MEDIPPD: VISION-LANGUAGE METROLOGY MEETS DERMAL INTERPRETATION IN PURIFIED PROTEIN DERIVATIVE SKIN TESTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated Purified Protein Derivative (PPD) skin-test assessment requires both millimeter-scale quantification of clinician-measured induration and localized interpretation of clinically relevant dermal findings. However, these objectives present two challenges: (1) *Metric Fidelity Gap*, where semantic representations may lose the geometric and scale information required for accurate physical measurement; (2) *Spatial Grounding Failure*, where correct semantic predictions may lack localized evidence supporting the identified dermal findings. To systematically examine these challenges, we establish a unified benchmark that links clinician-defined induration to calibrated physical measurement and dermal-sign predictions to localized evidence, enabling joint evaluation of metric fidelity and clinically meaningful grounding. To address the distinct representational requirements revealed by this benchmark, we propose **MediPPD**: Vision-Language Metrology Meets Dermal Interpretation in Purified Protein Derivative Skin Tests, which adopts *selective separation with controlled interaction* to preserve geometry-sensitive physical metrology while allowing calibrated geometry to support semantic reasoning, thereby enabling accurate measurement together with spatially grounded dermal interpretation. Extensive experiments demonstrate that MediPPD consistently improves physical measurement and grounded dermal interpretation over competitive visual and multimodal methods. Our code and dataset are available at `https://anonymous.4open.science/r/MediPPD-anony`.

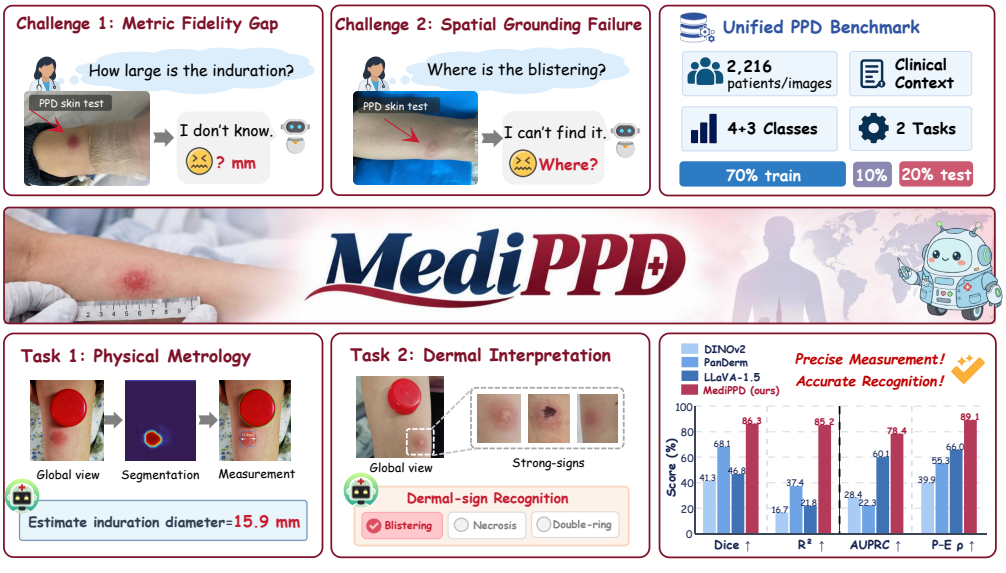


Figure 1: Overview of MediPPD. A unified PPD benchmark exposes the Metric Fidelity Gap and Spatial Grounding Failure, motivating two complementary tasks: physical metrology and dermal interpretation. MediPPD jointly achieves precise measurement and accurate dermal-sign recognition.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

The tuberculin skin test (TST), commonly performed with Purified Protein Derivative (PPD), remains an important tool for assessing tuberculosis infection (Organization, 2025; Konstantinos, 2010). Its interpretation is *quantitative* and *spatial*, requiring measurement of palpable induration and inspection of localized findings such as blistering, necrosis, and double-ring reaction. Since our annotations directly delineate the clinician-defined induration region, automated PPD assessment must recover its physical extent and the spatial evidence for dermal interpretation.

Existing digital TST systems mainly focus on image-based measurement using smartphone imaging, photogrammetry, or physical-reference calibration (Dendere et al., 2017; Naraghi et al., 2018; Akinola et al., 2024), while recent vision–language models (VLMs) show strong biomedical recognition and grounding capabilities (Radford et al., 2021; Li et al., 2022; 2023; Zhang et al., 2025). These methods target different objectives: measurement approaches emphasize geometric calibration and physical extent, whereas semantic models prioritize recognition and high-level abstraction. Unifying them therefore requires reconciling these distinct representational requirements.

As illustrated in Fig. 1, this setting presents two distinct challenges. The first is the *Metric Fidelity Gap*, where semantic representations may lose the geometric and scale information required for accurate physical measurement. Millimeter-scale assessment depends on preserving the spatial correspondence between the physical reference and the clinician-defined induration region, whereas semantic abstraction may weaken precisely this information. The second is *Spatial Grounding Failure*, where correct semantic predictions may lack localized evidence supporting the identified dermal findings. A model may correctly recognize a dermal finding while relying on global morphology or contextual correlations rather than the region that actually supports the prediction. These challenges reveal distinct representational requirements for physical metrology and dermal interpretation, motivating selective separation while retaining controlled information exchange.

To systematically examine these challenges, we establish a unified benchmark that links clinician-defined induration to calibrated physical measurement and dermal-sign predictions to localized evidence, enabling joint evaluation of metric fidelity and clinically meaningful grounding. Building on this benchmark, we propose **MediPPD**: Vision-Language Metrology Meets Dermal Interpretation in Purified Protein Derivative Skin Tests. MediPPD adopts *selective separation with controlled interaction* to preserve geometry-sensitive physical metrology while allowing calibrated geometry to support semantic reasoning. Specifically, Physical Scale Metrology maintains explicit physical calibration and induration geometry for millimeter-scale estimation, while Multiview Dermal Grounding derives localized semantic evidence from complementary visual views. This evidence is subsequently integrated with task-relevant clinical context through Task-Routed Clinical Fusion and used by Semantic Mask Refinement to improve induration segmentation. Importantly, semantic information is restricted from altering the calibrated diameter estimator, preserving metric fidelity while supporting spatially grounded dermal interpretation. Extensive experiments on the proposed benchmark demonstrate that MediPPD consistently improves physical measurement and grounded dermal interpretation over strong visual and multimodal methods.

Our contributions are summarized as follows:

- **Problem innovation:** We formulate automated PPD assessment as a joint visual-metrology and grounded dermal-interpretation problem, identifying the *Metric Fidelity Gap* and *Spatial Grounding Failure* as two central challenges.
- **Unified benchmark:** We establish a benchmark that links clinician-defined induration to calibrated physical measurement and dermal-sign predictions to localized evidence, enabling joint evaluation of metric fidelity and clinically meaningful grounding.
- **Novel paradigm:** We introduce MediPPD, a unified vision–language framework based on *selective separation with controlled interaction*, preserving geometry-sensitive metrology while enabling calibrated geometry to support semantic reasoning.
- **Comprehensive validation:** Extensive experiments demonstrate consistent improvements over strong visual and multimodal baselines across physical measurement, dermal-sign recognition, and spatial grounding.

2 Related Work

TST Assessment and Metrology. Deep learning has advanced medical image recognition, localization, and segmentation, including dermatological analysis (Esteva et al., 2017), yet automated tuberculin skin test (TST) assessment remains underexplored. Existing systems mainly combine image-based localization or reconstruction with physical measurement using smartphone photogrammetry, 3D reconstruction, semantic segmentation, and physical references (Dendere et al., 2017; Naraghi et al., 2018; Akinola et al., 2024). More broadly, physical metrology estimates real-world dimensions through calibration, geometric priors, learned scale representations, or known object geometry (Hold-Geoffroy et al., 2018; Ponimatkin et al., 2022; Lee et al., 2023). In PPD imaging, a nearby reference object enables explicit pixel-to-millimeter calibration, but existing systems largely focus on measurement rather than localized dermal interpretation. MediPPD preserves calibrated metrology while separating geometry-sensitive measurement from semantics-sensitive reasoning.

Medical Vision–Language Grounding. Medical vision–language learning increasingly links visual findings with clinical concepts through multimodal alignment and reasoning (Huang et al., 2021; Wu et al., 2023; Li et al., 2023; Moor et al., 2023; Zhang et al., 2025), with recent work emphasizing fine-grained localization (Farooq et al., 2025; Yan et al., 2025a). However, semantic correctness does not guarantee faithful grounding, and recent studies report substantial localization limitations in medical MLLMs (Liu et al., 2026; Nützel et al., 2025). MediPPD addresses this gap using language-conditioned semantic priors and global–local alignment, with grounding directly supervised by spatial annotations and integrated into dermal interpretation.

