WILLIE: A Unified Framework and Benchmark for Wound Classification, Segmentation, and Localization

Gopi Trinadh Maddikunta GMADDIKU@COUGARNET.UH.EDU
Department of Computer Science
University of Houston
Houston, Texas, USA

Peizhu Qian PQIAN@UH.EDU
Department of Computer Science
University of Houston
Houston, Texas, USA

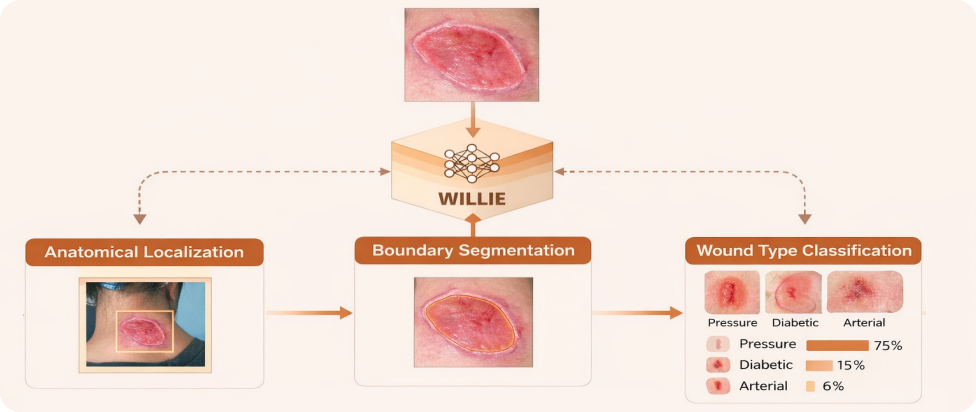

Figure 1: Clinical wound assessment involves three coupled image analysis tasks: anatomical localization, boundary segmentation, and wound type classification. `WILLIE` provides a unified framework for producing these outputs with a single pipeline.

Abstract

Chronic wound management affects over 8.2 million patients in the United States and imposes substantial clinical and economic burden. Clinical wound assessment commonly involves three coupled tasks: identifying wound type, delineating wound boundaries, and localizing the wound region for measurement and monitoring. Despite this clinical coupling, existing machine learning approaches typically address wound classification, segmentation, and localization using separate models. We present **WILLIE**, a unified framework and benchmark for wound classification, segmentation, and localization that enables systematic evaluation of multi-task wound analysis under a common protocol. WILLIE harmonizes three public wound datasets into a shared benchmark and compares unified models across three scaling configurations against 10 single-task baselines. The best model achieves 91.88% classification accuracy, 91.41% Dice, and 96.23% AP@0.5 while producing all three outputs in a single forward pass. Beyond aggregate performance, our results show that segmentation-derived localization outperforms dedicated detection baselines in this benchmark, suggesting that box-based localization may be unnecessary for spatially coherent wound targets. Our findings highlight that effective multi-task learning in healthcare imaging depends not only on shared representations, but also on task formulation, compatibility, and benchmark design.

1. Introduction

Chronic wounds remain a major challenge in healthcare delivery (Frykberg and Banks, 2015; Gould et al., 2015; Martinengo et al., 2019; Olsson et al., 2019). In the United States alone, chronic wound management affects over 8.2 million patients annually and incurs more than $28 billion in treatment costs (Nussbaum et al., 2018; Sen, 2021). The Agency for Healthcare Research and Quality (AHRQ) estimates that more than 2.5 million people develop pressure ulcers annually in U.S. hospitals (Berlowitz et al., 2011), and updated Medicare claims data indicate that chronic wound prevalence among beneficiaries rose from 14.5% in 2014 to 16.3% in 2019 (Carter et al., 2023). These wounds often require prolonged care, repeated assessment, and frequent treatment adjustment over weeks or months. The burden is especially high in settings with limited access to wound specialists, where delays in assessment can lead to infection, hospitalization, amputation, or other complications (Armstrong et al., 2020; Lavery et al., 2006; Swanson et al., 2022; Padula and Delarmente, 2019; Boulton et al., 2005).

Clinical wound assessment is inherently multi-faceted, demonstrated in Figure 1. When evaluating a wound image, clinicians identify wound type, delineate wound boundaries, and localize the wound relative to surrounding anatomy for documentation and longitudinal comparison (Berlowitz et al., 2011; Keast et al., 2004; Sibbald et al., 2021). Wound extent and anatomical context may inform wound type, while accurate boundary delineation is necessary for measuring progression over time. These tasks correspond to classification, segmentation, and localization, respectively. A computational approach that treats these tasks jointly therefore better reflects the way wound assessment is performed in practice. Despite this intuitive coupling, *most existing machine learning approaches address wound classification, segmentation, and detection in isolation.*

In this work, we present **WILLIE**: **W**ound **I**maging, **L**ocalization, and **I**ntegrated **E**valuation, *a unified framework and benchmark for wound classification, segmentation, and localization.* We position WILLIE as both a unified wound analysis framework and a common experimental setting for systematically evaluating multi-task wound analysis. WILLIE integrates three public wound datasets into a shared benchmark and supports direct comparison between unified and single-task models under a common evaluation protocol. This setting allows us to examine not only predictive performance, but also broader questions about task compatibility, foundation-model scaling, and when localization should be learned directly versus derived from segmentation. Our work makes four contributions:

1. We introduce a unified wound analysis benchmark by harmonizing three public datasets into a common evaluation setting for classification, segmentation, and localization.

2. We present WILLIE, a unified framework that supports all three wound-analysis tasks within a single pipeline.

3. We provide a systematic comparison of unified and single-task models across three scaling configurations and 10 baselines under a shared protocol.

4. We identify generalizable insights about multi-task learning in healthcare imaging, including task-dependent scaling behavior, the strengths of segmentation-derived localization, and the importance of task compatibility in benchmark design.

Generalizable Insights about Machine Learning in the Context of Healthcare

This study yields several generalizable insights for healthcare image analysis, multi-task learning, and system design:

- For spatially coherent medical targets, accurate segmentation may provide a stronger foundation for localization than a separately trained detection head. In our benchmark, segmentation-derived localization outperforms dedicated detection baselines, suggesting that when the target is contiguous and clinically defined by extent and boundary, mask prediction may be more effective than box regression alone.
- The benefits of scaling are task-dependent, even within a single clinical application. In our benchmark, classification performance saturates earlier, whereas segmentation and localization continue to improve with larger pretrained backbones. This suggests model selection should not assume uniform returns from scaling across tasks.
- Successful multi-task learning depends not only on shared inputs, but also on compatibility among task objectives. Our results suggest that classification and segmentation benefit from shared learning, whereas localization framed as box regression may be less compatible and may be better derived post hoc.
- Residual error can reflect genuine clinical ambiguity rather than purely algorithmic weakness. In our benchmark, persistent confusion between pressure and venous wounds suggests limits of image-only assessment, where visual evidence alone may be insufficient to resolve etiology without additional clinical context. In such settings, models should support uncertainty flagging and escalation to human review rather than present all predictions as equally definitive.

