arXiv:2604.06099v1 [cs.CV] 7 Apr 2026

Extending ZACH-ViT to Robust Medical Imaging: Corruption and Adversarial Stress Testing in Low-Data Regimes

Athanasios Angelakis
BioML Lab, RI CODE, UniBw, Munich, Germany
AUMC, Amsterdam, Netherlands
athanasios.angelakis@unibw.de

Marta Gomez-Barrero
BioML Lab, RI CODE, UniBw, Munich, Germany
marta.gomez-barrero@unibw.de

Abstract

The recently introduced ZACH-ViT (Zero-token Adaptive Compact Hierarchical Vision Transformer) formalized a compact permutation-invariant Vision Transformer for medical imaging and argued that architectural alignment with spatial structure can matter more than universal benchmark dominance. Its design was motivated by the observation that positional embeddings and a dedicated class token encode fixed spatial assumptions that may be suboptimal when spatial organization is weakly informative, locally distributed, or variable across biomedical images. The foundational study established a regime-dependent clean performance profile across MedMNIST, but did not examine robustness in detail. In this work, we present the first robustness-focused extension of ZACH-ViT by evaluating its behavior under common image corruptions and adversarial perturbations in the same low-data setting. We compare ZACH-ViT with three scratch-trained compact baselines, ABMIL, Minimal-ViT, and TransMIL, on seven MedMNIST datasets using 50 samples per class, fixed hyperparameters, and five random seeds. Across the benchmark, ZACH-ViT achieves the best overall mean rank on clean data (1.57) and under common corruptions (1.57), indicating a favorable balance between baseline predictive performance and robustness to realistic image degradation. Under adversarial stress, all models deteriorate substantially; nevertheless, ZACH-ViT remains competitive, ranking first under FGSM (2.00) and second under PGD (2.29), where ABMIL performs best overall. These results extend the original ZACH-ViT narrative: the advantages of compact permutation-invariant transformers are not limited to clean evaluation, but can persist under realistic perturbation stress in low-data medical imaging, while adversarial robustness remains an open challenge for all evaluated models.

Accepted at CVPR 2026 Workshop (PHAROS-AIF-MIH)

1. Introduction

Vision Transformers are now widely used in computer vision, but their standard formulation still assumes that spatial order should be encoded explicitly through positional embeddings and token-based aggregation. The recently introduced ZACH-ViT [1] questioned whether this assumption is always appropriate for medical imaging, where diagnostically relevant structure may be weakly ordered, locally distributed, or variable across acquisitions. In that setting, the foundational ZACH-ViT study argued that architectural inductive biases should align with the spatial characteristics of the target data rather than assume uniform spatial relevance, and showed a regime-dependent clean performance profile under strict few-shot training.

That first paper primarily established the architectural motivation and the clean performance profile of the model. A related question remained open: *does the same inductive-bias choice remain advantageous when robustness is evaluated directly?* This question is practically relevant because robustness has become a central concern in healthcare machine learning, where performance may shift under acquisition variability, data-quality degradation, and other departures from development conditions [2]. Realistic failures can arise from blur, compression artifacts, brightness or contrast shifts, partial occlusion, and sensor noise rather than only from idealized held-out examples. For compact models intended for deployment close to acquisition or analysis pipelines, clean performance alone is therefore insufficient if performance degrades sharply under such per-

turbations. In low-data clinical settings, demonstrating stability under such conditions is also important for the broader goal of trustworthy deployment.

This problem is especially relevant in light of recent efforts to benchmark robustness under medically meaningful corruptions. MedMNIST-C was introduced precisely to assess robustness under realistic image degradations—including Gaussian noise, blur, contrast adjustment, and compression artifacts—across MedMNIST-style biomedical benchmarks, reflecting the broader recognition that clean-test evaluation alone is not enough to characterize deployment behavior [3]. In parallel, surveys of medical AI robustness continue to emphasize that adversarial fragility and robustness under distribution shift address different, but complementary, failure modes [2, 4]. These perspectives motivate a direct robustness analysis of ZACH-ViT rather than an additional clean-only benchmark.

In this workshop paper, we address that question by treating robustness as the first focused extension of the ZACH-ViT research line. We retain the controlled MedMNIST setting used in the foundational manuscript and compare four scratch-trained compact backbones—ABMIL, Minimal-ViT, TransMIL, and ZACH-ViT—across seven datasets under four evaluation regimes: clean test data, common-corruption means, FGSM means, and PGD means. This design separates three practically relevant aspects of model behavior: baseline predictive performance, stability under realistic non-adversarial degradation, and resilience under gradient-based adversarial stress.

