Same Errors, Different Models? Cross-FM Failure Agreement in Surgical Phase Recognition

Anonymous

Anonymous Institution

Abstract. Surgical phase recognition remains challenging because visually similar phases, variable surgical styles, and temporal continuity blur phase boundaries. Recent work has improved aggregate accuracy through stronger visual backbones and domain-specific surgical pretraining. However, aggregate scores do not reveal whether the models confuse the same surgical phase pairs. We compare DINOv2, a general-purpose backbone pretrained on natural images, and LemonFM, a surgical foundation model pretrained on large-scale surgical video, in a frozen-feature setting across three laparoscopic datasets: Cholec80, MultiBypass140, and AutoLaparo. LemonFM improves linear-probe accuracy by 4.8–5.7 percentage points (pp) over DINOv2. Beyond accuracy, we treat the off-diagonal structure of each confusion matrix as a behavioural signature and measure cross-model *failure agreement*. The two models show high agreement on Cholec80 ($r = 0.81$) and MultiBypass140 (MB140; $r = 0.75$), and a lower point estimate on AutoLaparo ($r = 0.39$). This ordering persists after resolution and preprocessing controls, with $|\Delta r| \leq 0.05$. These findings show that the failure agreement between FMs varies across surgical workflows, even under the same model pair and evaluation protocol. High agreement suggests that simply replacing one FM with another may not resolve shared confusions, whereas the lower AutoLaparo estimate points to potentially more model-specific failure patterns that should be confirmed on larger datasets.

Keywords: Error consistency · Failure agreement · Foundation models · Model comparison · Surgical phase recognition.

1 Introduction

Surgical foundation models (FMs) have rapidly proliferated in the last two years. FM refers to a large-scale pretrained visual backbone that can be reused for downstream surgical tasks through feature extraction, probing, or fine-tuning. Despite progress in both surgical pretraining and general-purpose visual representation learning, current evaluations still rely largely on aggregate performance measures, such as precision and macro-F1, often reported at the level of individual datasets [10,14,6,3].

Aggregate accuracy alone cannot determine whether two foundation models fail in the same way. Surgical phases follow an ordered but variable workflow,

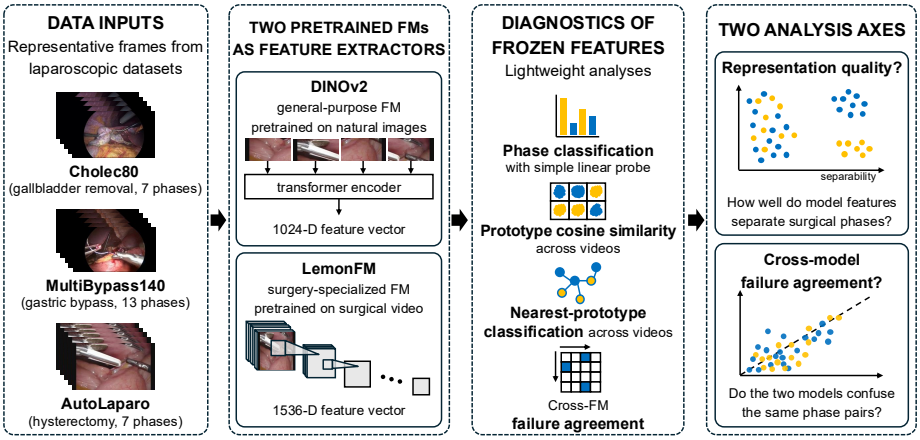

Fig. 1: Overview of the analysis framework. We compare surgical foundation models using both representation quality, which measures how well features separate surgical phases, and failure agreement, which measures whether models confuse the same phase pairs.

where visually similar stages, blurred temporal boundaries, and annotation uncertainty can make specific phase confusions intrinsically difficult. In contrast, model-specific errors may reveal differences in the representations learned by each foundation model. Models with similar aggregate performance can still differ substantially in where they fail. Therefore, we compare DINOv2 and LemonFM beyond aggregate performance by measuring the agreement of their specific phase confusions in surgical workflows.