Data and Code Availability

We release the unified and cleaned benchmark developed in this work and the preprocessing pipeline used to construct the benchmarks. We also provide scripts to reproduce the main results reported in this paper. WILLIE benchmark: https://huggingface.co/datasets/QianGroup/willie-benchmark. Trained models: https://huggingface.co/QianGroup/willie-weights. Code: https://github.com/Qian-Group-HRI/Willie

2. Related Work

Our work builds on three closely related lines of research: wound image analysis, multi-task learning in medical imaging, and vision foundation models for healthcare.

2.1. Wound Image Analysis

Recent wound-image studies have focused primarily on task-specific models for classification, segmentation, or detection. Deep convolutional networks established strong baselines for wound classification (Anisuzzaman et al., 2022; Aldoulah et al., 2023), while the FUSeg challenge (Wang et al., 2022) and subsequent U-Net-based methods (Dhar et al., 2024) standardized evaluation for wound segmentation. More recent work has explored stronger pretrained backbones, ensemble methods (Kiprono, 2025; Ullah et al., 2025), and limited use of clinical metadata (Patel et al., 2024). Wound localization, however, remains less

studied and is typically treated as a standalone detection task based on bounding boxes, without integration with diagnostic classification or pixel-level delineation. Overall, the state of the art remains fragmented across separate tasks and datasets, making it difficult to assess the value of unified wound analysis (Alhababi et al., 2025).

2.2. Multi-Task Learning in Medical Imaging

Multi-task learning has shown promise in medical imaging by leveraging shared structure across related tasks such as classification, segmentation, and detection. Prior work has reported gains in representation quality, data efficiency, and regularization (Mlynarski et al., 2019; Esteva et al., 2017), especially in settings where labeled data are limited. At the same time, successful multi-task learning depends on effective balancing of task objectives (Kendall et al., 2018; Chen et al., 2018; Guo et al., 2018), and existing work has shown that related clinical tasks do not always interact favorably during joint optimization (Standley et al., 2020; Fifty et al., 2021; Yu et al., 2020). Most prior medical imaging studies focus on two-task settings, with fewer examining three-task combinations under a common evaluation protocol. Our work extends this literature by studying classification, segmentation, and localization jointly in a clinically grounded wound-analysis benchmark.

2.3. Vision Foundation Models for Medical Imaging

Large pretrained vision models are increasingly important in medical imaging. Self-supervised models such as DINOv2 (Oquab et al., 2024) provide strong transferable representations for image classification, while ConvNeXt (Liu et al., 2022) remains competitive through strong local inductive biases. Segmentation foundation models such as SAM (Kirillov et al., 2023) and SAM2 (Ravi et al., 2024) have further advanced mask prediction with limited task-specific adaptation, including medical adaptations such as MedSAM (Ma et al., 2024). These developments are especially relevant to wound analysis, where datasets remain modest in size and tasks differ substantially in visual and structural demands. Our work builds on these advances by comparing foundation-model backbones within a unified framework and benchmark for wound classification, segmentation, and localization.

3. The WILLIE Benchmark

A central contribution of this work is the construction of a unified benchmark for wound classification, segmentation, and localization by harmonizing three publicly available datasets: FUSeg, AZH, and Medetec, representative samples shown in Figure 2. The benchmark approximates real-world wound care diversity: FUSeg provides clinical photographs with segmentation masks, AZH provides expert-annotated classification across wound types, and Medetec supplements underrepresented classes. Together, they enable evaluation of clinically coupled wound-analysis tasks under a shared experimental framework.

3.1. Source Datasets

FUSeg. The FUSeg dataset (Wang et al., 2022) was collected through a collaboration with the Advancing the Zenith of Healthcare (AZH) Wound and Vascular Center and the

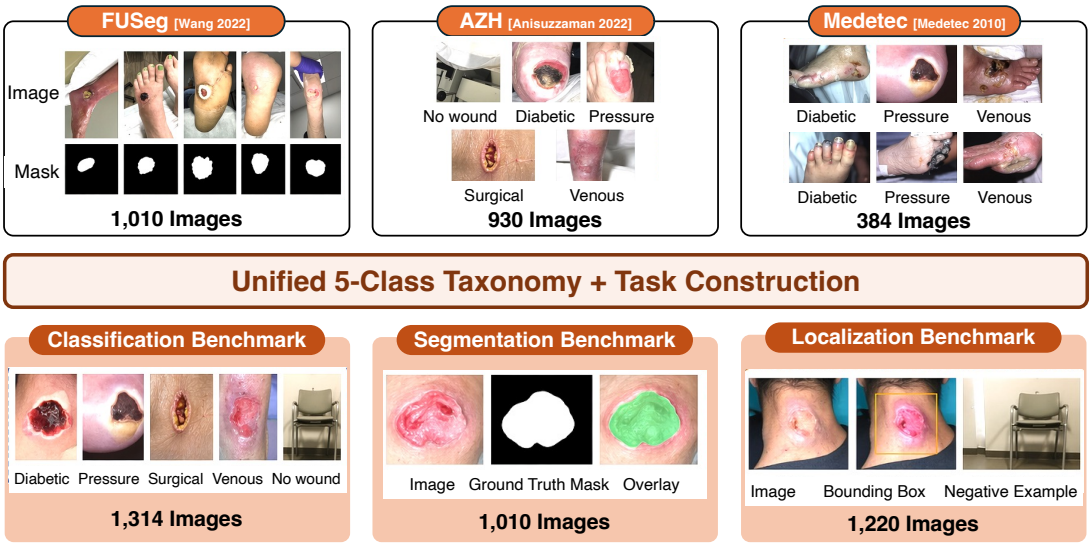

Figure 2: Construction of the WILLIE benchmark from three public wound datasets. FUSeg provides binary masks for segmentation and derived localization, AZH provides expert-annotated wound categories, and Medetec supplements underrepresented wound classes for classification. These sources are harmonized into a unified 5-class taxonomy and shared benchmark supporting classification, segmentation, and localization under a common evaluation protocol.

University of Wisconsin-Milwaukee. This dataset contains 1,010 chronic wound images with binary segmentation masks at 512×512 resolution.

AZH. The AZH dataset (Anisuzzaman et al., 2022) contains 930 clinical wound photographs annotated for classification across six categories: diabetic (185), pressure (134), surgical (164), venous (247), background (100), and no_wound (100). The dataset provides an official split of 696 training images and 234 test images, which we preserve for held-out evaluation. Image dimensions vary substantially, ranging from 93×93 to 259×259 pixels.

Medetec. The Medetec dataset (Medetec, 2010) contains 384 free stock images representing diabetic (81), pressure (170), and venous (133). We use Medetec to supplement underrepresented wound classes without introducing cross-dataset evaluation bias.

