitivity ΔAcc = Acc(50%) − Acc(100%) is largest for temporally structured artifacts (motion, packet loss, smoke). Across tuning settings, RobustEndoCLIP's VeRA retains higher Mean-C at both coverages, indicating gains persist under denser corruption.

5 Conclusion

We introduced Endo-C6, a compact, fixed-severity corruption benchmark and evaluation protocol for assessing the robustness of temporal CLIP-style VLMs on endoscopy videos. Across GI and laparoscopic domains, Endo-C6 exhibits robust mean and worst-case fragility in off-the-shelf surgical TVLM baselines, while RobustEndoCLIP, a lightweight VeRA-based few-shot adaptation, substantially improves corrupted performance under matched supervision. We expect Endo-C6 to serve as a practical reporting template for robustness claims in endoscopic video understanding, emphasizing both average behavior and tail failures.

Limitations and future work. This study focuses on prompt-based *classification* with CLIP-style temporal VLMs; extending the benchmark to additional endoscopy tasks (e.g., action triplets, detection/segmentation) and to *generative* multimodal models is a natural next step to explore.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ali, S., Zhou, F., Bailey, A., Braden, B., East, J.E., Lu, X., Rittscher, J.: A deep learning framework for quality assessment and restoration in video endoscopy. Medical image analysis **68**, 101900 (2021) 2
2. Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R.D., et al.: An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Scientific reports **10**(1), 2748 (2020) 2
3. Chen, T., Lyu, Q., Bai, L., Guo, E., Gao, H., Yang, X., Ren, H., Zhou, L.: Lightdiff: Surgical endoscopic image low-light enhancement with t-diffusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 369–379. Springer (2024) 2
4. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019) 2
5. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (ICLR) (2019) 2
6. Honarmand, M., Jamal, M.A., Mohareri, O.: Vidlpro: a video-language pre-training framework for robotic and laparoscopic surgery. In: Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond (2024) 1
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022) 2, 5
8. Imam, R., Marew, R., Yaqub, M.: On the robustness of medical vision-language models: Are they truly generalizable? In: Annual Conference on Medical Image Understanding and Analysis. pp. 233–256. Springer (2025) 2

9. Jaspers, T.J., Boers, T.G., Kusters, C.H., Jong, M.R., Jukema, J.B., De Groof, A.J., Bergman, J.J., de With, P.H., van der Sommen, F.: Robustness evaluation of deep neural networks for endoscopic image analysis: Insights and strategies. Medical Image Analysis **94**, 103157 (2024) 2
10. Kim, B.S., Lee, H.J., Park, J., Jeong, S.Y., Kim, M.J., Ahn, J.S., Park, J.W., Kim, S.: Automated endoscopic image quality assessment for robust colorectal lesion detection. BMC Medical Imaging **25**(1), 473 (2025) 2
11. Kopiczko, D.J., Blankevoort, T., Asano, Y.M.: Vera: Vector-based random matrix adaptation. arXiv preprint arXiv:2310.11454 (2023) 2, 3, 5
12. Lavanchy, J.L., Ramesh, S., Dall'Alba, D., Gonzalez, C., Fiorini, P., Müller-Stich, B.P., Nett, P.C., Marescaux, J., Mutter, D., Padoy, N.: Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. International journal of computer assisted radiology and surgery **19**(11), 2249–2257 (2024) 2
13. Lawrentschuk, N., Fleshner, N.E., Bolton, D.M.: Laparoscopic lens fogging: a review of etiology and methods to maintain a clear visual field. Journal of endourology **24**(6), 905–913 (2010) 2
14. Mohamadipanah, H., Kearse, L., Wise, B., Backhus, L., Pugh, C.: Generating rare surgical events using cyclegan: Addressing lack of data for artificial intelligence event recognition. Journal of Surgical Research **283**, 594–605 (2023) 2
15. Pan, Y., Bano, S., Vasconcelos, F., Park, H., Jeong, T.T., Stoyanov, D.: Desmokelap: improved unpaired image-to-image translation for desmoking in laparoscopic surgery. International Journal of Computer Assisted Radiology and Surgery **17**(5), 885–893 (2022) 2
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 2
17. Reiter, W.: Domain generalization improves end-to-end object detection for real-time surgical tool detection. International Journal of Computer Assisted Radiology and Surgery **18**(5), 939–944 (2023) 2
18. Shor, J., Johnston, N.: Does video compression affect cade polyp detectors? Gastrointestinal Endoscopy **97**(6), AB768 (2023) 2
19. Walimbe, S., Baby, B., Srivastav, V., Padoy, N.: Adaptation of multi-modal representation models for multi-task surgical computer vision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–33. Springer (2025) 1
20. Wang, A., Yin, H., Cui, B., Xu, M., Ren, H.: Benchmarking robustness of endoscopic depth estimation with synthetically corrupted data. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 45–57. Springer (2024) 2
21. Wei, H., Rudzicz, F., Fleet, D., Grantcharov, T., Taati, B.: Intraoperative adverse event detection in laparoscopic surgery: stabilized multi-stage temporal convolutional network with focal-uncertainty loss. In: Machine Learning for Healthcare Conference. pp. 283–307. PMLR (2021) 2
22. Yuan, K., Navab, N., Padoy, N., et al.: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. Advances in Neural Information Processing Systems **37**, 122952–122983 (2024) 1, 2, 3
23. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: International Conference

on Medical Image Computing and Computer-Assisted Intervention. pp. 306–316. Springer (2024) 1, 2, 3
24. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J.L., Marescaux, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures. Medical Image Analysis **105**, 103644 (2025) 1, 2, 3, 4