The resulting picture is consistent, although not uniform across settings. ZACH-ViT performs best overall on clean data and under common corruptions, which is the part of the robustness story most directly connected to likely deployment conditions. Under FGSM and PGD, performance declines sharply for all models, and no architecture is consistently best across all datasets. Even so, ZACH-ViT remains competitive and is the best-ranked transformer-based baseline on average under PGD. Taken together, the results do not overturn the original ZACH-ViT narrative. Rather, they refine it by showing that the architectural choices behind ZACH-ViT are not only associated with regime-dependent clean performance, but can also remain advantageous under realistic perturbation stress, whereas adversarial robustness remains a broader open problem.

Contributions. This paper makes three contributions. First, it presents the first robustness-focused extension of the recently introduced ZACH-ViT architecture by revisiting the MedMNIST benchmark under clean, corruption, FGSM, and PGD evaluation regimes. Second, it shows that ZACH-ViT consistently achieves strong mean-rank performance on clean data and under common corruptions across seven MedMNIST datasets, supporting a strong clean-corruption trade-off in a deployment-relevant setting. Third, it provides a cautious adversarial analysis: ZACH-ViT remains competitive under FGSM and PGD, but the results do not support broad adversarial-robustness claims, thereby identifying a clear direction for future work within the broader ZACH-ViT program.

2. Related Work

ZACH-ViT as a compact permutation-invariant backbone. The formal ZACH-ViT manuscript introduced the architecture as a compact Vision Transformer that removes both positional embeddings and the dedicated `[CLS]` token, replacing token-based aggregation with global average pooling over patch representations. The central claim of that work was not universal superiority, but *regime-dependent inductive-bias alignment*: when spatial layout is weakly informative or inconsistent, reducing reliance on fixed spatial priors may be beneficial, whereas stronger positional support can become mildly useful as spatial structure increases. The present paper does not introduce a new architecture; instead, it examines how the same design behaves under robustness stress.

Compact transformers and MIL-style baselines in medical imaging. Transformer-based models in medical imaging include both compact end-to-end backbones and multiple-instance learning (MIL) formulations. TransMIL and ABMIL are especially relevant comparators because they also aggregate information across local representations, although they differ from ZACH-ViT in objective, architectural role, and intended use [5, 6]. ABMIL employs attention-based pooling over instance features, while TransMIL models correlations between instances through a transformer architecture. These distinct aggregation philosophies provide substantive baselines for assessing whether any robustness advantage is specific to a compact permutation-invariant transformer or also emerges in MIL-based alternatives.

Robustness under realistic corruptions and adversarial attacks. Robustness in medical imaging is commonly studied along two complementary axes. Common corruptions approximate realistic variation arising from acquisition differences, compression, nuisance artifacts, and signal degradation [3, 7], whereas adversarial attacks probe sensitivity to worst-case local perturbations [8, 9]. These settings are informative for different reasons and should not be treated as interchangeable. A model may be comparatively stable under realistic corruptions without being strongest under gradient-based attacks. In medical imaging specifically, adversarial fragility and broader robustness failures have both been recognized as meaningful barriers to reliable

deployment [2, 4]. Since compact medical imaging models are more likely to encounter acquisition variability than adaptive white-box attackers in routine use, both perspectives are informative, but they address different deployment questions.

3. Methods

3.1. Backbone models

We evaluate four scratch-trained compact baselines: ABMIL, Minimal-ViT, TransMIL, and ZACH-ViT. ZACH-ViT is the primary method of interest. As introduced in the foundational manuscript, it is a compact permutation-invariant transformer without positional embeddings or a dedicated class token, and it uses adaptive residual projections across dimensional transitions together with global average pooling over patch tokens. The comparison set was selected to remain consistent with the compact-model regime of the foundational ZACH-ViT study: all four models are trained from scratch and have parameter counts below 1M. This restriction keeps the comparison focused on low-capacity backbones that are relevant to low-data and deployment-oriented settings, while still spanning different design philosophies, including MIL-style aggregation and a more conventional compact transformer baseline.

3.2. Datasets and few-shot protocol

We retain the seven MedMNIST datasets used in the main ZACH-ViT study: BloodMNIST, PathMNIST, BreastMNIST, PneumoniaMNIST, DermaMNIST, OCTMNIST, and OrganAMNIST. Binary tasks are evaluated with AUC@0.5 and multi-class tasks with Macro-F1. All models are trained under the same few-shot protocol: 50 training samples per class, unchanged validation and test splits, batch size 16, 23 epochs, and five random seeds $\{3, 5, 7, 11, 13\}$. This low-data setup is intended to emphasize robustness under data scarcity rather than maximize performance through extensive tuning.