3.2. Unified Class Taxonomy

We unify the classification labels into a 5-class taxonomy: diabetic, pressure, surgical, venous, and no_wound. The AZH `background` category is merged with `no_wound`, since both represent the absence of a wound requiring analysis. This yields a moderately imbalanced classification task, with a maximum class ratio of 2.32:1 (venous to surgical). To mitigate this imbalance, we use inverse-frequency class weights and weighted sampling during classification training. Table 1 summarizes the class distribution across datasets and the inverse-frequency weights used during training.

Table 1: Classification class distribution with inverse-frequency training weights.

	Train			Test	
Class	**AZH**	**Medetec**	**Total**	**AZH**	**Weight**
Diabetic	139	81	220	46	0.982
Pressure	100	170	270	34	0.802
Surgical	122	—	122	42	1.765
Venous	185	133	318	62	0.680
No Wound	150	—	150	50	1.434
Total	696	384	1,080	234	—

3.3. Task-Specific Benchmark Construction

Classification. The unified classification set contains 1,314 images. We split these into 918 training images, 162 validation images, and 234 test images using stratified sampling, where the 234-image test set corresponds exactly to the official AZH test split. For 5-fold cross-validation, the combined training and validation pool of 1,080 images is partitioned into five stratified folds, while the AZH test set remains fixed and held out across all folds.

Segmentation. Segmentation experiments use the official FUSeg release, comprising 610 training images and 400 validation images with binary masks at 512×512 resolution. The official FUSeg test set is excluded because its ground truth is not publicly available.

Localization. Localization annotations are derived from FUSeg segmentation masks by converting connected wound regions into bounding boxes, and are supplemented with negative images from AZH. This yields 1,220 localization images: 810 for training (600 positive and 210 negative) and 400 for validation (386 positive and 14 negative).

3.4. Preprocessing and Augmentation

We consider three model scales throughout the benchmark: **MINI**, **BASE**, and **XL**. These variants differ in model capacity and are used to examine the effect of scaling across classification, segmentation, and localization tasks. More details are provided in the next section.

For classification, the **MINI** configuration operates on images resized to 224×224, whereas the **BASE** and **XL** configurations use 384×384 inputs. Training-time augmentation includes random horizontal flips ($p = 0.5$), random vertical flips ($p = 0.3$), random rotations within $\pm 20°$, color jitter (brightness $= 0.2$, contrast $= 0.2$, saturation $= 0.1$, hue $= 0.05$), and random erasing ($p = 0.1$). All images are normalized using ImageNet statistics.

For segmentation, images and masks are processed at 512×512 resolution. We apply the same spatial augmentations as in classification, with nearest-neighbor interpolation for masks to preserve discrete labels.

For localization, images are resized to 512×512 and augmented using mosaic augmentation and HSV perturbation. When bounding boxes are derived from segmentation masks, we use nearest-neighbor interpolation during resizing to maintain spatial consistency.

4. The WILLIE Framework

WILLIE is a unified framework for wound classification, segmentation, and localization. As demonstrated in Figure 3, given an input wound image $\mathbf{x}$, WILLIE produces three outputs: a wound-type prediction $\hat{y}_{\text{cls}}$, a segmentation mask $\hat{\mathbf{M}}$, and a localization result $\hat{\mathbf{B}}$ derived from the predicted mask. The framework is evaluated in three model scales:

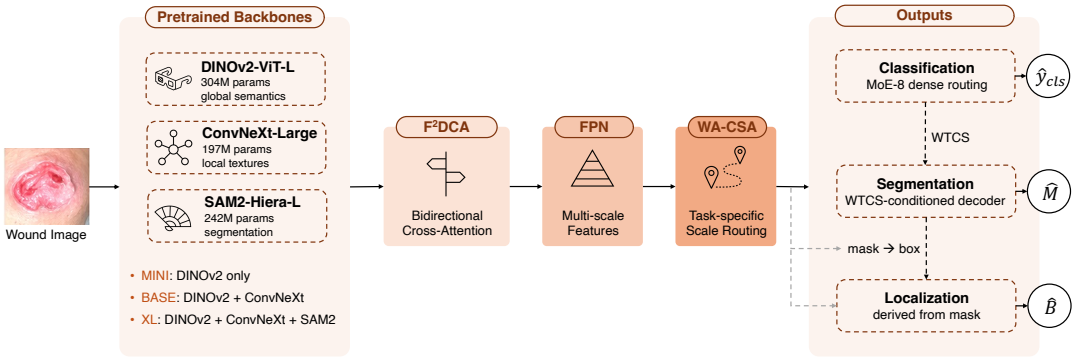

Figure 3: Overview of WILLIE architecture. A wound image is processed through shared pretrained backbones, cross-backbone fusion, multi-scale feature construction, and task-specific scale routing to produce three outputs: wound type classification, boundary segmentation, and localization.

MINI, **BASE**, and **XL**, to study how backbone capacity affects multi-task performance under a common benchmark. Our emphasis is not only on model design, but also on using a shared framework to evaluate when joint modeling is effective across clinically coupled wound-analysis tasks.

4.1. Model Scales and Backbone Configurations

We consider three model scales with increasing backbone capacity, summarized by Table 2.

Table 2: Model variant parameter statistics.

Variant	Backbone(s)	Total	Trainable	Frozen	% Trainable
MINI	DINOv2-S	34.3M	12.3M	22.1M	35.7
BASE	DINOv2-L + ConvNeXt-L	520.4M	19.8M	500.6M	3.8
XL	DINOv2-L + ConvNeXt-L + SAM2-Hiera-L	762.5M	360.2M	402.3M	47.3

WILLIE-MINI, 34.3M parameters. MINI uses a single DINOv2-Small (Oquab et al., 2024) (ViT-S/14) backbone with 22.06M frozen parameters and 12.26M trainable parameters in the task heads. It omits F^2DCA and uses a simplified feature pyramid to upsample the single-backbone output to multiple scales. This lightweight variant serves as the lowest-capacity point in our scaling analysis and a resource-efficient configuration.

WILLIE-BASE, 520.4M parameters. BASE uses two pretrained backbones: DINOv2-ViT-Large (304M) for global semantics and ConvNeXt-Large (Liu et al., 2022) (197M) for local texture and boundary cues. Only 19.8M parameters (3.8%) are trainable, including the F^2DCA layers, feature pyramid, task heads, and WTCS parameters. This design enables complementary feature extraction while limiting overfitting on a modest-sized benchmark.

WILLIE-XL, 762.5M parameters. XL adds SAM2-Hiera-Large (Ravi et al., 2024) (242M) as a third backbone, forming a triple-backbone architecture. DINOv2 and ConvNeXt remain frozen, while selected SAM2 decoder layers are fine-tuned, yielding 360.2M trainable parameters (47.3%). This variant tests whether a larger, segmentation-oriented backbone primarily benefits dense prediction rather than all wound-analysis tasks uniformly.