Multi-Task Representation Learning. Multi-task learning (MTL) balances shared and task-specific representations through feature sharing, routing, attention, and optimization (Misra et al., 2016; Liu et al., 2019; Kendall et al., 2018; Yu et al., 2020), with recent work exploring adaptive specialization and robust optimization (Baek et al., 2025; Qin et al., 2025; Chai et al., 2025). Our setting differs in that the tasks require different information to remain explicit: physical metrology depends on spatial correspondence, calibration, and absolute scale, whereas dermal interpretation benefits from semantic abstraction and appearance invariance. MediPPD therefore adopts *selective separation with controlled interaction*, preserving geometry-sensitive and semantics-sensitive representations while exchanging only complementary information.

3 Benchmark

3.1 Task Formulation

We formulate automated PPD assessment as a structured multi-output problem. Given a PPD image I with a physical reference object and target-independent clinical context $\mathbf{c}$, the model predicts:

$$\{\hat{M}^{\mathrm{ind}}, \hat{d}^{\mathrm{ind}}, \hat{\mathbf{y}}, \hat{\mathcal{G}}\} = f_\theta(I, \mathbf{c}), \tag{1}$$

where $\hat{M}^{\mathrm{ind}}$ is the clinician-defined induration mask, $\hat{d}^{\mathrm{ind}}$ its transverse diameter in millimeters, $\hat{\mathbf{y}}$ the dermal-sign predictions, and $\hat{\mathcal{G}}$ their grounded evidence regions.

The benchmark evaluates four capabilities: induration segmentation, physical metrology, dermal-sign recognition, and spatial grounding. Segmentation and metrology target the same clinician-defined induration, providing complementary spatial and physical-scale supervision.

3.2 Benchmark Protocol

The benchmark contains 2,216 PPD images from 2,216 distinct patients, with one image per patient. We use a fixed, patient-disjoint split of 1,551/221/444 cases for training, validation, and held-out testing (approximately 70/10/20), stratified to approximately preserve rare dermal-sign prevalence and the distribution of clinician-measured induration diameter. The training set is used for parameter optimization, while the validation set is used for checkpoint selection, early stopping, hyperparameter tuning, routing, evidence aggregation, mask refinement, and clinical-context configuration. All settings are fixed before test evaluation; the held-out test set is used only for final reporting across MediPPD, baselines, and ablations. Each image is annotated with a clinician-defined induration region, transverse induration diameter, case-level dermal findings, and localized evidence

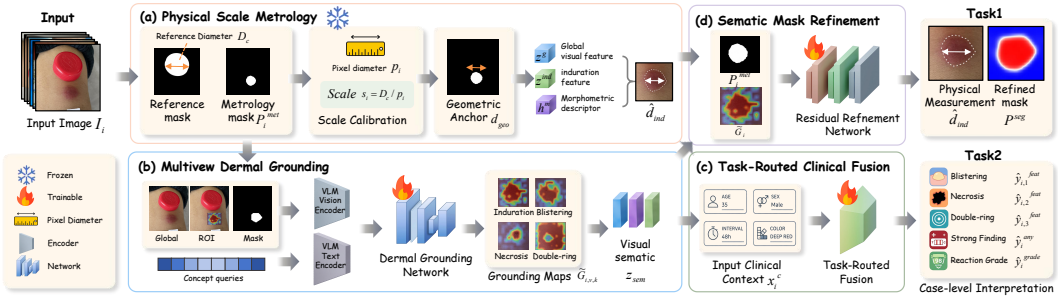


Figure 2: Overview of MediPPD, including physical scale metrology, multiview dermal grounding, task-routed clinical fusion and semantic mask refinement. MediPPD separates physical measurement from semantic reasoning while allowing calibrated geometry to support dermal interpretation.

regions. Clinical inputs exclude target-derived information, including measured diameter, target dermal labels, final clinical assessment outcomes. Detailed dataset statistics, annotation protocols, and clinical-information leakage controls are provided in Appendix A.

4 MediPPD

In this section, we briefly introduce the proposed MediPPD, including *Physical Scale Metrology*, *Multiview Dermal Grounding*, *Task-Routed Clinical Fusion* and *Semantic Mask Refinement*, as illustrated in Fig. 2. The detailed methodology is provided in Appendix B.

4.1 Physical Scale Metrology

For the i-th PPD image I_i, a geometry-sensitive network predicts a reference mask P_i^c and a *metrology induration mask* P_i^{met}. Let A_i^c and A_i^{met} denote their soft pixel areas, and let D_c be the known real-world diameter of the circular reference object. The reference pixel diameter $d_{i,c}^{\text{px}}$ and image-specific pixel-to-millimeter scale s_i are

$$d_{i,c}^{\text{px}} = \sqrt{\frac{4A_i^c}{\pi}}, \qquad s_i = \frac{D_c}{d_{i,c}^{\text{px}}}, \tag{2}$$

where s_i is measured in mm/pixel. Applying this scale to the predicted induration gives the calibrated geometric anchor

$$d_i^{\text{geo}} = s_i \sqrt{\frac{4A_i^{\text{met}}}{\pi}}. \tag{3}$$

Because this area-equivalent diameter may differ from the clinically measured transverse diameter, we apply a bounded residual correction:

$$\hat{d}_i^{\text{ind}} = d_i^{\text{geo}} + \delta_{\max} \tanh\big(f_d([\mathbf{z}_i^g; \mathbf{z}_i^{\text{ind}}; \mathbf{h}_i^m])\big), \tag{4}$$

where $\mathbf{z}_i^g$ and $\mathbf{z}_i^{\text{ind}}$ are frozen visual features from the full image and an induration-centered region, respectively. $\mathbf{h}_i^m$ encodes calibrated morphometric descriptors extracted from P_i^{met}, including physical area, major/minor axes, shape ratio, and compactness. Here, f_d is a scalar regressor, $\delta_{\max}$ bounds the correction magnitude, and $\hat{d}_i^{\text{ind}}$ is the final transverse diameter prediction. No language-conditioned semantic information enters this estimator.

4.2 Multiview Dermal Grounding

Using P_i^{met} as geometric guidance, we construct three views: *Global*, preserving the full image; *ROI*, centered on the predicted induration; and *Masked*, suppressing irrelevant background. We ground four concepts: induration, blistering, necrosis, and double-ring reaction, with the latter three forming the $K_s = 3$ dermal findings used for case-level interpretation.

For patch p from view v, let $\mathbf{P}_{i,v,p}$ denote its frozen visual feature and $\mathbf{q}_k$ the query for concept k. Their language-conditioned response is

$$\ell_{i,v,k,p} = \frac{\mathbf{q}_k^\top (W_p \mathbf{P}_{i,v,p} + \mathbf{e}_v)}{\sqrt{d}}, \tag{5}$$

where W_p projects visual features into a d-dimensional grounding space, $\mathbf{e}_v$ encodes view identity, and $\ell_{i,v,k,p}$ is the grounding logit. The responses are reconstructed and mapped to original-image coordinates as grounding maps $\widetilde{G}_{i,v,k}$.

Grounding is supervised by

$$\mathcal{L}_{\text{ground}} = \mathcal{L}_{\text{Dice}}^{+} + \mathcal{L}_{\text{BCE}}^{\text{bal}}, \tag{6}$$

where the two terms denote positive-map Dice and class-balanced BCE losses.

4.3 Task-Routed Clinical Fusion

Let $\mathbf{x}_i^c$ denote the permitted target-independent clinical context, comprising age, sex, post-injection measurement interval, recorded color, and palpation/texture description. Measured diameters, target dermal descriptions, final reaction assessments, and explicit metrology variables are excluded to reduce shortcut learning.

The visual representations are fused with clinical context as

$$\mathbf{h}_i = \mathrm{LN}\big(F_c([\mathbf{z}_i^g; \mathbf{z}_i^{\text{ind}}]) + E_c(\mathbf{x}_i^c)\big), \tag{7}$$

where F_c and E_c encode visual and clinical information, LN denotes layer normalization, and $\mathbf{h}_i$ is the fused semantic representation. Separate prediction heads estimate blistering, necrosis, double-ring reaction, the presence of any strong finding, and the overall four-class reaction grade.

For each dermal finding k, the aligned multiview grounding maps are aggregated using a Top-q operator to obtain an evidence score $g_{i,k} \in [0,1]$. Let $\hat{y}_{i,k}$ denote the corresponding case-level prediction. Their consistency is encouraged by

$$\mathcal{L}_{\text{cons}} = \frac{1}{K_s} \sum_{k=1}^{K_s} \mathrm{SmoothL1}\left(\hat{y}_{i,k}, g_{i,k}\right), \qquad K_s = 3, \tag{8}$$

linking semantic confidence to directly supervised localized evidence.