3.3. Robustness evaluation

Each model is evaluated under four conditions: (i) **Clean**, corresponding to performance on the original test set; (ii) **Corruption mean**, defined as the average performance across the evaluated common corruptions, namely Gaussian noise, Gaussian blur, brightness-contrast adjustment, JPEG compression, and cutout, each at three severity levels; (iii) **FGSM mean**, defined as the average performance across the evaluated FGSM perturbation strengths ($\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$); and (iv) **PGD mean**, defined as the average performance across the evaluated PGD perturbation strengths using the same ϵ values, 10 attack steps, and step size $\epsilon/4$. Adversarial examples are clipped to the valid input range, and PGD additionally projects perturbed inputs back into the corresponding L_∞ ball around the clean sample. For each dataset and condition, we report the mean and standard deviation over five seeds. These seed-level summaries provide empirical variability estimates, but they should not be interpreted as formal uncertainty quantification. To obtain dataset-agnostic summaries, we compute mean ranks across the seven datasets, where lower rank indicates better performance.

Statistical considerations. To provide a basic indication of variability, we report mean and standard deviation across five random seeds. While the present study is not designed as a formal statistical comparison, the consistency of ranking patterns across datasets suggests that the observed differences are not driven by a single split. A more comprehensive statistical testing framework is left for future work.

4. Results

4.1. The robustness extension preserves the original clean performance profile

Table 1 shows that the clean performance profile remains broadly consistent with the original ZACH-ViT manuscript. The model performs particularly well on DermaMNIST and OCTMNIST, remains competitive on BloodMNIST and BreastMNIST, and stays close to the top-performing model on OrganAMNIST and PneumoniaMNIST. TransMIL remains strongest on PathMNIST, while ABMIL performs notably well on BreastMNIST. The most informative dataset-agnostic summary is the mean rank: ZACH-ViT attains the best overall clean rank (1.57), narrowly ahead of TransMIL (1.71), with ABMIL and Minimal-ViT trailing. This indicates that the robustness extension begins from the same central observation as the foundational paper—namely, that architectural alignment matters more than universal dominance across all regimes.

4.2. ZACH-ViT shows the strongest overall profile under realistic corruptions

The main additional result of this study emerges under common corruptions. Figure 1 and Table 1 show that ZACH-ViT again attains the best mean rank across the seven datasets (1.57), ahead of TransMIL (2.00), ABMIL (3.14), and Minimal-ViT (3.29). This pattern is not attributable to a single dataset. ZACH-ViT remains particularly strong on DermaMNIST, OCTMNIST, and PneumoniaMNIST, while staying close to the best model on BreastMNIST and OrganAMNIST. TransMIL retains a clear advantage on PathMNIST, but the broader pattern is consistent: the inductive-bias choice underlying ZACH-ViT remains favorable when robustness is defined in a deployment-relevant manner.

This result is the most practically relevant finding in the present study. In medical imaging, realistic failures are

more likely to arise from compression, scanner variation, contrast changes, blur, noise, or partial occlusion than from carefully optimized white-box attacks. From that perspective, corruption robustness is not merely an auxiliary benchmark, but an important stress test for compact models intended for deployment-oriented evaluation. In our benchmark, this advantage is observed across multiple corruption families, including noise, blur, brightness-contrast shifts, JPEG compression, and cutout. The present results therefore strengthen the original ZACH-ViT claim by showing that the architectural advantages observed under clean evaluation can persist under plausible perturbations encountered in practice.

4.3. Adversarial robustness remains limited for all models

The adversarial results are more heterogeneous and should be interpreted carefully. Figure 2 and Table 2 show that performance declines sharply for every model under both FGSM and PGD. Under FGSM, ZACH-ViT attains the best overall mean rank (2.00), ahead of TransMIL (2.43), Minimal-ViT (2.57), and ABMIL (3.00). At the level of individual datasets, however, the strongest model varies: ABMIL performs best on BloodMNIST and PneumoniaMNIST, TransMIL on BreastMNIST, and ZACH-ViT on DermaMNIST and OCTMNIST. Under PGD, ABMIL is strongest overall (mean rank 2.00), whereas ZACH-ViT ranks second (2.29), remaining the best-ranked transformer-based baseline on average.

The most appropriate interpretation is therefore comparative rather than absolute. The present robustness extension does not support claims that ZACH-ViT is universally robust to adversarial attacks. Instead, it shows that the model remains among the stronger compact backbones when the threat model becomes substantially harsher. This interpretation is more restrained and more consistent with the empirical evidence. The architectural choices that appear beneficial for clean data and realistic corruptions do not, on their own, resolve white-box robustness.

4.4. Corruption-specific severity plots support the multi-family robustness pattern

To complement the dataset-agnostic summaries, Figure 3 shows representative corruption-specific severity plots drawn from the benchmark outputs. These examples provide direct evidence that the robustness pattern is not limited to a single degradation type or dataset. Across the selected examples, ZACH-ViT remains competitive as severity increases under Gaussian noise, Gaussian blur, brightness-contrast adjustment, JPEG compression, and cutout. The purpose of this figure is not to claim uniform dominance at every severity level, but to show that the favorable corruption profile of ZACH-ViT is visible across multiple corruption families rather than arising from a single artifact of averaging.