In general computer vision, previous work has compared models beyond aggregate accuracy. Geirhos et al. [5] introduced trial-level error consistency, and TIDE [2] decomposes object-detection errors into interpretable error types. In surgical phase recognition, Funke et al. [4] and surveys of surgical workflow analysis [8] have also argued that evaluation should consider more than a single summary score. However, surgical-FM comparisons still mainly report aggregate performance, and it is unclear whether different FMs fail on the same surgical phase pairs across datasets.

Contributions.

1. A pairwise FM failure-agreement analysis for surgical phase recognition, measuring whether different FMs confuse the same ordered phase pairs.
2. Quantification of failure agreement across three laparoscopic datasets using video-level nearest-prototype confusion matrices and bootstrap intervals.
3. Matched-resolution and common-preprocessing controls show that the agreement pattern is stable, together with an analysis relating high-agreement errors to visually similar surgical stages.

2 Related Work

Recent Surgical FMs spans endoscopic-image masked autoencoder (MAE) pretraining [1], surgical-video pretraining [12,10], surgical vision language pretraining [14], self-supervised learning (SSL) scaling [6], and LemonFM [3]. In parallel, DINOv2 [9] provides a general-purpose self-supervised visual backbone for comparison. These models are still mainly compared by aggregate performance. More broadly, prior work has studied evaluation beyond accuracy, including error consistency [5], TIDE-style error decomposition [2], and surgical workflow metrics [4,8]. We extend this perspective to surgical FMs by comparing whether different backbones confuse the same surgical phase pairs.

3 Method

3.1 Model inputs and feature extraction

We evaluate two FMs in frozen-feature mode, extracting one feature vector per frame at 1 fps:

- **LemonFM** [3] – a surgery-pretrained ConvNeXt-Large FM with 1536-dimensional features and a native 224×224 input pipeline: resize to 256 pixels, center crop to 224×224, and normalize using the ImageNet channel mean and standard deviation.
- **DINOv2 ViT-L/14** [9] – a general-purpose FM with 1024-dimensional features and a native 518×518 input pipeline: resize to 518×518 with bicubic interpolation, center crop to 518×518, and normalize using the ImageNet channel mean and standard deviation.

We use DINOv2 as a representative general-purpose self-supervised comparator, which is widely used, publicly available vision backbone with an established feature-extraction setting. Since each model was evaluated at its recommended input resolution, the comparison includes a resolution difference between the two FMs. Evaluating each FM at its author-recommended input resolution is common when comparing heterogeneous backbones [14,12]. We control for potential confounding caused by differences in resolution in Section 4.3.

3.2 Datasets

We used three laparoscopic phase-recognition datasets with different anatomies, workflow structures, numbers of annotated phases, and dataset sizes: Cholec80 [11] (cholecystectomy; 80 videos, 7 phases, ∼184k frames @1fps); MultiBypass140 (MB140) [7] (gastric bypass; 140 videos, 13 phases, Bern and Strasbourg centers, ∼778k frames); and AutoLaparo [13] (hysterectomy; 21 videos, 7 phases, ∼83k frames). We perform frame-level classification and analyze phase-confusion patterns, without explicitly modeling the surgical workflow as a temporal graph.

3.3 Diagnostics and the failure-agreement metric

We use the diagnostics to answer two linked questions: whether one FM gives more separable phase representations, and whether the two FMs fail on the same phase pairs. For each dataset and FM, we report class-balanced linear-probe performance using accuracy and macro-F1, the unweighted mean of class-wise F1 scores. To characterize the frozen feature space, we compute cross-video prototype cosine similarity and evaluate a video-level leave-one-out cross-validation (LOOCV) nearest-prototype classifier. Cross-video prototype cosine similarity is computed by first averaging frame features into L2-normalized video-specific phase centroids, excluding video–phase pairs with fewer than five frames. For each phase, we average the pairwise cosine similarities between centroids from different videos, and then average these phase-wise values.