4.4 Semantic Mask Refinement

Semantic Mask Refinement uses only the induration grounding maps from the Global, ROI, and Masked views, denoted jointly by $\widetilde{\mathcal{G}}_i$, to refine the metrology mask without modifying physical measurement.

Let L_i^{met} denote the logits underlying P_i^{met}. The refined logits and final segmentation are

$$L_i^{\text{seg}} = L_i^{\text{met}} + \lambda_{\text{mask}} \tanh\Big(R_\theta([P_i^{\text{met}}; \widetilde{\mathcal{G}}_i])\Big), \qquad P_i^{\text{seg}} = \sigma(L_i^{\text{seg}}), \tag{9}$$

where R_θ is a residual refinement network, λ_{mask} bounds the semantic correction, and P_i^{seg} is the final refined induration mask.

The reported physical diameter remains determined by the frozen metrology pathway,

$$\hat{d}_i^{\text{ind}} = f_{\text{met}}(P_i^{\text{met}}, I_i, s_i), \tag{10}$$

where f_{met} denotes the complete calibrated diameter estimator; P_i^{seg} is never fed back into it.

To preserve geometric consistency, let $d_i^{\text{seg-geo}}$ denote the calibrated area-equivalent diameter derived from P_i^{seg}. We constrain it around the frozen geometric anchor using

$$\mathcal{L}_{\text{pres}} = \max\left(0, |d_i^{\text{seg-geo}} - \mathrm{sg}(d_i^{\text{geo}})| - \epsilon_d\right), \tag{11}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and ϵ_d is the allowed geometric tolerance. With M_i^{ind} denoting the clinician-defined induration mask, the refinement objective is

$$\mathcal{L}_{\text{SMR}} = \mathcal{L}_{\text{seg}}(P_i^{\text{seg}}, M_i^{\text{ind}}) + \lambda_{\text{pres}} \mathcal{L}_{\text{pres}}, \tag{12}$$

where $\mathcal{L}_{\text{seg}}$ is the segmentation loss and λ_{pres} controls geometric preservation. Thus, P_i^{met} supports calibrated measurement, whereas P_i^{seg} provides the final semantic-refined segmentation.

Table 1: Induration segmentation and physical metrology results. Best and second-best results are shown in **bold** and underline, respectively.

Method	Dice ↑	IoU ↑	MAE (mm) ↓	RMSE (mm) ↓	R^2 ↑	WTR@2mm ↑	WTR@5mm ↑
Segmentation-based Methods							
DeepLabv3+ (Chen et al., 2018)	0.7436	0.5919	4.0480	6.6768	0.5421	0.4775	0.7027
SegFormer-B2 (Xie et al., 2021)	0.7454	0.5941	4.2304	7.0442	0.4903	0.4865	0.7387
YOLOv8s-Seg (Jocher et al., 2023)	0.7525	0.6033	4.4622	7.9550	0.3499	0.4775	0.7297
nnWNet (Zhou et al., 2025)	0.7318	0.5770	4.4879	6.8671	0.5156	0.4414	0.6577
SegMoTE (Lu et al., 2026)	0.6636	0.4966	4.3440	6.9827	0.4991	0.4595	0.6757
PMRNet (Kang, 2026)	0.6910	0.5279	4.6967	7.3391	0.4467	0.4324	0.6667
EoMT (Kerssies et al., 2025)	0.1732	0.0318	11.3784	15.0604	-1.3299	0.3063	0.3243
Foundation and Multimodal Methods							
DINOv2 (Oquab et al., 2023)	0.4127	0.2600	7.1733	9.0067	0.1667	0.1982	0.4144
PanDerm (Yan et al., 2025d)	0.6812	0.5165	4.9077	7.8055	0.3741	0.4505	0.6757
LLaVA-1.5 (Liu et al., 2024)	0.4680	0.3054	6.9042	8.7262	0.2178	0.1441	0.4324
MedCLIPSeg (Koleilat et al., 2026)	0.7047	0.5440	4.2192	6.2916	0.5934	0.4144	0.6937
MediPPD	**0.8631**	**0.7592**	**2.3299**	**3.7901**	**0.8524**	**0.5946**	**0.8378**

Table 2: Dermal-sign recognition and spatial grounding results. AUROC and AUPRC correspond to the strong-reaction target; G-Dice denotes Rare G-Dice.

Method	Blistering AP ↑	Necrosis AP ↑	Double-ring AP ↑	Macro AP ↑	AUROC ↑	AUPRC ↑	G-Dice ↑	Loc. mAP@0.5 ↑	P–E ρ ↑
General and Medical Vision Models									
YOLOv8s-Seg (Jocher et al., 2023)	0.0360	0.0902	0.0450	0.0571	0.5822	0.1025	0.0012	0.0060	0.5831
Medic-AD (Park et al., 2026)	0.2770	0.0360	0.0450	0.1194	0.5408	0.0998	0.0000	0.0000	-0.0739
DINOv2 (Oquab et al., 2023)	0.0517	0.1504	0.2909	0.1644	0.8105	0.2844	0.0332	0.0019	0.3993
PanDerm (Yan et al., 2025d)	0.1156	0.1536	0.3858	0.2183	0.8148	0.2231	0.2646	0.0320	0.5529
Vision–Language Models									
LLaVA-1.5 (Liu et al., 2024)	0.5563	0.1090	0.2451	0.3035	0.9118	0.6011	0.1835	0.1389	0.6601
LLaVA-Med v1.5 (Li et al., 2023)	0.6917	0.1159	0.3908	0.3995	0.9020	0.5971	0.1616	0.0833	0.7038
DermLIP-PanDerm (Yan et al., 2025b;d)	0.8750	0.2627	0.4578	0.5318	0.9107	0.6201	0.1479	0.0294	0.6266
MAKE (Yan et al., 2025c)	0.8875	0.2546	0.3974	0.5132	0.9434	0.7358	0.0899	0.0022	0.3169
O-MAKE (Li et al., 2026)	0.8929	**0.3163**	0.4501	0.5531	0.9009	0.7075	0.0726	0.0049	-0.1654
MediPPD	**0.9167**	0.2677	**0.5089**	**0.5644**	**0.9719**	**0.7837**	**0.3768**	**0.2471**	**0.8913**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Dataset and protocol. We evaluate MediPPD on 2,216 patient-level PPD images using the fixed split in Section 3.2. Training optimizes model parameters, validation determines all development choices, and the held-out test set is used only for final evaluation. The strong-reaction target denotes blistering, necrosis, or double-ring reaction and is independent of induration diameter.

Evaluation metrics. We report Dice and IoU for segmentation; MAE, RMSE, R^2, WTR@2mm, and WTR@5mm for metrology; AP, Macro AP, AUROC, and AUPRC for dermal interpretation; and Rare G-Dice and localized mAP@0.5 for grounding. MAE is the primary metrology metric, and Spearman correlation (ρ) measures prediction–evidence consistency.

Baselines. We compare MediPPD with segmentation models, visual foundation and dermatology models, general/medical VLMs, and dermatology-specific VLMs. For metrology, applicable baselines use the same reference-based calibration protocol. All methods share the same data split, annotation targets, and evaluator.

Implementation. MediPPD follows the staged training procedure in Section 4: metrology is optimized and frozen first, followed by dermal grounding and interpretation, and finally Semantic Mask Refinement. Spatial outputs are restored to original-image coordinates before evaluation. Further implementation details are provided in the Appendix C.

5.2 MAIN RESULTS

Induration Segmentation and Physical Metrology. As shown in Table 1, MediPPD outperforms all baselines in segmentation and physical measurement, achieving 0.863 Dice, 2.33 mm MAE and $R^2 = 0.852$. Strong segmentation baselines can still exhibit large diameter errors, showing that spatial overlap alone does not ensure metric fidelity. These results support explicit reference calibration and bounded correction for reliable image-to-millimeter measurement.

Table 3: Ablation of physical metrology.

Variant	Geometry	Ref. Calib.	Correction	MAE ↓	RMSE ↓	R^2 ↑	WTR@2mm ↑	WTR@5mm ↑
DR	—	—	Direct	5.1840	8.7420	0.2147	0.3423	0.6126
PG	✓	—	Learned	4.7860	8.2140	0.3038	0.3964	0.6757
RCG	✓	✓	None	3.5172	7.9550	0.3499	0.4775	0.7297
RCG + UR	✓	✓	Unbounded	2.5140	4.1260	0.8249	0.5653	0.8198
RCG + BR	✓	✓	**Bounded**	**2.3299**	**3.7901**	**0.8524**	**0.5946**	**0.8378**

Table 4: Dermal interpretation ablation analysis.