Mean rank is used as a dataset-agnostic summary to avoid over-interpreting absolute metric differences across heterogeneous tasks. We therefore interpret mean rank as a coarse comparative signal rather than a definitive measure of superiority.

4.5. Mean-rank and retention summaries clarify the trade-off

Figure 4 summarizes the benchmark from two complementary perspectives. The left panel shows mean ranks across the four evaluation regimes. ZACH-ViT ranks first on clean data and common corruptions, first under FGSM, and second under PGD. No other model matches that overall profile. The right panel reports retention relative to each model's own clean performance. Here ABMIL shows the highest proportional retention under common corruptions (0.96), FGSM (0.31), and PGD (0.30), but it also starts from a weaker clean baseline on most datasets. ZACH-ViT retains 0.92 of its clean performance under common corruptions, 0.23 under FGSM, and 0.18 under PGD, remaining close to or above the two transformer-based comparators.

This comparison clarifies the central message of the paper: robustness is not fully captured by a single scalar summary. A model may show high retention because it begins from a lower clean baseline, whereas another may still be preferable in absolute terms because its starting point is stronger. In the present benchmark, ZACH-ViT provides the best overall clean-corruption trade-off, while ABMIL deserves explicit credit for stronger adversarial retention.

5. Discussion

The main contribution of this paper is that it extends the formal introduction of ZACH-ViT in a way that remains consistent with the original scientific claim. The foundational manuscript argued that architectural performance should be interpreted through inductive-bias alignment with data structure, rather than through universal benchmark dominance alone. Within that framework, removing fixed positional priors can be beneficial when spatial structure is weakly informative or unstable. The present results suggest that this advantage is not limited to clean held-out evaluation. It also carries over, to a meaningful extent, to robustness under realistic corruptions, precisely the regime in which compact deployment-oriented models are likely to encounter difficulties. A controlled ablation isolating the effect of permutation invariance on robustness remains an important direction for future work.

This is relevant for the broader development of ZACH-ViT as a research line. A foundational architecture paper becomes more informative when its first extension is not simply another clean benchmark, but a focused stress

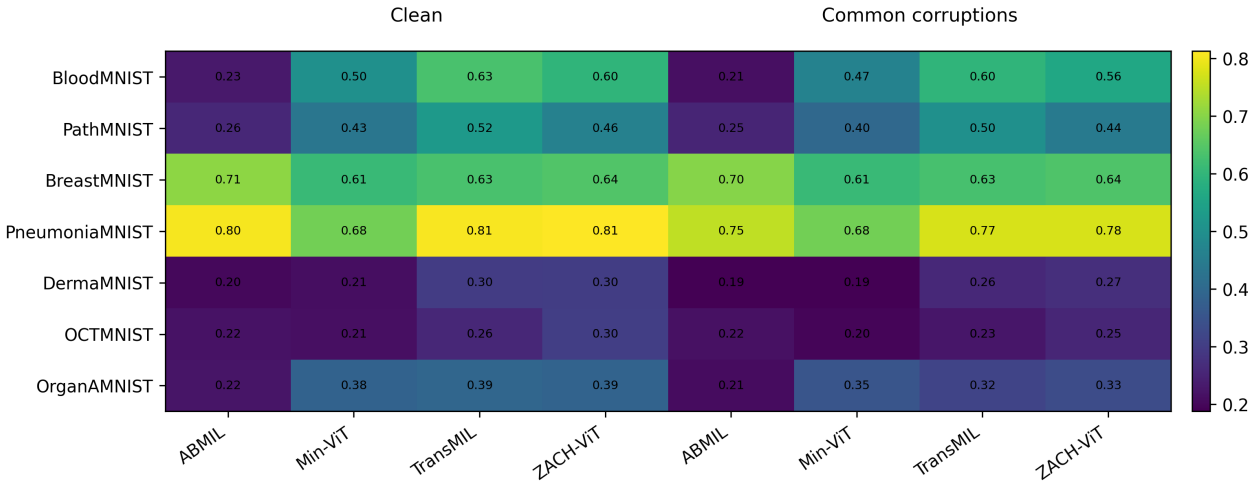

Figure 1. Clean performance and corruption-averaged robustness across the seven MedMNIST datasets. The clean performance pattern reported in the foundational ZACH-ViT study is largely preserved, and the corruption analysis further shows that ZACH-ViT provides the most favorable overall balance between baseline predictive performance and robustness to realistic image degradation.