4.2. Shared Representation and Feature Fusion

WILLIE combines multi-backbone configurations through cross-backbone and cross-scale fusion. Let F_A and F_B denote feature maps from two backbones at a given scale. F^2DCA addresses the semantic gap between these representations through bidirectional cross-attention:

$$F'_A = F_A + \alpha \cdot \text{WoundGate}(F_A) \odot \text{CrossAttn}(F_A, F_B)$$

with an analogous update for F'_B. Here, $\text{WoundGate}(F) = \sigma(W_{\text{gate}} \cdot \text{GAP}(F))$ is a learned sigmoid gating function, where GAP denotes global average pooling. The gate suppresses cross-attention in non-wound regions and amplifies it where complementary features are most useful. The scaling factor α is initialized to 0.1 and typically converges to 0.15–0.25 during training, so that cross-attention corrections remain a bounded refinement rather than dominating the original backbone features.

For XL, F^2DCA is applied pairwise across all three backbone combinations at each scale. The resulting enhanced features are merged by learned weighted summation within the feature pyramid (Lin et al., 2017). At scale s, the fused pyramid feature is

$$\text{FPN}_s = \sum_b \alpha_b^s \cdot \text{Proj}_b^s(F_b^s), \qquad \sum_b \alpha_b^s = 1,$$

where $\text{Proj}_b^s(\cdot)$ denotes a 1×1 projection to the common pyramid dimension $d_{\text{fpn}} = 256$. The weights α_b^s are produced by a gating network,

$$\alpha^s = \text{softmax}\Big(W^s \cdot [\text{GAP}(F_1^s); \text{GAP}(F_2^s); \ldots]\Big).$$

This allows the framework to emphasize different backbones at different scales. Empirically, higher-resolution levels place greater weight on ConvNeXt for local boundary detail, lower-resolution levels place greater weight on DINOv2 for global wound-type structure, and intermediate levels place greater weight on SAM2 in the XL model.

To support task-specific use of the shared pyramid, WILLIE applies Wound-Aware Cross-Scale Attention (WA-CSA). Given a task query $\mathbf{q}$ and pyramid features $\mathbf{P} = \{P_1, P_2, P_3, P_4\}$, WA-CSA computes

$$\text{WA-CSA}(\mathbf{q}, \mathbf{P}) = \sum_j w_j(\mathbf{q}) \cdot \text{Attn}(\mathbf{q}, P_j),$$

where $w_j(\mathbf{q}) = \text{softmax}_j(\text{MLP}(\mathbf{q}))$ is a learned per-query, per-scale weighting. This mechanism allows wound boundary queries to emphasize higher-resolution levels and wound-type queries to emphasize lower-resolution levels.

4.3. Task Heads

Classification. The classification branch applies global pooling to the shared representation and predicts one of five wound classes using a mixture-of-experts classifier with eight expert subnetworks (MoE-8) (Jacobs et al., 1991). Per-variant expert counts are detailed in Appendix E. Let $\mathbf{h}_{\text{cls}}$ be the pooled classification feature; each expert is a two-layer MLP with hidden dimension 512:

$$\text{Expert}_k(\mathbf{h}_{\text{cls}}) = W_2^k \,\text{GELU}(W_1^k \mathbf{h}_{\text{cls}} + b_1^k) + b_2^k.$$

A router produces soft weights $\mathbf{w} = \text{softmax}(W_{\text{router}}\mathbf{h}_{\text{cls}})$, and the final classification logits are $\mathbf{o}_{\text{cls}} = \sum_{k=1}^{8} w_k \cdot \text{Expert}_k(\mathbf{h}_{\text{cls}})$. Then the classification result:

$$\hat{y}_{\text{cls}} = \arg\max \mathbf{o}_{\text{cls}}.$$

WILLIE uses dense routing, so all experts contribute to every prediction. An auxiliary load-balancing loss encourages more even expert utilization and reduces expert collapse.

Segmentation. The segmentation branch follows a U-Net decoder (Ronneberger et al., 2015) with projected spatial and channel squeeze-excitation (P-scSE) blocks for feature recalibration at each decoder level. To couple diagnostic context with boundary prediction, the decoder is modulated by Wound-Type Conditioned Segmentation (WTCS), implemented with feature-wise linear modulation (FiLM) (Perez et al., 2018). At decoder level ℓ, the penultimate classification feature vector $\mathbf{z}$ computes a per-channel scale and shift:

$$\text{WTCS}(F_\ell, \mathbf{z}) = \gamma_\ell \odot F_\ell + \beta_\ell,$$

where $\gamma_\ell = W_\gamma^\ell \mathbf{z} + b_\gamma^\ell$ and $\beta_\ell = W_\beta^\ell \mathbf{z} + b_\beta^\ell$. This conditioning allows wound-type information to influence mask prediction without requiring hard class assignments, while preserving gradient flow between the segmentation and classification branches. The modulated decoder features are then mapped to the predicted wound mask $\hat{\mathbf{M}}$:

$$\hat{\mathbf{M}} = \text{Decoder}_{\text{seg}}(\{\, \text{WTCS}(F_\ell, \mathbf{z}) \,\}_\ell).$$

Localization. Rather than training a separate detection head, WILLIE derives wound localization from the predicted segmentation mask. The segmentation-to-localization pipeline thresholds the predicted mask $\hat{\mathbf{M}}$ at 0.5, extracts connected components, computes one axis-aligned bounding box per component, and removes components smaller than 20 pixels, yielding the localization output $\hat{\mathbf{B}}$. Full construction details are provided in Appendix D. This produces localization as a derived output of segmentation rather than as an independently trained regression task.

Why localization is derived from segmentation. We treat localization as a derived output rather than a separately trained detection objective for both clinical and empirical reasons. In wound images, the target region is typically contiguous and defined by its extent and boundary, making segmentation a natural basis for localization. In preliminary experiments, a dedicated FCOS head (Tian et al., 2019) achieved only 6.6% mAP@0.5 within the shared architecture, whereas segmentation-derived localization performed substantially better. This design also avoids introducing a separate detection loss into the multi-task objective. Within the benchmark, localization is therefore evaluated as an output of the unified framework, while standard detection baselines are retained as external references.

4.4. Training Objective

WILLIE is trained with a composite multi-task loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{bal}}\mathcal{L}_{\text{balance}},$$

where $\mathcal{L}_{\text{cls}}$ is weighted cross-entropy computed from $\hat{y}_{\text{cls}}$, $\mathcal{L}_{\text{seg}}$ is the equally weighted sum of binary cross-entropy and Dice loss computed from $\hat{\mathbf{M}}$, $\mathcal{L}_{\text{seg}} = 0.5\,\mathcal{L}_{\text{BCE}} + 0.5\,\mathcal{L}_{\text{Dice}}$, and

$\mathcal{L}_{\text{balance}}$ is the MoE load-balancing regularizer. The task weights λ_{cls} and λ_{seg} are learnable parameters initialized to 1.0.