GV	RV	DG	TR	Macro AP ↑	AUPRC ↑	G-Dice ↑	mAP@0.5 ↑
✓	—	—	—	0.3880	0.6120	0.1720	0.0890
✓	✓	—	—	0.4660	0.6900	0.2490	0.1420
✓	✓	✓	—	0.5290	0.7520	0.3410	0.2210
✓	✓	✓	✓	**0.5644**	**0.7837**	**0.3768**	**0.2471**

Table 5: Mask ablation.

VP	$\tilde{\mathcal{G}}^g$	$\tilde{\mathcal{G}}^r$	$\tilde{\mathcal{G}}^m$	Dice ↑	IoU ↑
✓	—	—	—	0.7525	0.6033
✓	✓	—	—	0.8140	0.6863
✓	✓	✓	—	0.8420	0.7271
✓	✓	✓	✓	**0.8631**	**0.7592**

Dermal-sign Recognition and Spatial grounding. Table 2 shows that MediPPD achieves the best performance on eight of nine metrics, including 0.564 Macro AP, 0.972 AUROC, and 0.891 prediction–evidence correlation. Notably, localized mAP@0.5 improves from the strongest baseline value of 0.139 to 0.247, while several competitive recognition models remain weak in localization. This gap confirms that accurate case-level recognition does not necessarily yield spatially faithful evidence, while MediPPD improves semantic prediction and spatial grounding jointly.

5.3 Ablation Study

We ablate the key components of MediPPD using the same evaluation protocol as the main experiments. Additional ablation studies are provided in the Appendix D.

Physical metrology. In Table 3, DR denotes direct appearance-to-diameter regression, PG uses pixel-space induration geometry without reference calibration, RCG uses the reference-calibrated geometric anchor, and UR/BR denote unbounded/bounded residual correction. Performance improves progressively from DR to PG and RCG, showing that explicit geometry and image-specific calibration both contribute to reliable measurement. Adding residual correction substantially reduces error, while bounding the correction around the calibrated anchor further improves MAE from 2.51 to 2.33 mm and achieves the best overall performance. These results support the proposed measurement chain of explicit geometry, reference calibration, and constrained residual correction.

Dermal interpretation. As shown in Table 4, ROI support, Dermal Grounding (DG), and Task-Routed fusion (TR) provide complementary improvements in both recognition and localization, increasing Macro AP from 0.388 to 0.564 and localized mAP@0.5 from 0.089 to 0.247. The consistent gains indicate that geometry-derived spatial support, explicit grounding supervision, and selectively routed clinical context jointly strengthen semantic prediction and spatial evidence quality.

Semantic mask refinement. Table 5 shows that progressively adding aligned grounding evidence from the Global, ROI, and Masked views improves the visual-prior segmentation from 0.753 to 0.863 Dice and from 0.603 to 0.759 IoU. The monotonic gains suggest that the three views provide complementary spatial cues for bounded mask refinement, while the physical-metrology pathway remains unaffected.

5.4 Additional Analysis

Shortcut and Clinical Information Audits. Figure 3 shows that directly providing induration diameter or morphometric variables does not improve dermal interpretation and degrades further when these variables are shuffled, supporting their exclusion from semantic heads. In contrast, correctly aligned target-independent clinical context improves AUPRC from 0.741 to 0.784 and Macro AP from 0.545 to 0.564, whereas shuffled context harms performance. These controls support selective routing of clinically relevant context while preventing geometric shortcuts.

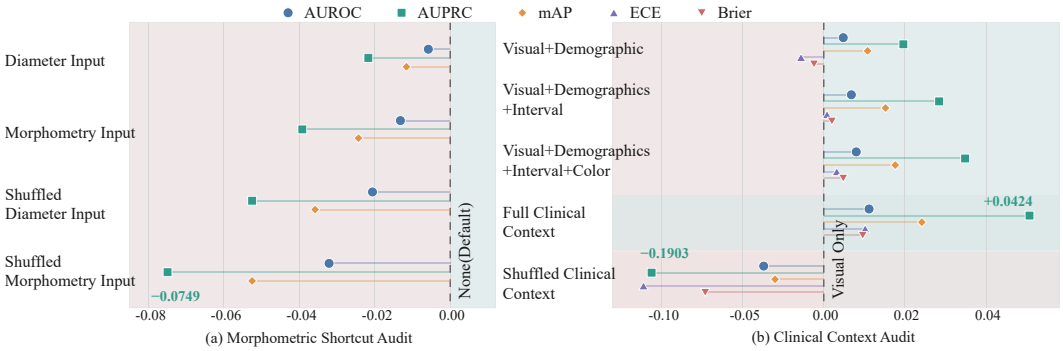


Figure 3: Audits of geometric shortcuts and target-independent clinical context.

Induration Segmentation and Physical Metrology. As shown in Table 1, MediPPD achieves the best segmentation and metrology performance, with 0.863 Dice, 2.33,mm MAE, $R^2 = 0.852$, and $WTR@5mm = 83.78\%$. Strong segmentation baselines still show larger diameter errors, indicating that spatial overlap alone does not guarantee metric fidelity. The high R^2 and $WTR@5mm$ further demonstrate that MediPPD produces measurements that are both accurate and consistent with clinician-defined diameters. This supports explicit reference calibration and bounded correction for reliable physical measurement.

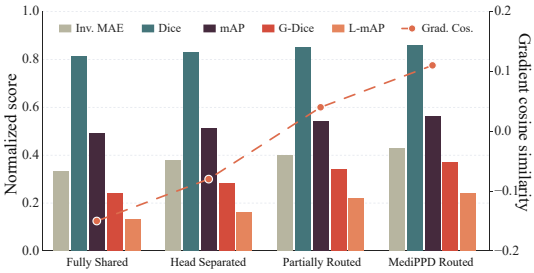


Figure 4: Selective routing Effects.

Geometry-Guided VLM Interpretation. Figure 5 shows that a vanilla VLM provides only coarse image-level interpretation, whereas MediPPD uses the geometry-derived reaction ROI to produce more localized dermal evidence. This geometry guidance narrows semantic attention to clinically relevant regions, improving the spatial correspondence between predicted findings and their visual evidence. This interaction is directional: geometry guides semantic reasoning, while VLM outputs do not alter the reference-calibrated diameter estimator.

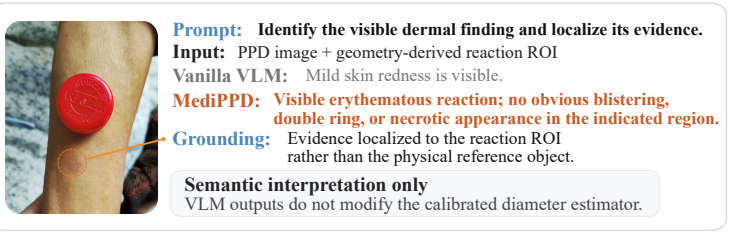


Figure 5: Geometry-derived localization guides dermal reasoning without altering calibrated metrology.

Robustness to Image Perturbations. Figure 6 shows that bounded residual refinement consistently outperforms the visual prior, direct semantic prediction, and unbounded refinement under all tested perturbations. Performance is most affected by Gaussian noise and center occlusion, but the bounded strategy remains the most robust. This supports using semantic evidence as a constrained correction rather than an independent segmentation signal.

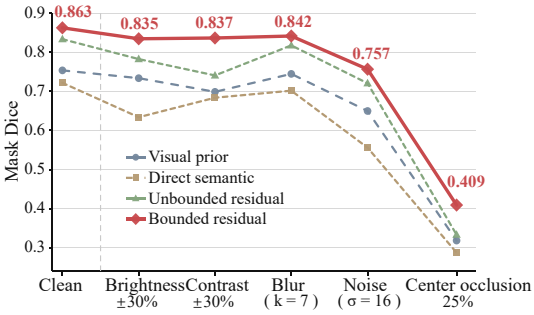
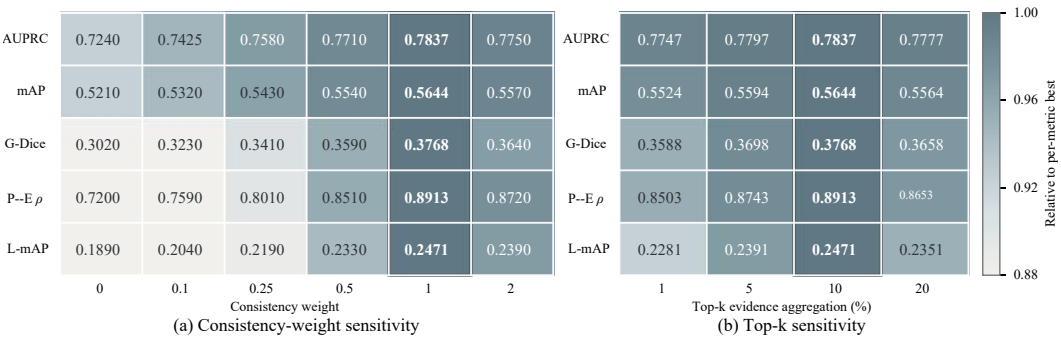


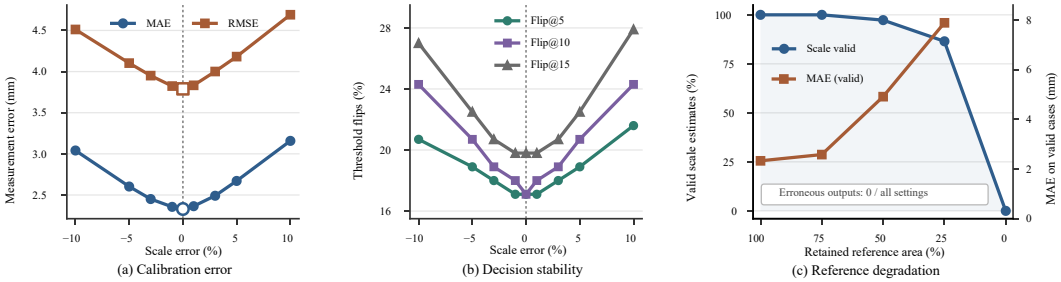
Figure 6: Robustness to Image Perturbations.