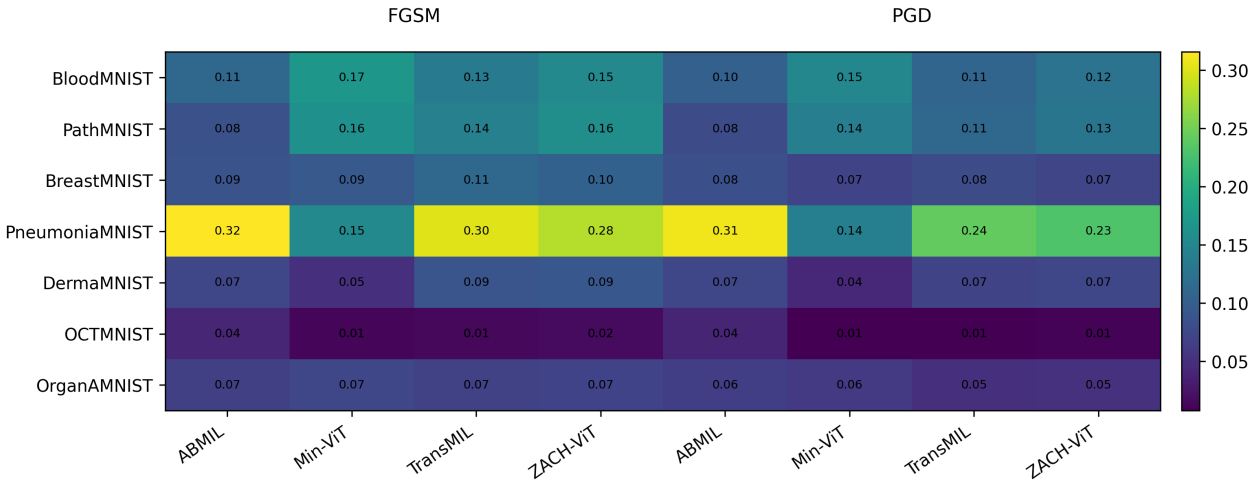

Figure 2. Adversarial stress testing under FGSM and PGD. All models degrade substantially relative to the clean setting. ZACH-ViT remains competitive and is the best-ranked transformer-based baseline on average under PGD, but no architecture is consistently best across all datasets.

test that clarifies where the design is and is not advantageous. In that sense, the current study refines the identity of ZACH-ViT: the model is not defined by dominance in every evaluation regime, but by a balanced combination of compactness, clean predictive performance, and corruption stability in low-data medical imaging. One plausible explanation is that permutation-invariant aggregation reduces reliance on brittle or acquisition-dependent positional correlations when medically relevant signal is distributed rather than tightly anchored to fixed spatial structure, encouraging greater reliance on local visual evidence and compositional statistics, although the present study was not designed to isolate this mechanism directly.

The corruption results also suggest a plausible connection to Edge AI and deployment-oriented medical imaging. Edge and near-device inference are often motivated by latency, connectivity, and privacy constraints in healthcare settings [11, 12]. Within that broader context, a compact model that performs well under realistic corruptions is a reasonable candidate backbone for further evaluation in production-oriented and edge-aware clinical AI pipelines. Our results therefore support ZACH-ViT as a compact backbone that may be relevant to edge-oriented medical imaging workflows, especially when robustness to acqui-