Since localization is derived from segmentation rather than trained as a separate head, there is no localization loss. For checkpoint selection, we use a combined validation score:

$$\text{Combined} = 0.4 \times \text{cls\_acc} + 0.4 \times \text{seg\_dice} + 0.2 \times \text{loc\_AP@0.5},$$

where loc_AP@0.5 is computed from the seg-to-loc pipeline on the validation set. This criterion reflects the central goal of the benchmark: evaluating unified performance across clinically coupled wound-analysis tasks rather than optimizing any one task in isolation.

5. Experiments

We evaluate WILLIE under a common benchmark protocol and compare it against established task-specific baselines for wound classification, segmentation, and localization.

5.1. Hypotheses

We posit the following hypotheses.

H1: A unified multi-task framework can match or exceed strong task-specific baselines while producing all three outputs in a single pipeline. **H1** will be evaluated by comparing WILLIE against task-specific baselines on classification accuracy, macro F1, and macro AUC-ROC; segmentation Dice and IoU; and localization AP@0.5.

H2: Localization derived from segmentation is more effective than training a separate detection head or relying solely on standard detection architectures. **H2** will be evaluated by comparing WILLIE's segmentation-derived localization performance against baselines, and against preliminary experiments with a dedicated FCOS detection head.

H3: The benefits of scaling are task-dependent: larger and more specialized pretrained backbones provide larger gains for segmentation and localization than for classification. **H3** will be evaluated by comparing task-specific performance across the MINI, BASE, and XL variants of WILLIE using classification, segmentation, and localization metrics.

H4: Joint modeling is most effective for task combinations with compatible objectives; in particular, classification and segmentation benefit from shared representations, whereas localization is better treated as a derived output from segmentation than as a third jointly optimized regression task. **H4** will be evaluated by comparing unified and task-specific performance, and by examining whether segmentation-derived localization is more effective than adding a separately optimized detection objective within the shared framework.

5.2. Baselines

We compare WILLIE against established baselines in each task, all trained under our unified 5-class protocol with identical data splits and evaluation to ensure fair comparison.

Classification:

- **VGG-19** (124.9M) is included following Anisuzzaman et al. (2022).
- **ResNet-50** (24.6M) serves as a standard CNN baseline.
- **EfficientNet-B4** (18.5M) provides a stronger CNN baseline with recent EfficientNet-based classifiers (Aldoulah et al., 2023; Ullah et al., 2025).
- **DINOv2 + Linear** (22.2M) uses a frozen DINOv2 with a linear head, isolating the contribution of self-supervised pretrained features without multi-task modeling.

Segmentation:

- **U-Net with EfficientNet-B3 encoder** (13.2M parameters) provides a standard medical-image segmentation baseline (Ronneberger et al., 2015).
- **MedSAM (ViT-B)** (93.7M total, 4.1M trainable) represents a medical adaptation of SAM using prompt-free inference (Ma et al., 2024).
- **SAM2.1 (Hiera-S)** (46.1M total, 11.7M trainable) represents a more recent segmentation foundation model (Ravi et al., 2024).

We compare SAM2.1 and MedSAM using a Wilcoxon signed-rank test, with $p < 0.001$, indicating a statistically significant advantage for SAM2.1 on this benchmark.

Localization:

- **RT-DETR-L** (32.8M parameters) provides a transformer-based localization baseline on bounding boxes derived from FUSeg masks (Zhao et al., 2024) .
- **YOLOv8m** (25.9M parameters) provides a strong real-time detection baseline trained under the same conditions (Jocher et al., 2023).

5.3. Training Protocol

All WILLIE variants are trained using 5-fold stratified cross-validation on the 1,080-image training and validation pool, with approximately 864 images for training and 216 for validation in each fold. The official 234-image AZH test split is held out across all folds and used only for final classification evaluation.

For the optimizer, we use AdamW with a weight decay of 0.01. Learning rates are set to 1.5×10^{-5} for fine-tuned backbone components and 1×10^{-4} for task-specific heads. Training uses cosine annealing with warm restarts ($T_0 = 10$, $T_{\text{mult}} = 1$), for a maximum of 50 epochs, with early stopping based on the combined validation metric (patience = 12).

Batch size is 8 for MINI and 16 for BASE and XL. Mixed-precision training is used throughout. We additionally apply exponential moving average model averaging with decay 0.999. All experiments are run on NVIDIA Tesla V100-PCIE-32GB GPUs. Latency and peak-memory characterization is reported in Appendix B.

5.4. Evaluation Metrics

We report task-specific and unified metrics chosen to test the hypotheses above.

Classification. We report accuracy, macro F1, and macro AUC-ROC using a one-vs-rest formulation. These metrics are used to evaluate **H1**. Macro F1 and macro AUC-ROC are particularly important because the unified 5-class setting is moderately imbalanced.

Table 3: Classification comparison on the held-out AZH test set ($n = 234$).

Model	Type	Params	Acc (%)	F1 (%)	AUC
VGG-19	Baseline	124.9M	78.63	78.14	0.956
EfficientNet-B4	Baseline	18.5M	83.76	83.36	0.975
DINOv2+Linear	Ablation	22.2M	86.75	87.01	0.986
WILLIE-MINI (ens)	Unified	34.3M	88.90	87.60	0.976
ResNet-50	Baseline	24.6M	88.03	87.82	0.973
WILLIE-XL (TTA)	Unified	762.5M	**91.88**	90.73	0.986
WILLIE-BASE (TTA)	Unified	520.4M	**91.88**	**91.14**	**0.992**

Table 4: Segmentation comparison on FUSeg validation ($n = 400$).

Model	Type	Params	Mean Dice	Median Dice
U-Net (Eff-B3)	Baseline	13.2M	84.59	91.20
WILLIE-MINI	Unified	34.3M	81.80	87.50
WILLIE-BASE	Unified	520.4M	86.36	92.89
MedSAM (ViT-B)	Baseline	93.7M	90.93	93.33
WILLIE-XL	Unified	762.5M	91.41	93.50
SAM2.1 (Hiera-S)	Baseline	46.1M	**91.74**	**93.76**

Segmentation. We report Dice coefficient and intersection-over-union (IoU), averaged across images. These metrics are used to evaluate both **H1** and **H3**. They test whether unified modeling preserves strong boundary delineation relative to task-specific segmentation baselines, and whether segmentation continues to benefit from larger and more specialized pretrained backbones.

Localization. We report AP@0.5. For baselines, AP@0.5 is computed from detector outputs. For WILLIE, AP@0.5 is computed from the segmentation-to-localization pipeline. This metric is used to evaluate **H1** and **H2** and to test whether segmentation-derived localization is an effective alternative to dedicated detection.