Sensitivity of Spatial grounding. Figure 7 reports hyperparameter sensitivity on the validation set. Performance improves as the prediction–evidence consistency weight increases to 1, while stronger coupling produces a mild decline. Similarly, Top-$k = 10\%$ provides the strongest overall balance between case-level recognition and spatial evidence quality. We therefore fix the consistency weight to 1 and Top-k to 10% for all subsequent test-set evaluations.



(a) Consistency-weight sensitivity (b) Top-k sensitivity

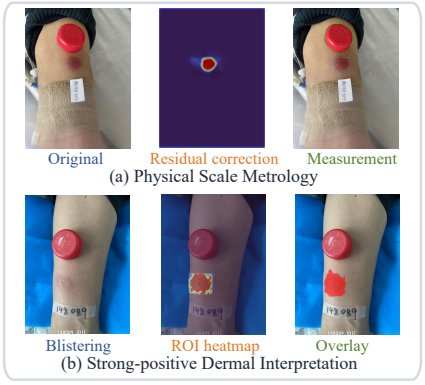
Figure 7: Sensitivity to consistency weighting and Top-k evidence aggregation.



(a) Calibration error (b) Decision stability (c) Reference degradation

Figure 8: Measurement reliability under scale perturbation and reference-object degradation.

Physical Measurement Reliability. Figure 8 shows that MediPPD remains stable under mild scale perturbations, while larger calibration errors gradually increase diameter error and threshold-flip rates. Progressive reference-object degradation likewise reduces measurement reliability, and no physical estimate is produced when the reference is unavailable. This predictable failure pattern avoids unsupported metric outputs when calibration becomes unreliable. This behavior reflects conservative degradation of the explicit reference-calibrated measurement pathway.



(a) Physical Scale Metrology (b) Strong-positive Dermal Interpretation

Figure 9: Qualitative metrology and grounded dermal interpretation.

Qualitative Metrology and Dermal Grounding. Figure 9 provides compact qualitative results for both physical-scale metrology and grounded dermal interpretation. The top row visualizes representative metrology cases using the original image, signed residual map, and final measurement result. The bottom row shows representative strong-positive dermal findings through ROI-centered input, grounding heatmap, and final evidence overlay. Across cases, MediPPD produces spatially localized dermal evidence while preserving accurate reference-calibrated diameter estimation. Additional analysis results are provided in the Appendix E and additional qualitative results are provided in the Appendix F .

6 Conclusion

In this paper, we propose **MediPPD**, a unified vision–language framework for automated PPD assessment that addresses the *Metric Fidelity Gap* and *Spatial Grounding Failure*. MediPPD preserves calibrated geometry for physical metrology while using geometry-derived evidence to support semantic reasoning without affecting diameter estimation. Experiments demonstrate consistent improvements in physical measurement, dermal-sign recognition, and spatial grounding over strong visual and multimodal baselines. The proposed benchmark further provides a unified basis for evaluating both metric fidelity and clinically meaningful grounding in automated PPD assessment. More broadly, our results highlight the importance of preserving measurement-critical geometry when integrating physical quantification with semantic interpretation.

REFERENCES

Olubunmi Adewale Akinola, Joseph Folorunsho Orimolade, Akindele Segun Afolabi, Habeeb Kehinde Shopeju, Emmanuel Adetiba, and Adeyinka Ajao Adewale. Analysis and measurement of tuberculin skin test induration using deep neural network. *Int. J. Online Biomed. Eng.*, 20: 137–159, 2024.

Seungmin Baek, Soyul Lee, Hayeon Jo, Hyesong Choi, and Dongbo Min. Tadformer: Task-adaptive dynamic transformer for efficient multi-task learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14858–14868. IEEE, 2025.

Junyi Chai, Shenyu Lu, and Xiaoqian Wang. Identifying and mitigating spurious correlation in multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25698–25707, 2025.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision*, pp. 833–851. Springer, 2018.

Ronald Dendere, Tinashe Mutsvangwa, Rene Goliath, Molebogeng X Rangaka, Ibrahim Abubakar, and Tania S Douglas. Measurement of skin induration size using smartphone images and photogrammetric reconstruction: pilot study. *JMIR Biomedical Engineering*, 2(1):e8333, 2017.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Hasan Farooq, Murtaza Taj, Mehwish Nasim, and Arif Mahmood. Localization lens for improving medical vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 341–350. Springer, 2025.

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matthew Fisher, Emiliano Gambaretto, Sunil Hadap, and Jean-François Lalonde. A perceptual measure for deep single image camera calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2354–2363, 2018.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3922–3931. IEEE, 2021.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolo. *Zenodo*, 2023.

Boce Kang. Pmrnet: Physics-informed multi-scale refinement network for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15659–15668, 2026.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan De Geus. Your vit is secretly an image segmentation model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25303–25313. IEEE, 2025.

Taha Koleilat, Hojat Asgariandehkordi, Omid Nejatimanzari, Berardino Barile, Yiming Xiao, and Hassan Rivaz. Medclipseg: probabilistic vision-language adaptation for data-efficient and generalizable medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1406–1417, 2026.

Anastasios Konstantinos. Testing for tuberculosis. *Australian prescriber*, 33(1), 2010.

Byeong-Uk Lee, Jianming Zhang, Yannick Hold-Geoffroy, and In So Kweon. Single view scene scale estimation using scale field. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21435–21444. IEEE, 2023.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022.

Xieji Li, Siyuan Yan, Yingsheng Liu, H Peter Soyer, Monika Janda, Victoria Mar, and Zongyuan Ge. Multi-aspect knowledge-enhanced dermatological vision-language pretraining with multi-agent data generation. *IEEE Transactions on Medical Imaging*, 2026.

Guimeng Liu, Tianze Yu, Somayeh Ebrahimkhani, Shawn Lin, Kok Pin Ng, and Man Cheung. How do medical mllms fail? a study on visual grounding in medical images. In *International Conference on Learning Representations*, volume 2026, pp. 95215–95268, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 1871–1880. IEEE, 2019.

Yujie Lu, Jingwen Li, Sibo Ju, Yanzhou Su, He Yao, Yisong Liu, Min Zhu, and Junlong Cheng. Segmote: Token-level mixture of experts for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36332–36342, 2026.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3994–4003, 2016.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023.

Safa Naraghi, Tinashe Mutsvangwa, René Goliath, Molebogeng X Rangaka, and Tania S Douglas. Mobile phone-based evaluation of latent tuberculosis infection: Proof of concept for an integrated image capture and analysis system. *Computers in biology and medicine*, 98:76–84, 2018.

Felix Nützel, Mischa Dombrowski, and Bernhard Kainz. Generate to ground: Multimodal text conditioning boosts phrase grounding in medical vision-language models. *arXiv preprint arXiv:2507.12236*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

World Health Organization. *WHO consolidated guidelines on tuberculosis. Module 3: diagnosis*. World Health Organization, 2025.

Woohyeon Park, Jaeik Kim, Sunghwan Steve Cho, Pa Hong, Wookyoung Jeong, Yoojin Nam, Namjoon Kim, Ginny Y Wong, Ka Chun Cheung, and Jaeyoung Do. Medic-ad: Towards medical vision-language model's clinical intelligence. *arXiv preprint arXiv:2603.27176*, 2026.

Georgy Ponimatkin, Yann Labbé, Bryan Russell, Mathieu Aubry, and Josef Sivic. Focal length and object pose estimation via render and compare. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3815–3824. IEEE, 2022.