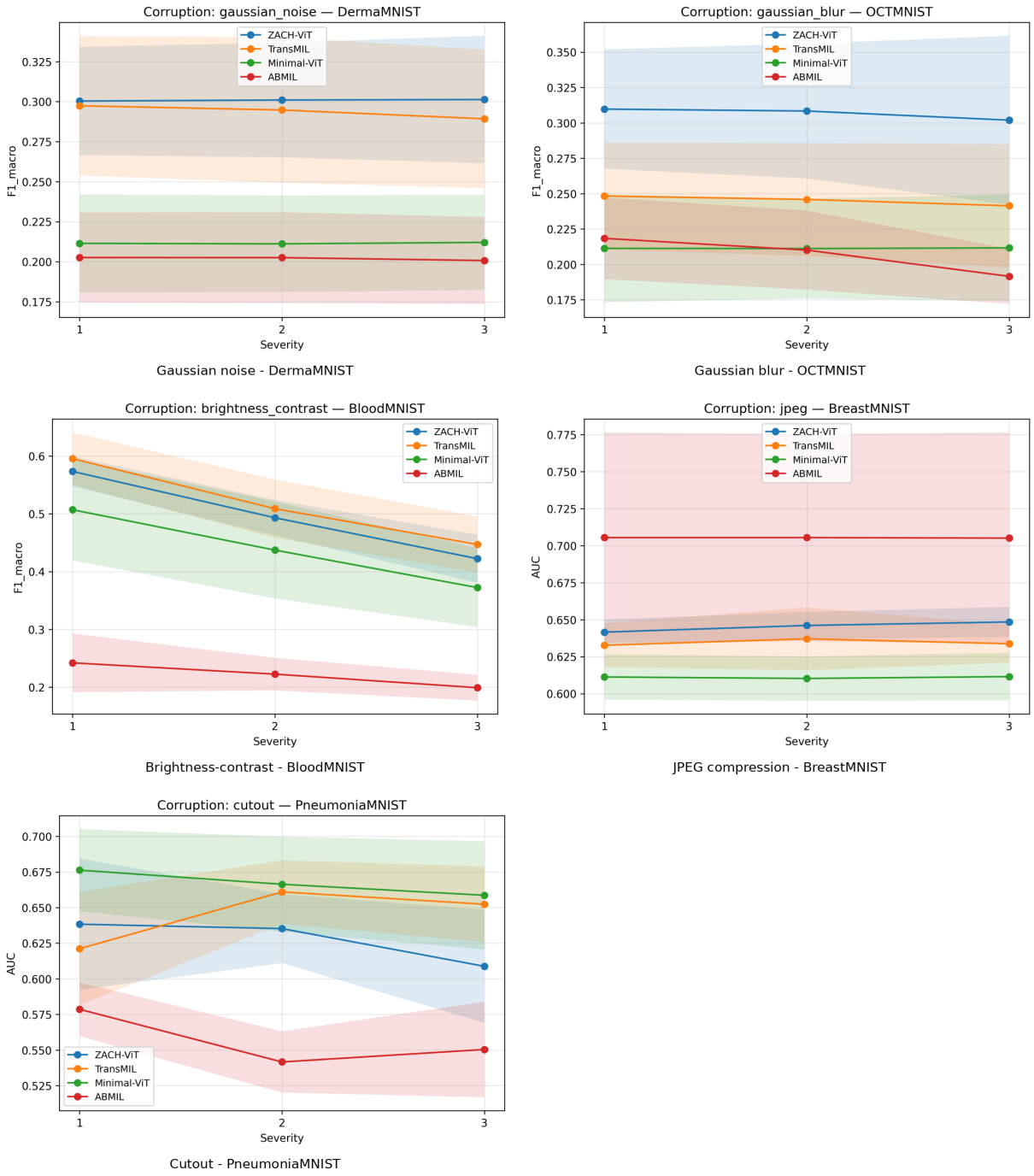

Figure 3. Representative corruption-specific severity plots from the benchmark outputs. The selected panels illustrate that the favorable corruption profile of ZACH-ViT is visible across multiple corruption families, including Gaussian noise, Gaussian blur, brightness-contrast adjustment, JPEG compression, and cutout.

sition variability is important. This interpretation should, however, remain bounded: the present study does not evaluate inference latency, memory footprint, energy use, thermal constraints, integration into medical-device software, or prospective clinical performance. Those aspects would require dedicated hardware and translational validation before any deployment claim could be justified.

For the PHAROS community, where the intersection of

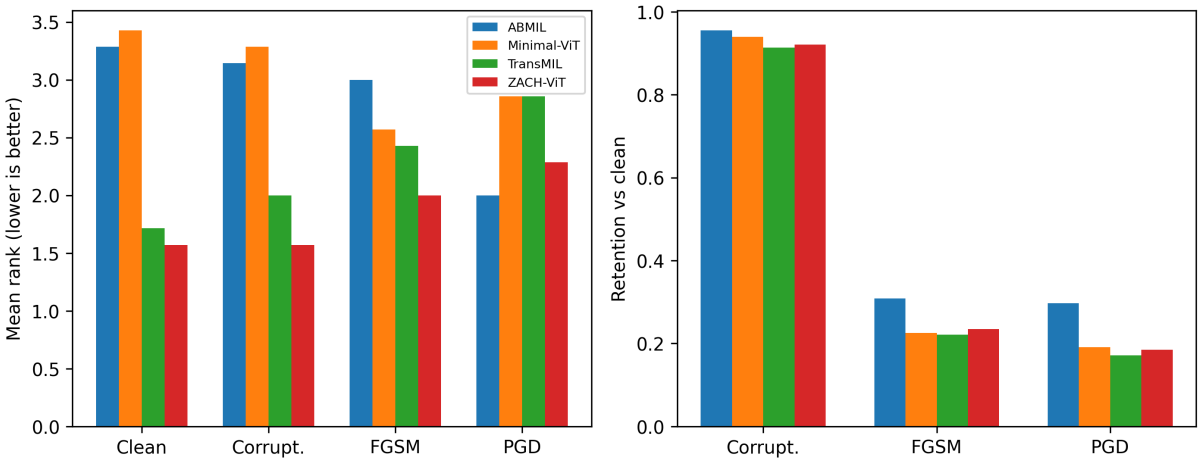

Figure 4. Dataset-agnostic summary of robustness behavior. **Left:** mean rank across clean, corruption, FGSM, and PGD regimes (lower is better). **Right:** average retention relative to each model's own clean performance. ZACH-ViT is the most balanced model overall, while ABMIL shows the strongest proportional retention under adversarial stress.