In the LOOCV classifier, one video v is held out. For each phase c, we pool all phase-c frame features from the remaining videos and compute a frame-weighted prototype:

$$\boldsymbol{\mu}_c^{(-v)} = \frac{\sum_{(x,y)\in D^{(-v)},\, y=c} \mathbf{f}(x)}{\left\| \sum_{(x,y)\in D^{(-v)},\, y=c} \mathbf{f}(x) \right\|_2},$$

where $\mathbf{f}(x)$ is the frozen feature vector and $D^{(-v)}$ denotes all frames outside the held-out video. Each held-out frame is L2-normalized and assigned by maximum cosine similarity:

$$\hat{y} = \operatorname*{argmax}_c \left(\frac{\mathbf{f}(x)}{\|\mathbf{f}(x)\|_2} \right)^\top \boldsymbol{\mu}_c^{(-v)}.$$

Aggregating predictions over all held-out videos gives a confusion matrix, which is row-normalized by ground-truth phase so that $C_{ij} = P(\hat{y} = j \mid y = i)$.

Our main metric is *cross-FM failure agreement*. For each dataset and model $m \in \{\textsc{LemonFM}, \mathrm{DINOv2}\}$, we extract the off-diagonal entries of the row-normalized LOOCV confusion matrix:

$$\mathbf{e}^m = \mathrm{vec}_{i\neq j}(C_{ij}^m),$$

where each entry is an ordered phase pair from ground-truth phase i to predicted phase j. Failure agreement is the Pearson correlation

$$r = \mathrm{corr}(\mathbf{e}^{\textsc{LemonFM}}, \mathbf{e}^{\mathrm{DINOv2}}).$$

High r means that the two FMs confuse similar phase pairs, whereas low r means that their confusion patterns differ. We use Pearson correlation because the off-diagonal entries are continuous confusion rates. CIs for r are estimated with a video-level bootstrap. We resample whole videos with replacement 2,000 times, aggregate the corresponding confusion matrices, and recompute r for each resample. This treats videos, rather than frames, as the independent sampling units.

Table 1: Frozen-feature comparison. Acc, Macro-F1, and nearest-prototype accuracy (Proto Acc) under LOOCV are performance metrics, where higher is better. Cross-vid sim is the mean cross-video prototype cosine similarity and is reported as a feature-space diagnostic.

Dataset	FM	Acc	Macro-F1	Proto Acc	Cross-vid sim
Cholec80	DINOv2	0.727	0.645	0.477	0.937
	LemonFM	**0.775**	**0.689**	**0.658**	0.659
MB140	DINOv2	0.725	0.611	0.465	0.952
	LemonFM	**0.782**	**0.652**	**0.621**	0.660
AutoLaparo	DINOv2	0.669	0.561	0.499	0.939
	LemonFM	**0.725**	**0.646**	**0.704**	0.662

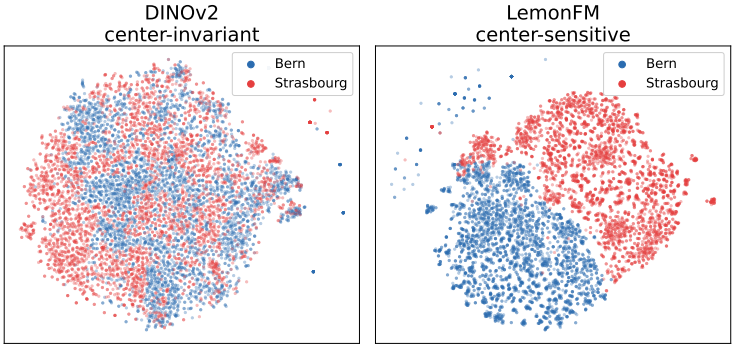

Fig. 2: Cross-center t-SNE of the MB140 frozen feature space, colored by center. The Bern and Strasbourg centers are more intermixed for DINOv2 and more separated for LemonFM, consistent with the different feature-space behavior observed in Table 1.