Unified evaluation. To assess overall multi-task performance, we use the combined validation score defined in Section 4.4.

5.5. Results

We report results using the evaluation protocol described above.

Result 1: Unified approaches outperform or perform comparably relative to task-specific baselines. Table 3 shows that the unified WILLIE models achieve the strongest classification performance. BASE and XL attain the highest accuracy, and BASE achieves the strongest overall macro-F1 and AUC, indicating that unified modeling does not compromise diagnostic classification performance. Notably, even the lightweight MINI remains competitive with the strongest baseline. These results support H1 for classification.

Table 4 shows that segmentation performance improves substantially with model scale and backbone specialization. XL achieves segmentation performance comparable to the strongest task-specific baselines, exceeding both U-Net and MedSAM and approaching SAM2.1. In contrast, the lower-capacity unified variants lag behind the strongest segmentation specialists, suggesting that dense wound delineation places greater demands on representation quality than classification. These results support H1 for segmentation.

Table 5: Localization comparison on FUSeg validation ($n = 400$).

Model	Type	Params	AP@0.5	Prec.	Recall	F1
RT-DETR-L	Dedicated	32.8M	87.95	88.15	89.38	88.76
YOLOv8m	Dedicated	25.9M	91.22	87.45	88.47	87.96
WILLIE-MINI s→l	Unified	34.3M	84.60	89.03	88.66	88.84
WILLIE-BASE s→l	Unified	520.4M	89.91	92.40	95.40	93.87
WILLIE-XL s→l	Unified	762.5M	**96.23**	**96.23**	**94.64**	**95.43**

s→l: segmentation-to-localization pipeline.

Table 5 shows that segmentation-derived localization is highly effective. XL achieves the strongest localization performance, surpassing YOLOv8m and RT-DETR-L, and BASE remains competitive with dedicated detectors despite not optimizing a separate box-regression objective. These results support H1 for localization and directly support H2.

Result 2: Scaling benefits are strongly task-dependent, with the largest gains observed for segmentation and localization. Figure 4 shows that scaling behavior is strongly task-dependent. Classification improves from MINI to BASE but does not improve further at XL, indicating early saturation for wound-type recognition under the current benchmark. By contrast, segmentation and localization continue to improve across all three configurations, with the largest gains observed for the dense prediction tasks. This pattern suggests that increased model capacity and backbone specialization provide their clearest benefit for segmentation and segmentation-derived localization, whereas classification appears to benefit primarily from the transition from the lightweight MINI configuration to the stronger BASE configuration. These results support H3.

Result 3: Unified modeling is most effective when task objectives are compatible, with localization best handled as a derived output of segmentation. The strongest localization results are achieved when localization is treated as a derived output of segmentation rather than as a separately optimized detection objective, suggesting that these tasks are not equally compatible under joint optimization. In particular, classification and segmentation appear to benefit from shared learning, whereas localization framed as box regression is better handled post hoc in this benchmark. This result supports H4.

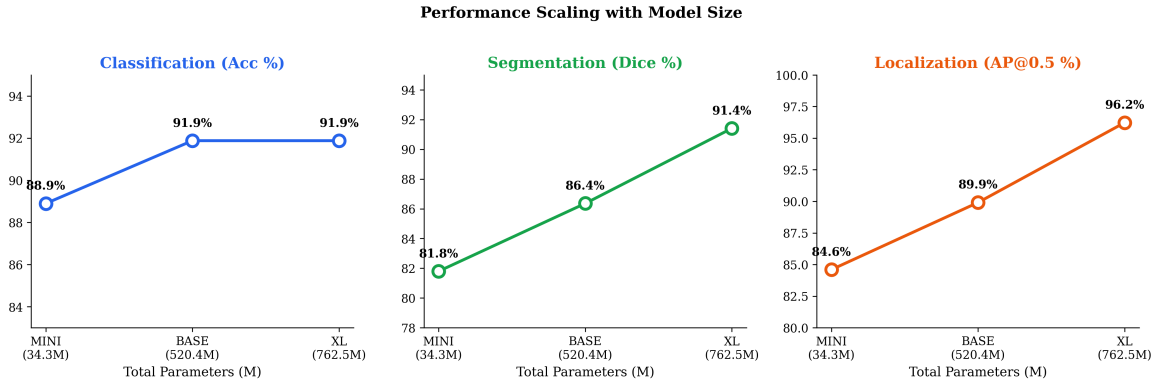

Figure 4: Task-dependent scaling behavior across WILLIE variants. Classification accuracy improves from MINI to BASE and then saturates at XL, whereas segmentation Dice and localization AP@0.5 continue to improve with model scale.

Table 6: Per-class results for WILLIE-BASE (top-3 TTA).

Class	Prec.	Recall	F1	Support	Train
Diabetic	97.3	78.3	86.7	46	220
Pressure	76.5	76.5	76.5	34	270
Surgical	88.9	95.2	92.0	42	122
Venous	91.0	98.4	94.6	62	318
No Wound	98.0	100.0	**99.0**	50	150

Result 4: Residual classification error is concentrated in clinically challenging wound categories rather than uniformly distributed across classes. Table 6 shows that classification performance is strongest for no wound, venous, and surgical categories, while pressure wounds remain the most challenging class despite substantial training representation. The dominant residual error pattern is confusion between pressure and venous wounds, localizing the primary source of remaining classification error within the benchmark. This pattern suggests that some residual error may reflect genuine clinical ambiguity in image-only wound assessment rather than uniformly distributed model weakness. Bootstrap 95% confidence intervals for per-class F1 are reported in Appendix C, and a source-level analysis of the pressure–venous confusion is provided in Appendix F.

6. Discussion

Across tasks, WILLIE remained competitive with strong task-specific baselines while revealing broader patterns about task formulation, scaling, and residual error. These findings are relevant not only to wound analysis, but also to healthcare imaging settings in which multiple clinically coupled outputs must be generated from the same input.

6.1. Key Findings

Finding 1: Segmentation-derived localization is well suited to wound analysis. One of our clearest findings is that localization derived from segmentation can outperform dedicated detection baselines in wound images. WILLIE-XL achieved the strongest localization performance overall, exceeding YOLOv8m and RT-DETR-L, whereas a preliminary FCOS head performed poorly within the shared framework. This is consistent with the clinical structure of wounds: unlike many generic detection targets, wounds are contiguous regions whose extent and boundary are central to interpretation. More broadly, this result suggests that for healthcare imaging tasks in which the target is spatially coherent and clinically defined by region extent, localization may be more effective as a derived output of segmentation than as an independently optimized objective.

Finding 2: Scaling benefits are strongly task-dependent. A second finding is that scaling does not benefit all wound-analysis tasks equally. Classification improved from MINI to BASE but then saturated at XL, whereas segmentation and localization continued to improve at the largest scale. This indicates that model capacity and backbone specialization matter differently across tasks, even within a single clinical workflow. For classification, the combination of DINOv2 and ConvNeXt in BASE appears sufficient to capture most of

the image's discriminative information. For segmentation and localization, the additional features introduced by SAM2 continue to provide measurable gains. More generally, these results suggest that scaling decisions in healthcare imaging should be guided by task requirements rather than by the assumption that larger models improve all outputs uniformly.