Xiaohan Qin, Xiaoxing Wang, and Junchi Yan. Towards consistent multi-task learning: Unlocking the potential of task-specific parameters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10067–10076. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 21315–21326. IEEE, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Huimin Yan, Xian Yang, Liang Bai, and Jiye Liang. Local alignment for medical vision-language pre-training. *IEEE Transactions on Image Processing*, 34:7321–7334, 2025a.

Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12681–12690. IEEE, 2025b.

Siyuan Yan, Xieji Li, Ming Hu, Yiwen Jiang, Zhen Yu, and Zongyuan Ge. Make: Multi-aspect knowledge-enhanced vision-language pretraining for zero-shot dermatological assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 369–379. Springer, 2025c.

Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, Vincent Tang, Aik Beng Ng, David Powell, Paul Bonnington, Simon See, Elisabetta Magnaterra, Peter Ferguson, Jennifer Nguyen, Pascale Guitera, Jose Banuls, Monika Janda, Victoria Mar, Harald Kittler, H. Peter Soyer, and Zongyuan Ge. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31(8):2691–2702, 2025d.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1):AIoa2400640, 2025.

Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20852–20862. IEEE, 2025.

Appendix Contents

A Benchmark and Annotation Details . . . 14
A.1 Dataset Statistics and Splits . . . 14
A.2 Annotation Protocol and Quality Control . . . 15
A.3 Clinical Context and Leakage Control . . . 16
B Detailed Methodology . . . 17
B.1 Physical Scale Metrology . . . 17
B.2 Multiview Dermal Grounding . . . 18
B.3 Task-Routed Clinical Fusion . . . 20
B.4 Semantic Mask Refinement . . . 21
B.5 Training . . . 22
C Implementation Details . . . 22
C.1 Network Configuration . . . 22
C.2 Training Strategy and Optimization . . . 23
C.3 Hyperparameter Settings . . . 24
C.4 Inference and Evaluation . . . 24
D Additional Ablation Studies . . . 26
D.1 Physical Metrology Design Analysis . . . 27
D.2 Morphometric Feature Contribution . . . 27
D.3 Multiview Configuration Analysis . . . 28
D.4 Semantic Grounding Design Analysis . . . 28
E Reliability and Sensitivity Analysis . . . 29
E.1 Multi-Seed Stability . . . 30
E.2 Diameter Correction Sensitivity . . . 30
E.3 Spatial Configuration Sensitivity . . . 31
E.4 Probability Calibration and Operating Points . . . 32
F Additional Qualitative Results . . . 33
F.1 Physical Metrology Visualization . . . 33
F.2 Dermal Grounding Visualization . . . 33
F.3 Additional Visual Diagnostics . . . 41

A BENCHMARK AND ANNOTATION DETAILS

This section provides additional details of the benchmark construction, annotation targets, data partitioning, and clinical-information control used throughout our experiments. The benchmark is designed to jointly evaluate physical-scale induration measurement and spatially grounded dermal interpretation while maintaining strict separation between development data and the test set.

A.1 DATASET STATISTICS AND SPLITS

The benchmark contains $N = 2{,}216$ PPD images collected from $2{,}216$ distinct patients, with one image retained for each patient. Each sample contains the PPD reaction site together with a physical reference object of known size and is associated with the supervision required for induration segmentation, physical measurement, dermal-sign recognition, and spatial grounding.

The clinical outcomes are organized into four categories: strong positive, moderate positive, positive, and negative. As summarized in Table 6, strong-positive reactions constitute the largest group, accounting for 40.43% of the benchmark, followed by negative (30.69%), moderate-positive (22.16%), and positive (8.08%) cases.

Table 6: Distribution of the four clinical outcome categories in the MediPPD benchmark.

Clinical Outcome	**Images**	**Proportion**
Strong positive	896	40.43%
Moderate positive	461	20.80%
Positive	179	8.08%
Negative	680	30.69%
Total	2,216	100.00%

In addition to the mutually exclusive clinical outcome categories, the benchmark contains multi-label image annotations for the reference object and visible reaction characteristics. These annotations are not mutually exclusive: a single image may simultaneously contain erythema/induration, blistering, necrosis, and a double-ring reaction. Therefore, the category counts in Table 7 should not be summed to obtain the total number of benchmark images.

Table 7: Distribution of image-level annotation categories. Because the annotations are multi-label, one image may contain multiple dermal findings.

Annotation Category	**Images**	**Proportion**
Reference cap	2,216	100.00%
Erythema/induration	1,533	69.18%
Blistering	79	3.56%
Necrosis	57	2.57%
Double-ring reaction	72	3.25%

The relatively low prevalence of blistering, necrosis, and double-ring reaction highlights the pronounced class imbalance of the strong-reaction findings. In contrast, erythema/induration is present in more than two thirds of the benchmark, while the physical reference cap is available in nearly all images. These distributions motivate the class-aware evaluation of dermal-sign recognition and the explicit treatment of the reference object in the physical metrology pathway.

We use a fixed patient-disjoint partition

$$\mathcal{D} = \mathcal{D}_{\mathrm{tr}} \cup \mathcal{D}_{\mathrm{val}} \cup \mathcal{D}_{\mathrm{te}}, \qquad \mathcal{D}_a \cap \mathcal{D}_b = \varnothing \quad (a \neq b), \tag{13}$$

with $1{,}551$ training cases, 221 validation cases, and 444 held-out test cases, corresponding to approximately $70\%/10\%/20\%$ of the complete benchmark (Table 8). Because each patient contributes only one image, the patient-disjoint constraint also prevents image-level duplication across partitions.

The partition is stratified to approximately preserve both the prevalence of the rare dermal findings and the distribution of clinician-measured induration diameter. This is important because blistering,

Table 8: Patient-level partition of the MediPPD benchmark.

Split	Patients	Images	Proportion
Training	1,551	1,551	69.99%
Validation	221	221	9.97%
Test	444	444	20.04%
Total	2,216	2,216	100.00%

necrosis, and double-ring reaction are substantially less frequent than ordinary PPD reactions, while induration diameter forms a continuous target with a broad physical range.

The three partitions have strictly separated roles. $\mathcal{D}_{\mathrm{tr}}$ is used to optimize model parameters and to estimate all training-dependent preprocessing statistics. $\mathcal{D}_{\mathrm{val}}$ is used for checkpoint selection, early stopping, hyperparameter selection, evidence-aggregation settings, routing design, mask-refinement configuration, and clinical-context configuration. All such choices are fixed before accessing $\mathcal{D}_{\mathrm{te}}$, which is reserved exclusively for the final comparison of MediPPD, baselines, and ablations.

For each case i, the benchmark provides the supervision tuple

$$\mathcal{A}_i = \left\{ M_i^{\mathrm{ind}}, d_i^{\mathrm{ind}}, \mathbf{y}_i, \mathcal{G}_i \right\}, \tag{14}$$

where M_i^{ind} is the clinician-defined induration mask, $d_i^{\mathrm{ind}} \in \mathbb{R}_{\geq 0}$ is the recorded transverse induration diameter in millimeters, $\mathbf{y}_i$ contains the case-level dermal-sign labels, and $\mathcal{G}_i$ contains the corresponding spatial evidence annotations. This unified annotation structure allows segmentation, physical metrology, recognition, and grounding to be evaluated on the same patients rather than as disconnected tasks.

A.2 Annotation Protocol and Quality Control

The annotations are organized around two complementary forms of supervision: *physical measurement supervision*, which specifies the clinically measured extent of induration, and *semantic–spatial supervision*, which specifies both the presence and location of clinically relevant dermal findings.

Induration annotation. The induration target is represented by a binary spatial mask

$$M_i^{\mathrm{ind}} \in \{0,1\}^{H_i \times W_i}, \tag{15}$$

which delineates the clinician-defined induration region in the image. Importantly, the corresponding transverse diameter d_i^{ind} is retained as an independent millimeter-scale clinical target rather than being defined by the pixel area of M_i^{ind}. This distinction is central to the benchmark: a segmentation can achieve high spatial overlap while still producing an inaccurate physical measurement after geometric calibration. It therefore permits segmentation quality and metric fidelity to be evaluated separately.

The image additionally contains a circular reference object with known real-world diameter D_c. In the current implementation, $D_c = 30$ mm. The reference establishes an image-specific conversion between pixel space and physical space and is used only by the geometry-sensitive metrology pathway. Thus, the benchmark does not assume a globally fixed pixel-to-millimeter ratio across images.

Dermal-sign annotation. We consider three clinically relevant dermal findings for case-level interpretation: *blistering*, *necrosis*, and *double-ring reaction*. Their binary labels are represented as

$$\mathbf{y}_i = \left[y_i^{\mathrm{bli}}, y_i^{\mathrm{nec}}, y_i^{\mathrm{ring}} \right] \in \{0,1\}^3. \tag{16}$$

For evaluation of the composite strong-reaction target, we deterministically define

$$y_i^{\mathrm{strong}} = \max\left(y_i^{\mathrm{bli}}, y_i^{\mathrm{nec}}, y_i^{\mathrm{ring}} \right), \tag{17}$$

so that a case is positive whenever at least one of the three strong dermal findings is present. The strong-reaction target therefore does not depend on the measured induration diameter.