4 Results

4.1 Which FM is better: consistent surgical-FM advantage

Across all three datasets, the surgical FM provides the more discriminative frozen representation (Table 1). LemonFM leads the linear probe by +4.8 to +5.7 pp in accuracy and by +4.1 to +8.5 pp in macro-F1. It also achieves higher per-video LOOCV nearest-prototype accuracy on most videos (paired t-test with videos, $p \leq 2.5 \times 10^{-10}$). LemonFM's higher linear-probe accuracy, macro-F1, and nearest-prototype accuracy support stronger phase discriminability. DINOv2 produces higher same-phase cross-video prototype similarity (≈ 0.94, Table 1), whereas LemonFM shows lower values (≈ 0.66). This indicates that DINOv2 forms more video-consistent same-phase prototypes, while LemonFM's same-phase prototypes vary more across videos. Therefore, we treat cross-video similarity as a consistency diagnostic.

Table 2: Cross-FM failure agreement from row-normalized LOOCV nearest-prototype confusion matrices. $n_{\text{off-diag}} = K(K-1)$ is the number of ordered off-diagonal phase-pair entries used to compute Pearson r. CIs are estimated by video-level bootstrap with 2,000 resamples. r common-preproc. denotes Pearson r after applying LemonFM's 224×224 preprocessing to DINOv2.

Dataset	$n_{\text{off-diag}}$	#vid	Pearson r	95% CI (boot.)	r common preproc.
Cholec80	42	80	0.806	[0.74, 0.84]	0.853
MB140	156	140	0.750	[0.68, 0.78]	0.742
AutoLaparo	42	21	0.394	[0.14, 0.81]	0.364

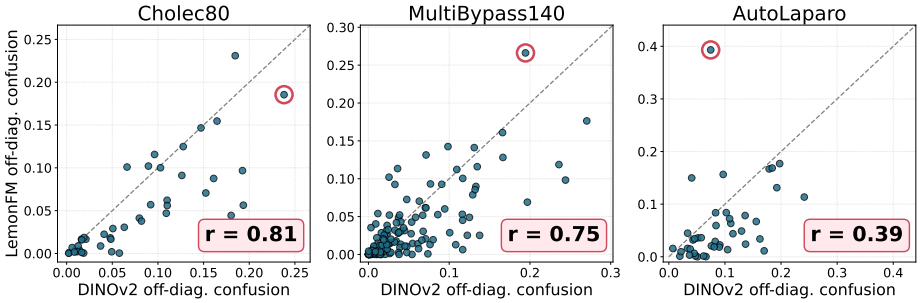

Fig. 3: Cross-FM failure agreement. Each point is one ordered phase pair; axes show row-normalized off-diagonal confusion rates for DINOv2 and LemonFM. Points near the identity line are confused at similar rates by both FMs. Agreement is high on Cholec80 and MB140 and lower on AutoLaparo, although AutoLaparo has a wide bootstrap interval. Circled points mark the most strongly shared confused pair in each dataset.

4.2 Failure agreement varies across surgical procedure types

Table 2 reports the cross-FM failure-agreement correlation r, computed from the off-diagonal entries of the row-normalized LOOCV nearest-prototype confusion matrices. The table also reports a video-level bootstrap confidence interval (CI). All three datasets show positive correlations, indicating that the two FMs tend to confuse some of the same phase pairs. The point estimates decrease from Cholec80 ($r = 0.806$) to MB140 ($r = 0.750$) and AutoLaparo ($r = 0.394$). The bootstrap intervals support high and relatively tight agreement for Cholec80 and MB140. By contrast, AutoLaparo has a wider interval because it contains only 21 videos, so we describe it as a lower but less certain point estimate. Because input resolution is fixed for each FM across datasets, the change in failure-agreement r likely reflects dataset-specific anatomy and workflow structure rather than resolution alone.