Finding 3: Unified performance depends on task compatibility. The results further suggest that strong unified performance depends not only on feature sharing, but also on whether task objectives remain compatible within a shared architecture. In WILLIE, classification and segmentation can be learned jointly while preserving strong performance on both tasks. Localization is also effective within the unified framework, but only when treated as a derived output of segmentation; when framed as a separately optimized detection objective, performance deteriorates sharply. This pattern highlights an important principle for healthcare multi-task learning: clinically related tasks are not necessarily equally compatible during optimization. A unified model may succeed when tasks share useful structure, but degrade when one objective is poorly aligned with the others.

Finding 4: Residual error reflects clinically meaningful ambiguity. Although WILLIE achieves strong overall classification performance, pressure wounds remain the most difficult class, and confusion between pressure and venous wounds is the dominant residual error pattern. This is clinically plausible, as visual appearance alone may not always provide enough information to distinguish etiology when morphology overlaps or when images lack contextual information such as anatomical site, patient history, vascular assessment, or temporal progression. This observation highlights an important limit of image-only prediction. In healthcare imaging, some residual error may reflect ambiguity in the modality itself rather than insufficient model capacity alone. For wound analysis, this suggests that future systems may benefit from integrating images with metadata, longitudinal evidence, or clinician-in-the-loop escalation rather than treating image-only classification as fully self-sufficient.

6.2. Limitations

This study has several limitations. First, the benchmark is constructed from a relatively small number of public wound datasets, and the resulting taxonomy remains coarse relative to the heterogeneity of wound care in practice. Second, segmentation evaluation is limited to the public FUSeg training/validation release because official test-set labels are not available. Third, localization is derived from segmentation masks and assumes connected wound regions, which may be less appropriate for images containing multiple disjoint wounds or more complex spatial structure. Fourth, although the framework contains several architectural components, we provide a per-component ablation in Appendix A rather than a full factorial ablation of every module. Finally, we do not provide prospective clinical validation, workflow evaluation, or latency characterization for real-world deployment.

6.3. Future Work

Several directions follow naturally from this work. One priority is to extend the benchmark with larger and more diverse wound datasets, including broader wound categories, richer metadata, and images collected across institutions, devices, and care settings. A second is to

evaluate unified wound analysis prospectively in realistic clinical workflows, including how model outputs support longitudinal monitoring, triage, and specialist escalation. A third is to study more explicitly which task combinations are compatible in shared healthcare imaging systems, including whether compatibility can be predicted or improved through alternative objective design. Finally, the lightweight MINI configuration suggests a possible path toward lower-resource deployment, but this requires dedicated study of efficiency, robustness, and clinical usability beyond retrospective benchmark evaluation.

7. Conclusion

We presented WILLIE, a unified framework and benchmark for wound classification, segmentation, and localization that enables systematic study of multi-task wound analysis under a shared, clinically grounded protocol. We show that a single framework can achieve strong performance across all three tasks while producing unified outputs in one pipeline. Beyond aggregate performance, the benchmark yields broader insights for healthcare imaging: segmentation-derived localization outperforms dedicated detection baselines in this setting, scaling benefits are strongly task-dependent, and successful unified learning depends not only on shared representations but also on compatibility among task formulations.

References

Z. A. Aldoulah, H. Malik, and R. Molyet. A novel fused multi-class deep learning approach for chronic wounds classification. *Applied Sciences*, 13(21):11630, 2023.

Mustafa Alhababi, Gregory Auner, Hafiz Malik, Muteb Aljasem, and Zaid Aldoulah. Unified wound diagnostic framework for wound segmentation and classification. *Machine Learning with Applications*, 19:100616, 2025.

D. M. Anisuzzaman, Chuanbo Wang, Behrouz Rostami, Sandeep Gopalakrishnan, Jeffrey Niezgoda, and Zeyun Yu. Image-based artificial intelligence in wound assessment: A systematic review. *Advances in Wound Care*, 11(12):687–709, 2022.

David G Armstrong, Matilde A Swerdlow, Alexandria A Armstrong, Michael S Conte, William V Padula, and Sicco A Bus. Five-year mortality and direct costs of care for people with diabetic foot complications are comparable to cancer. *Journal of Vascular Surgery*, 71(3):1055–1061, 2020.

Dan Berlowitz, Luisa VanDeusen Lukas, Victoria Parker, et al. Preventing pressure ulcers in hospitals: A toolkit for improving quality of care. Technical report, Agency for Healthcare Research and Quality, Rockville, MD, 2011. AHRQ Publication No. 11-0087-EF.

Andrew JM Boulton, Loretta Vileikyte, Gunnel Ragnarson-Tennvall, and Jan Apelqvist. The global burden of diabetic foot disease. *The Lancet*, 366(9498):1719–1724, 2005.

Marissa J Carter, Joan DaVanzo, Randall Haught, et al. Chronic wound prevalence and the associated cost of treatment in Medicare beneficiaries: Changes between 2014 and 2019. *Journal of Medical Economics*, 26(1):894–901, 2023.

Zhao Chen et al. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, pages 794–803, 2018.

T. Dhar et al. FUSegNet: A deep convolutional neural network for foot ulcer segmentation. *Biomedical Signal Processing and Control*, 92:106057, 2024.

Andre Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In *NeurIPS*, pages 27503–27516, 2021.

Robert G Frykberg and James Banks. Challenges in the treatment of chronic wounds. *Advances in Wound Care*, 4(9):560–582, 2015.

Lisa Gould, Pam Abadir, Harold Brem, et al. Chronic wound repair and healing in older adults: current status and future research. *Journal of the American Geriatrics Society*, 63(3):427–438, 2015.

Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *ECCV*, pages 270–287, 2018.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. GitHub, https://github.com/ultralytics/ultralytics, 2023.

David H Keast, Charles K Bowering, A Burrows Evans, Gary L Mackean, Catherine Burrows, and Loretta D'Souza. MEASURE: A proposed assessment framework for developing best practice recommendations for wound assessment. *Wound Repair and Regeneration*, 12(s1):S1–S17, 2004.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018.

Moses Kiprono. WoundNet-Ensemble: A novel IoMT system integrating self-supervised deep learning and multi-model fusion for wound classification. *arXiv preprint arXiv:2512.18528*, 2025.

Alexander Kirillov et al. Segment anything. In *ICCV*, pages 4015–4026, 2023.

Lawrence A Lavery, David G Armstrong, Robert P Wunderlich, Michael J Mohler, Courtney S Wendel, and Brent A Lipsky. Risk factors for foot infections in individuals with diabetes. *Diabetes Care*, 29(6):1288–1293, 2006.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

Zhuang Liu et al. A ConvNet for the 2020s. In *CVPR*, pages 11976–11986, 2022.