In addition to case-level labels, localized evidence is annotated for the induration region and the three dermal findings. We write

$$\mathcal{G}_i = \left\{ G_i^{\mathrm{ind}}, G_i^{\mathrm{bli}}, G_i^{\mathrm{nec}}, G_i^{\mathrm{ring}} \right\}, \tag{18}$$

Table 9: Clinical-context allowlist and leakage-control policy.

Category	Variables	Use
Permitted	Age, sex, post-injection interval	Target-independent clinical context
Permitted	Recorded color, palpation/texture	Target-independent reaction context
Excluded	Transverse, longitudinal, or mean induration diameter	Direct metrology target or derivative
Excluded	Explicit morphometric measurements	Potential geometric shortcut
Excluded	Target dermal-sign descriptions or labels	Direct semantic target leakage
Excluded	Final reaction assessment / outcome	Downstream clinical target leakage

where each G_i^k denotes the spatial evidence associated with concept k in the original-image coordinate system. These annotations provide direct supervision for Multiview Dermal Grounding and allow semantic correctness to be evaluated separately from localization fidelity.

This paired representation is particularly important for rare findings. A correct image-level prediction is not treated as sufficient evidence of successful grounding: the corresponding spatial response must also overlap the annotated region. Consequently, classification AP and localized AP measure distinct properties and are reported separately throughout the paper.

Annotation consistency. The benchmark preserves a clear separation between the different target types. The clinician-recorded transverse diameter remains the scalar reference for physical metrology, while the induration mask provides spatial supervision; dermal-sign labels provide case-level semantic targets, while their evidence regions provide localization targets. No target is reconstructed from another target at evaluation time. This avoids circular evaluation—for example, deriving the reference diameter from the same mask whose metric fidelity is being assessed.

All spatial annotations are mapped into the original-image coordinate system before evaluation. For multiview training, annotations are transformed into the corresponding Global, ROI, or Masked coordinate system only as an intermediate supervision step and are mapped back to the original coordinates for final spatial evaluation. This ensures that grounding performance is compared under a common coordinate system across methods.

A.3 Clinical Context and Leakage Control

MediPPD optionally incorporates a restricted set of target-independent clinical variables into the case-interpretation pathway. Let

$$\mathbf{x}_i^c = \left[a_i,\, t_i,\, s_i,\, c_i^{\text{color}},\, c_i^{\text{touch}}\right] \tag{19}$$

denote the permitted clinical context, where a_i is age, t_i is the post-injection measurement interval, s_i denotes sex, c_i^{color} is the recorded local color description, and c_i^{touch} is the recorded palpation/texture description.

Continuous variables are normalized using statistics computed exclusively from the training partition:

$$\widetilde{x}_{i,j} = \frac{x_{i,j} - \mu_j^{\text{tr}}}{\sigma_j^{\text{tr}} + \epsilon}, \tag{20}$$

where μ_j^{tr} and σ_j^{tr} are the training-set mean and standard deviation. Categorical vocabularies are likewise constructed only from the training set and are encoded independently of the held-out test data; unseen or missing categories are mapped to a dedicated `<MISSING>` entry.

To prevent target leakage, clinical inputs are defined by an explicit allowlist. Variables directly encoding measurement outcomes, target findings, or final clinical conclusions are excluded from model inputs. Table 9 summarizes this separation.

Leakage control is enforced not only at the input-variable level but also through task routing. In the default configuration, the case-interpretation branch does not receive the calibrated diameter,

geometric anchor, or morphometric vector used by Physical Scale Metrology. Formally, if

$$\mathbf{m}_i = [d_i^{\text{geo}}, A_i^{\text{geo}}, a_i^{\text{geo}}, b_i^{\text{geo}}, \ldots] \tag{21}$$

denotes metrology-derived information, the semantic predictor follows

$$\hat{\mathbf{y}}_i = f_{\text{sem}}\left(I_i, \mathbf{x}_i^c\right), \qquad \mathbf{m}_i \notin \text{Input}(f_{\text{sem}}), \tag{22}$$

unless a size-input ablation explicitly overrides this restriction. Geometry may determine the reaction ROI and thereby provide *spatial guidance*, but the resulting physical size itself is not supplied as a semantic shortcut.

The converse restriction is also maintained: semantic predictions and clinical context cannot modify the frozen calibrated diameter estimator. Consequently, information exchange is directional and controlled—geometry may support localization and semantic reasoning, whereas semantic evidence cannot alter the reported physical measurement.

Finally, all decisions concerning the clinical allowlist, preprocessing, routing, and evidence aggregation are made using the training and validation partitions only. The held-out test set contributes neither normalization statistics nor vocabulary construction, model selection, threshold selection, or architectural decisions. Together with the patient-disjoint split, these controls ensure that the reported test results evaluate generalization to previously unseen patients rather than reuse of target-derived or test-derived information.

B Detailed Methodology

This section provides additional architectural and optimization details of MediPPD. The implementation follows the principle of *selective separation with controlled interaction*: physical metrology retains an explicit reference-calibrated pathway, while semantic grounding and clinical interpretation operate on frozen vision–language representations. Information exchange is asymmetric: calibrated geometry may guide semantic reasoning, whereas semantic predictions and refined segmentation are prevented from modifying the reported physical measurement.

B.1 Physical Scale Metrology

Reference-based scale recovery. For image I_i, the geometry-sensitive parser produces a circular-reference probability mask P_i^c and a metrology-specific induration mask P_i^{met}. Their soft areas are computed as

$$A_i^c = \sum_u P_i^c(u), \qquad A_i^{\text{met}} = \sum_u P_i^{\text{met}}(u), \tag{23}$$

where u indexes image pixels. Assuming an approximately circular reference object, its area-equivalent pixel diameter is

$$d_{i,c}^{\text{px}} = \sqrt{\frac{4A_i^c}{\pi}}. \tag{24}$$

Given its known physical diameter $D_c = 30$ mm, the image-specific physical scale is

$$s_i = \frac{D_c}{d_{i,c}^{\text{px}}} \quad [\text{mm/pixel}]. \tag{25}$$

Unlike methods that assume a common pixel scale across images, Eq. 25 performs independent calibration for every case. When the reference object cannot be reliably recovered, the physical scale is marked invalid and no unsupported millimeter estimate is produced.

Calibrated geometric anchor. The metrology mask is first converted into an area-equivalent diameter in pixel space and subsequently into physical units:

$$d_i^{\text{geo}} = s_i \sqrt{\frac{4A_i^{\text{met}}}{\pi}}. \tag{26}$$

This quantity forms the explicit geometric anchor of the diameter estimator. It is intentionally computed before any semantic reasoning and therefore remains directly traceable to the reference object and predicted induration geometry.

In addition to d_i^{geo}, we extract a compact set of calibrated morphometric descriptors from P_i^{met}. Let a_i^{px} and b_i^{px} denote the major and minor axes of the lesion geometry and let L_i^{px} denote its perimeter. Their calibrated counterparts are

$$a_i^{\mathrm{mm}} = s_i a_i^{\mathrm{px}}, \qquad b_i^{\mathrm{mm}} = s_i b_i^{\mathrm{px}}, \qquad A_i^{\mathrm{mm}^2} = s_i^2 A_i^{\mathrm{met}}. \tag{27}$$

The resulting descriptor vector includes the calibrated diameter anchor, physical area, major and minor axes, aspect ratio, compactness, and mask confidence:

$$\mathbf{m}_i = \left[d_i^{\mathrm{geo}}, A_i^{\mathrm{mm}^2}, a_i^{\mathrm{mm}}, b_i^{\mathrm{mm}}, \frac{a_i^{\mathrm{px}}}{b_i^{\mathrm{px}} + \epsilon}, \frac{4\pi A_i^{\mathrm{met}}}{(L_i^{\mathrm{px}})^2 + \epsilon}, c_i^{\mathrm{met}} \right]. \tag{28}$$

Different dimensions are normalized before projection, and a lightweight encoder produces

$$\mathbf{h}_i^m = E_m(\mathbf{m}_i). \tag{29}$$

Visual support and bounded correction. Frozen global and induration-centered visual features are denoted by $\mathbf{z}_i^g$ and $\mathbf{z}_i^{\mathrm{ind}}$, respectively. Rather than directly regressing an unconstrained diameter from appearance, MediPPD predicts only a residual around the calibrated anchor:

$$\mathbf{h}_i^d = F_d([\mathbf{z}_i^g; \mathbf{z}_i^{\mathrm{ind}}; \mathbf{h}_i^m]), \tag{30}$$

$$\Delta d_i = \delta_{\max} \tanh\left(\mathbf{w}_d^\top \mathbf{h}_i^d + b_d\right), \tag{31}$$

$$\hat{d}_i^{\mathrm{ind}} = d_i^{\mathrm{geo}} + \Delta d_i. \tag{32}$$

We use $\delta_{\max} = 3$ mm in the default configuration. Therefore,

$$\left| \hat{d}_i^{\mathrm{ind}} - d_i^{\mathrm{geo}} \right| \le \delta_{\max}, \tag{33}$$

which prevents the learned appearance model from arbitrarily overriding the physically calibrated estimate.