Shared high-confusion pairs are closely related surgical stages. In the two high-agreement datasets, the phase pairs both FMs confuse most are

Table 3: Matched-resolution control on Cholec80. DINOv2 is evaluated at its native 518×518 resolution and under LemonFM's 224×224 preprocessing pipeline. LemonFM remains higher on linear-probe accuracy and nearest-prototype accuracy, while DINOv2 retains high cross-video prototype similarity.

Metric	LemonFM@224	DINOv2@518	DINOv2@224
linear-probe accuracy	**0.775**	0.727	0.688
nearest-prototype accuracy	**0.658**	0.477	0.445
cross-video prototype similarity	0.659	0.937	0.919

visually near-identical at the frame level: Cholec80 *CalotTriangleDissection* ↔ *GallbladderDissection*, MB140 *gastrojejunal_ anastomosis* ↔ *anastomosis_ test*. This suggests that part of the failure agreement reflects shared frame-level ambiguity, rather than only model-specific behavior. These pairs are highlighted in Fig. 3.

Lower agreement on AutoLaparo is less certain. AutoLaparo shows a lower point estimate ($r = 0.394$), and the most confused phase pairs overlap less between the two FMs than in Cholec80 and MB140. This may indicate more model-specific failure patterns. However, because AutoLaparo contains only 21 videos and has a wide bootstrap interval, this result should be treated as a hypothesis rather than a confirmed low-agreement regime.

4.3 Robustness: matched resolution and common preprocessing

The main comparison evaluates DINOv2 at 518×518 resolution and LemonFM at 224 × 224, so model architecture and input resolution differ together. To assess the effect of input resolution, we re-extracted DINOv2 features using LemonFM's 224 × 224 preprocessing pipeline and repeated the analyses on all three datasets. The main observations are unchanged: LemonFM remains higher on the evaluated performance metrics in the matched-resolution setting, and reducing DINOv2 to 224 × 224 increases the performance gap rather than reducing it (Table 3). On Cholec80, the linear-probe accuracy gap increases from +0.048 to +0.087, while DINOv2's high cross-video prototype similarity remains similar (0.937 → 0.919). Thus, the control does not support input resolution as the main explanation for LemonFM's performance advantage.

Under common preprocessing, the failure-agreement r changes by at most 0.05, preserving the ordering Cholec80 $\gtrsim$ MB140 > AutoLaparo (Table 2). The agreement ordering is therefore stable under common preprocessing, while the AutoLaparo estimate remains less certain because of its smaller size.

5 Discussion

Different representation behavior. LemonFM and DINOv2 emphasize different aspects of the frozen representation space. LemonFM achieves stronger

phase separability, whereas DINOv2 shows greater consistency across videos and centers. These results suggest that the two FMs organize the frozen feature space differently. LemonFM better supports phase classification. In contrast, DINOv2 produces more video-consistent same-phase representations. Therefore, the results show why surgical-FM evaluation should not rely on a single metric: aggregate accuracy, prototype-based generalization, cross-video consistency, and failure agreement capture different aspects of the representation.

Interpretation of failure agreement. Failure agreement characterizes prototype-based cross-video errors. Computation from video-level LOOCV nearest-prototype confusion matrices reflects whether frozen features form phase prototypes that generalize to unseen videos. High agreement observed in Cholec80 and MB140 suggests shared sources of difficulty, such as visual similarity, temporal ambiguity, or annotation uncertainty. The lower AutoLaparo estimate may indicate more model-specific errors, but requires further validation because the dataset contains only 21 videos and has a wide bootstrap interval.