Jun Ma et al. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

Lorainne Martinengo, Mervin Olsson, Ram Bajpai, et al. Prevalence of chronic wounds in the general population: systematic review and meta-analysis of observational studies. *Annals of Epidemiology*, 29:8–15, 2019.

Medetec. Medetec wound database. http://www.medetec.co.uk/files/medetec-image-databases.html, 2010. Accessed: 2025.

Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. Deep learning with mixed supervision for brain tumor segmentation. *Journal of Medical Imaging*, 6(3): 034002, 2019.

Samuel R Nussbaum, Marissa J Carter, Caroline E Fife, Joan DaVanzo, Randall Haught, Marcia Nusgart, and Donna Cartwright. An economic evaluation of the impact, cost, and Medicare policy implications of chronic nonhealing wounds. *Value in Health*, 21(1): 27–32, 2018.

Mervin Olsson, Krister Järbrink, Udhaya Divakar, et al. The humanistic and economic burden of chronic wounds: A systematic review. *Wound Repair and Regeneration*, 27(1): 114–125, 2019.

Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William V Padula and Becky A Delarmente. The national cost of hospital-acquired pressure injuries in the United States. *International Wound Journal*, 16(3):634–640, 2019.

Yash Patel, Tirth Shah, Mrinal Kanti Dhar, Tianyu Zhang, Jeffrey Niezgoda, Sandeep Gopalakrishnan, and Zeyun Yu. Integrated image and location analysis for wound classification: A deep learning approach. *Scientific Reports*, 14:7043, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI*, pages 3942–3951, 2018.

Nikhila Ravi et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

Chandan K. Sen. Human wound and its burden: Updated 2020 compendium of estimates. *Advances in Wound Care*, 10(5):281–292, 2021.

R Gary Sibbald, James A Elliott, Elizabeth A Ayello, and Ranjani Somayaji. Optimizing the moisture management tightrope with wound bed preparation 2021. *Advances in Skin and Wound Care*, 34(2):58–68, 2021.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *ICML*, pages 9120–9132, 2020.

Terry Swanson, Karen Ousey, Emily Haesler, et al. IWII wound infection in clinical practice consensus document: 2022 update. *Journal of Wound Care*, 31(Sup12):S10–S21, 2022.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019.

Sifat Ullah, Ali Javed, Muteb Aljasem, and Abdul Khader Jilani Saudagar. Eff-ReLU-Net: A deep learning framework for multiclass wound classification. *BMC Medical Imaging*, 25:257, 2025.

Chuanbo Wang et al. FUSeg: The foot ulcer segmentation challenge. *arXiv preprint arXiv:2201.00414*, 2022.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, pages 5824–5836, 2020.

Yian Zhao et al. DETRs beat YOLOs on real-time object detection. In *CVPR*, 2024.

Appendix A. Component Ablation Study

We ablate each component (F^2DCA, WA-CSA, MoE, WTCS) individually by retraining under the same 5-fold cross-validation protocol used in the main paper and evaluating on the validation folds. We report classification accuracy, segmentation Dice, and localization AP@0.5 (mask-derived). No test-time augmentation is used, so absolute numbers are lower than the TTA test-set results in Tables 3–5; this ablation isolates the *relative* contribution of each component. MINI omits F^2DCA by design, as it uses a single backbone.

Table 7: Components ablation: classification accuracy (%). 5-fold CV, validation folds, no TTA.

Configuration	MINI	BASE	XL
Full	86.6	88.2	89.8
– F^2DCA	–	87.8	89.9
– WA-CSA	86.6	86.9	90.2
– MoE	86.7	87.6	89.8
– WTCS	86.1	87.0	89.5

Table 8: Component ablation: segmentation Dice (%). 5-fold CV, validation folds, no TTA.

Configuration	MINI	BASE	XL
Full	82.3	84.1	83.2
– F^2DCA	–	84.9	83.6
– WA-CSA	82.9	84.7	84.0
– MoE	83.5	83.8	82.1
– WTCS	83.6	84.9	84.0

Table 9: Component ablation: localization AP@0.5 (%). 5-fold CV, validation folds, no TTA.

Configuration	MINI	BASE	XL
Full	84.1	85.6	85.0
– F^2DCA	–	86.8	85.7
– WA-CSA	85.5	86.7	85.2
– MoE	85.3	85.6	85.0
– WTCS	86.1	87.6	85.0

Across all three scales and all three tasks, removing any single component keeps performance within cross-validation noise. The pretrained backbones and multi-scale feature processing account for most of the performance. This is consistent with the paper's stated

contribution (Section 6.2): a unified multi-task framework and benchmark evaluated under a common protocol, rather than a demonstration that each internal module is individually essential.

Appendix B. Latency and Compute

The WILLIE-XL model performs a single forward pass at approximately 88 ms per image with 9.3 GB peak GPU memory on an NVIDIA V100. Parameter counts for all variants are reported in Table 2 (MINI 34.3M, BASE 520.4M, XL 762.5M). Test-time augmentation, used for the headline results, multiplies inference cost by the number of augmented views and is an evaluation choice rather than a deployment requirement.

Appendix C. Per-Class Confidence Intervals

To quantify small-sample uncertainty on the held-out AZH classification test set ($n = 234$), we report bootstrap 95% confidence intervals for per-class F1. Small classes carry wide intervals: the pressure class ($n = 34$) has F1 76.5 with 95% CI $[63.5, 86.7]$, while well-supported classes are tight (no-wound F1 99.0, 95% CI $[96.6, 100.0]$). These intervals contextualize the per-class results in Table 6.

Appendix D. Localization Construction

Localization boxes are derived from the predicted segmentation mask: the mask is thresholded at 0.5, connected components are extracted, components smaller than 20 pixels are removed, and one axis-aligned bounding box is computed per remaining component (Section 4.3). All localization methods, including the YOLOv8 and RT-DETR baselines, are evaluated against the same mask-derived ground-truth boxes. The finding that segmentation-derived localization outperforms dedicated detectors is therefore benchmark-level evidence for spatially coherent wound targets, not a general claim that detection is unnecessary.

Appendix E. Per-Variant Expert Configuration

The mixture-of-experts classification head uses eight experts with dense routing in the WILLIE-XL configuration described in Section 4.3. The number of experts is matched to backbone capacity across the smaller variants. The "MoE-8" designation refers specifically to the XL configuration.

Appendix F. Pressure–Venous Confusion

The dominant residual classification error is confusion between pressure and venous wounds (Section 6, Finding 4). Beyond clinical ambiguity, a contributing factor is data source: pressure images are drawn largely from Medetec while venous images are drawn largely from AZH, so part of the observed confusion may reflect differences in image source in addition to clinical similarity between the two wound types.