Table 1. Clean and corruption-averaged performance across the seven MedMNIST datasets. The reported metric is AUC for binary tasks and Macro-F1 for multiclass tasks. Values are mean ± standard deviation over five seeds after averaging across corruption settings within each seed. We note that absolute differences between top-performing models are often small and should be interpreted with caution in the absence of formal hypothesis testing.

	Clean				Common corruptions			
Dataset	ABMIL	Minimal-ViT	TransMIL	ZACH-ViT	ABMIL	Minimal-ViT	TransMIL	ZACH-ViT
BloodMNIST	0.232 ± 0.053	0.502 ± 0.079	0.634 ± 0.042	0.597 ± 0.025	0.214 ± 0.033	0.465 ± 0.073	0.596 ± 0.040	0.559 ± 0.023
PathMNIST	0.257 ± 0.028	0.432 ± 0.066	0.523 ± 0.067	0.461 ± 0.054	0.253 ± 0.031	0.398 ± 0.049	0.501 ± 0.097	0.443 ± 0.054
BreastMNIST	0.706 ± 0.071	0.611 ± 0.015	0.633 ± 0.016	0.642 ± 0.010	0.699 ± 0.064	0.611 ± 0.014	0.630 ± 0.011	0.639 ± 0.006
PneumoniaMNIST	0.804 ± 0.019	0.681 ± 0.031	0.808 ± 0.014	0.813 ± 0.012	0.753 ± 0.019	0.681 ± 0.029	0.774 ± 0.012	0.775 ± 0.010
DermaMNIST	0.202 ± 0.028	0.211 ± 0.029	0.298 ± 0.044	0.301 ± 0.035	0.188 ± 0.023	0.188 ± 0.025	0.259 ± 0.042	0.272 ± 0.035
OCTMNIST	0.219 ± 0.024	0.213 ± 0.039	0.258 ± 0.024	0.304 ± 0.041	0.217 ± 0.027	0.196 ± 0.028	0.226 ± 0.030	0.255 ± 0.028
OrganAMNIST	0.224 ± 0.025	0.384 ± 0.041	0.394 ± 0.024	0.388 ± 0.027	0.209 ± 0.020	0.353 ± 0.024	0.316 ± 0.025	0.334 ± 0.016
Mean rank	3.29	3.43	1.71	1.57	3.14	3.29	2.00	1.57

healthcare AI and practical deployment is central, this study offers two takeaways: (1) architectural choices that appear beneficial in clean evaluation can also confer robustness to realistic acquisition variability, contributing to the development of more trustworthy AI systems; and (2) adversarial robustness remains a distinct challenge requiring dedicated methods beyond inductive-bias tuning. Although this study does not address fairness explicitly, robustness in low-data medical settings should ultimately be evaluated together with subgroup reliability and equitable performance under deployment shift.

At the same time, the adversarial results impose important constraints on interpretation. Neither ZACH-ViT nor any comparator is robust in an absolute sense under FGSM or PGD. A more appropriate conclusion is that the architecture remains competitive under attack while showing clearer advantages under realistic perturbations. That is already a meaningful result for a workshop paper.

6. Limitations

This study has several limitations. First, the benchmark is intentionally restricted to scratch-trained compact baselines and does not include large pretrained foundation models or transfer learning settings, which are commonly used in medical imaging. The present study therefore focuses on controlled low-data behavior rather than state-of-the-art performance under large-scale pretraining. Second, robustness is summarized through averages over corruption types and attack strengths, which is suitable for a comparative workshop study but does not isolate the source of each performance difference. Third, the adversarial analysis is empirical rather than certified and should be interpreted as stress testing rather than formal robustness assurance. In addition, the adversarial evaluation is limited to standard FGSM and short-horizon PGD attacks and should not be interpreted as a comprehensive robustness assessment under stronger or adaptive threat models. Fourth, all datasets

Table 2. FGSM- and PGD-averaged adversarial performance across the seven MedMNIST datasets. Values are mean $\pm$ standard deviation over five seeds after averaging across attack strengths within each seed. FGSM uses $\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$. PGD uses the same ϵ values, 10 steps, and step size $\epsilon/4$. We note that absolute differences between top-performing models are often small and should be interpreted with caution in the absence of formal hypothesis testing.