Limitations. We compare only two FMs in a frame-level frozen-feature setting. Although matched-resolution and common-preprocessing controls support the main ordering, absolute accuracy can still be affected by native input design (224×224 for LemonFM and 518×518 for DINOv2). Temporal models may resolve some single-frame confusions. We quantified the failure agreement using the Pearson correlation, which may not fully capture the absolute magnitude of shared confusions. Alternative agreement metrics, such as Matthews correlation coefficient (MCC), may capture complementary aspects of shared errors. Future work should extend this analysis to additional FMs, fine-tuned models, temporal phase-recognition architectures, and alternative agreement metrics.

6 Conclusion

We reframed surgical foundation model comparison from aggregate performance ranking to failure-pattern analysis. Using video-level LOOCV nearest-prototype classification, we examined whether a surgical foundation model and a general-purpose foundation model confuse the same surgical phase pairs across three laparoscopic datasets. The models showed high failure agreement on cholecystectomy and gastric bypass, but a lower, statistically less certain point estimate on hysterectomy, an ordering that survives matched-resolution and common-preprocessing controls. Failure agreement complements aggregate accuracy by showing whether errors are shared across models or more model-specific. Surgical FM evaluation should therefore consider both overall performance and failure-pattern agreement.

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

References

1. Batić, D., Holm, F., Özsoy, E., Czempiel, T., Navab, N.: EndoViT: Pretraining vision transformers on a large collection of endoscopic images. International Journal of Computer Assisted Radiology and Surgery **19**(6), 1085–1091 (2024). https://doi.org/10.1007/s11548-024-03091-5
2. Bolya, D., Foley, S., Hays, J., Hoffman, J.: TIDE: A general toolbox for identifying object detection errors. In: Computer Vision – ECCV 2020. Lecture Notes in Computer Science, vol. 12348, pp. 558–573. Springer (2020). https://doi.org/10.1007/978-3-030-58580-8_33
3. Che, C., Wang, C., Vercauteren, T., Tsoka, S., Garcia-Peraza-Herrera, L.C.: LEMON: A large endoscopic MONocular dataset and foundation model for perception in surgical settings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42659–42669 (2026)
4. Funke, I., Rivoir, D., Speidel, S.: Metrics matter in surgical phase recognition. arXiv preprint arXiv:2305.13961 (2023). https://doi.org/10.48550/arXiv.2305.13961
5. Geirhos, R., Meding, K., Wichmann, F.A.: Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. In: Advances in Neural Information Processing Systems. vol. 33, pp. 13890–13902 (2020)
6. Jaspers, T.J.M., de Jong, R.L.P.D., Li, Y., Kusters, C.H.J., Bakker, F.H.A., van Jaarsveld, R.C., Kuiper, G.M., van Hillegersberg, R., Ruurda, J.P., Brinkman, W.M., Pluim, J.P.W., de With, P.H.N., Breeuwer, M., Al Khalil, Y., van der Sommen, F.: Scaling up self-supervised learning for improved surgical foundation models. Medical Image Analysis **108**, 103873 (2026). https://doi.org/10.1016/j.media.2025.103873
7. 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**, 2249–2257 (2024). https://doi.org/10.1007/s11548-024-03166-3
8. Li, Y., Zhao, Z., Li, R., Li, F.: Deep learning for surgical workflow analysis: A survey of progresses, limitations, and trends. Artificial Intelligence Review **57**, 291 (2024). https://doi.org/10.1007/s10462-024-10929-6
9. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
10. Schmidgall, S., Kim, J.W., Jopling, J., Krieger, A.: General surgery vision transformer: A video pre-trained foundation model for general surgery. arXiv preprint arXiv:2403.05949 (2024). https://doi.org/10.48550/arXiv.2403.05949
11. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: EndoNet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging **36**(1), 86–97 (2017). https://doi.org/10.1109/TMI.2016.2593957
12. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14221, pp. 101–111. Springer (2023). https://doi.org/10.1007/978-3-031-43996-4_10

13. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: AutoLaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Lecture Notes in Computer Science, vol. 13437, pp. 486–496. Springer (2022). https://doi.org/10.1007/978-3-031-16449-1_46
14. 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). https://doi.org/10.1016/j.media.2025.103644