	FGSM				PGD			
Dataset	ABMIL	Minimal-ViT	TransMIL	ZACH-ViT	ABMIL	Minimal-ViT	TransMIL	ZACH-ViT
BloodMNIST	0.110 ± 0.005	0.170 ± 0.015	0.134 ± 0.021	0.152 ± 0.018	0.102 ± 0.006	0.150 ± 0.015	0.108 ± 0.014	0.125 ± 0.020
PathMNIST	0.084 ± 0.010	0.162 ± 0.010	0.142 ± 0.026	0.160 ± 0.037	0.079 ± 0.008	0.139 ± 0.011	0.111 ± 0.021	0.127 ± 0.036
BreastMNIST	0.086 ± 0.064	0.094 ± 0.021	0.111 ± 0.056	0.103 ± 0.038	0.082 ± 0.061	0.071 ± 0.019	0.079 ± 0.050	0.071 ± 0.031
PneumoniaMNIST	0.316 ± 0.070	0.153 ± 0.055	0.301 ± 0.061	0.282 ± 0.015	0.309 ± 0.068	0.141 ± 0.051	0.242 ± 0.047	0.231 ± 0.011
DermaMNIST	0.074 ± 0.026	0.050 ± 0.021	0.089 ± 0.011	0.091 ± 0.015	0.073 ± 0.026	0.043 ± 0.018	0.070 ± 0.011	0.074 ± 0.015
OCTMNIST	0.041 ± 0.010	0.012 ± 0.007	0.014 ± 0.006	0.017 ± 0.011	0.040 ± 0.010	0.008 ± 0.006	0.008 ± 0.005	0.011 ± 0.007
OrganAMNIST	0.065 ± 0.020	0.073 ± 0.022	0.067 ± 0.016	0.070 ± 0.013	0.065 ± 0.020	0.060 ± 0.019	0.049 ± 0.016	0.051 ± 0.007
Mean rank	3.00	2.57	2.43	2.00	2.00	2.86	2.86	2.29

are drawn from MedMNIST, which provides a useful controlled benchmark but does not substitute for external validation, prospective evaluation, or silent deployment testing in intended-use environments [10]. Fifth, we do not assess subgroup fairness or demographic performance disparities, in part because such annotations are not available in the present benchmark. Finally, the Edge AI relevance discussed above is inferential rather than demonstrated, because no hardware-level measurements or workflow integration experiments are included.

7. Conclusion

We presented the first robustness-focused extension of the recently introduced ZACH-ViT architecture for low-data medical imaging. Across seven MedMNIST datasets and four evaluation regimes, ZACH-ViT achieved the best mean rank on clean data and under common corruptions while remaining competitive under FGSM and PGD attacks. The present study therefore extends, rather than replaces, the original ZACH-ViT narrative: design choices motivated by inductive-bias alignment with spatial structure can provide an attractive balance of predictive performance, realistic robustness, and efficiency in data-constrained medical imaging, even though adversarial robustness remains incomplete. For PHAROS and related healthcare-AI venues, these findings position ZACH-ViT as a principled compact backbone whose strengths are most visible under realistic perturbation stress, and they motivate further evaluation in edge-oriented medical imaging settings. The corruption/adversarial evaluation protocol presented here may also serve as a useful template for other PHAROS participants assessing compact medical imaging backbones.

References

[1] Athanasios Angelakis. ZACH-ViT: Regime-dependent inductive bias in compact vision transformers for medical imaging. arXiv:2602.17929, 2026.

[2] Alan Balendran, Céline Beji, Florie Bouvier, Ottavio Khalifa, Theodoros Evgeniou, Philippe Ravaud, and Raphaël Porcher. A scoping review of robustness concepts for machine learning in healthcare. *npj Digital Medicine*, 8(1):38, 2025.

[3] Francesco Di Salvo, Sebastian Doerrich, and Christian Ledig. MedMNIST-C: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions. arXiv:2406.17536, 2024.

[4] Kyriakos D. Apostolidis and George A. Papakostas. A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, 10(17):2132, 2021.

[5] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, 2021.

[6] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, 2018.

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[10] Jiancheng Yang, Rui Shi Huang, Jiajing Li, et al. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10:48, 2023.

[11] Atslands Rocha, Matheus Monteiro, César Mattos, Madson Dias, Jorge Soares, Regis Magalhães, and José Macedo. Edge AI for Internet of Medical Things: A literature review. *Computers & Electrical Engineering*, 116:109202, 2024.

[12] Vishal Lakshminarayanan, Aswathy Ravikumar, Harini Sriraman, Sujatha Alla, and Vijay Kumar Chattu. Health Care

Equity Through Intelligent Edge Computing and Augmented Reality/Virtual Reality: A systematic review. *Journal of Multidisciplinary Healthcare*, 16:2839–2859, 2023.