The final residual layer is initialized with zero weights and zero bias, so that $\Delta d_i = 0$ at initialization. Training therefore starts from the explicit calibrated solution and learns only limited corrections required to better match the clinician-recorded transverse diameter.

Diameter regression is optimized using

$$\mathcal{L}_{\mathrm{diam}} = \mathrm{SmoothL1}\left(\hat{d}_i^{\mathrm{ind}}, d_i^{\mathrm{ind}}\right). \tag{34}$$

After metrology optimization, the complete diameter pathway is frozen before semantic interpretation is trained.

B.2 Multiview Dermal Grounding

Geometry-guided view construction. The metrology mask provides spatial guidance for constructing three complementary image views:

$$\mathcal{V}_i = \{I_i^g, I_i^r, I_i^m\}, \tag{35}$$

corresponding to the *Global*, *ROI*, and *Masked* views.

The Global view is the complete image and retains contextual information. The ROI view is cropped around the predicted induration support and enlarged by 15% on each side to retain boundary context. If no reliable ROI can be recovered, the complete image is used as the deterministic fallback. The Masked view preserves pixels within the geometry-derived reaction region while replacing irrelevant background with a fixed neutral value, thereby reducing dependence on acquisition background and surrounding skin context.

These views are deliberately complementary: the Global view preserves large-scale context, the ROI view increases effective resolution around the reaction, and the Masked view suppresses potentially distracting background information.

Frozen vision–language representation. The default implementation uses a frozen LLaVA-1.5-7B visual tower as the vision–language backbone. The backbone parameters are not updated during MediPPD training, and no autoregressive text generation is required. Spatial features from the three views are extracted and adaptively organized on a 12×12 grid:

$$\mathbf{P}_i \in \mathbb{R}^{3\times 144\times D_v}, \tag{36}$$

where D_v is the frozen visual-feature dimension.
For each view $v \in \{g, r, m\}$, its global visual representation is obtained by spatial pooling:

$$\mathbf{z}_i^v = \frac{1}{144} \sum_{p=1}^{144} \mathbf{P}_{i,v,p}. \tag{37}$$

The Global and ROI representations correspond to $\mathbf{z}_i^g$ and $\mathbf{z}_i^{\text{ind}}$ used by the downstream branches.

Language-conditioned concept queries. We define four semantic concepts corresponding to induration, blistering, necrosis, and double-ring reaction. Each concept k is initialized from a descriptive text embedding $\mathbf{e}_k^s$ and transformed into a trainable grounding query:

$$\mathbf{q}_k = W_s \mathbf{e}_k^s + \Delta \mathbf{q}_k, \tag{38}$$

where W_s projects the frozen semantic embedding and $\Delta \mathbf{q}_k$ is a learnable query offset. This initialization gives each query a clinically interpretable semantic starting point while still allowing dataset-specific adaptation.

For patch p of view v, the grounding response is

$$\ell_{i,v,k,p} = \frac{\mathbf{q}_k^\top \left(W_p \mathbf{P}_{i,v,p} + \mathbf{e}_v \right)}{\sqrt{d}}, \tag{39}$$

where W_p maps the frozen visual representation into the d-dimensional grounding space and $\mathbf{e}_v$ is a learned view embedding. The resulting tensor has dimension

$$\boldsymbol{\ell}_i \in \mathbb{R}^{3 \times 4 \times 12 \times 12}. \tag{40}$$

Coordinate restoration. Global and Masked responses are directly resized to the target image coordinate system. ROI responses are first reconstructed in the crop coordinate system and then inversely mapped using the ROI transformation. We denote this operation by

$$\widetilde{G}_{i,v,k} = \mathcal{T}_{i,v}^{-1} \left(\sigma(\ell_{i,v,k}) \right), \tag{41}$$

where $\mathcal{T}_{i,v}^{-1}$ restores view-specific predictions to the original image. Consequently, semantic evidence from all views can be compared and fused in a common coordinate frame.

Sparse grounding supervision. Ground-truth spatial annotations are transformed into each view using the same geometric mapping and resized to the 12×12 grounding lattice. Dermal findings are sparse, particularly for blistering, necrosis, and double-ring reaction. A conventional unweighted BCE objective can therefore be dominated by negative patches.

For a grounding target containing N_+ positive and N_- negative patches, we define the positive weight as

$$w_+ = \text{clip}\left(\frac{N_-}{\max(N_+, 1)}, 1, 25 \right). \tag{42}$$

The class-balanced binary cross-entropy is then

$$\mathcal{L}_{\text{BCE}}^{\text{bal}} = -\frac{1}{N} \sum_p \left[w_+ G_p \log \hat{G}_p + (1 - G_p) \log(1 - \hat{G}_p) \right]. \tag{43}$$

For maps containing positive evidence, we additionally use

$$\mathcal{L}_{\text{Dice}}^{+} = 1 - \frac{2 \sum_p \hat{G}_p G_p + \epsilon}{\sum_p \hat{G}_p + \sum_p G_p + \epsilon}. \tag{44}$$

If no positive patch exists for a concept in a training sample, the positive-map Dice term is omitted while the balanced BCE remains active. The complete grounding objective is

$$\mathcal{L}_{\text{ground}} = \mathcal{L}_{\text{Dice}}^{+} + \mathcal{L}_{\text{BCE}}^{\text{bal}}. \tag{45}$$

This design discourages the trivial all-negative solution while preserving supervision for negative cases.

B.3 TASK-ROUTED CLINICAL FUSION

Selective task routing. Instead of concatenating all available features into a single shared representation, MediPPD explicitly restricts the information accessible to each task:

$$\begin{aligned} \mathcal{R}_{\text{met}} &: \{\mathbf{z}^g, \mathbf{z}^{\text{ind}}, \mathbf{h}^m\} \to \hat{d}^{\text{ind}}, \\ \mathcal{R}_{\text{ground}} &: \{\mathbf{P}^g, \mathbf{P}^r, \mathbf{P}^m, \mathbf{q}\} \to \widetilde{\mathcal{G}}, \\ \mathcal{R}_{\text{case}} &: \{\mathbf{z}^g, \mathbf{z}^{\text{ind}}, \mathbf{x}^c\} \to \hat{\mathbf{y}}. \end{aligned} \tag{46}$$

In particular, the case-level semantic pathway does not receive d_i^{geo}, $\hat{d}_i^{\text{ind}}$, or the explicit morphometric vector $\mathbf{m}_i$ in the default configuration. This prevents induration size from becoming a shortcut for predicting strong dermal findings. Geometry influences semantic interpretation only through spatial support such as ROI construction.

Visual and clinical fusion. The Global and ROI visual representations are first combined as

$$\mathbf{h}_i^v = F_c\left([\mathbf{z}_i^g; \mathbf{z}_i^{\text{ind}}]\right). \tag{47}$$

Permitted clinical variables include age, sex, post-injection measurement interval, recorded local color, and palpation/texture description. Continuous variables are standardized using training-set statistics, whereas categorical variables are encoded using vocabularies built exclusively from the training partition; missing or unseen values are mapped to a dedicated missing category.

The resulting clinical representation is

$$\mathbf{h}_i^c = E_c(\mathbf{x}_i^c), \tag{48}$$

and fusion is performed through residual addition followed by layer normalization:

$$\mathbf{h}_i = \text{LN}\left(\mathbf{h}_i^v + \mathbf{h}_i^c\right). \tag{49}$$

This formulation retains a visual prediction pathway while allowing target-independent clinical information to provide a controlled residual contribution.

Multi-head case prediction. Three prediction heads are applied to $\mathbf{h}_i$. The dermal-sign head predicts blistering, necrosis, and double-ring reaction independently:

$$\hat{\mathbf{y}}_i^s = \sigma\left(W_s^c \mathbf{h}_i + \mathbf{b}_s\right) \in [0,1]^3. \tag{50}$$

A separate binary head predicts the presence of any strong dermal finding:

$$\hat{y}_i^{\text{strong}} = \sigma\left(\mathbf{w}_a^\top \mathbf{h}_i + b_a\right), \tag{51}$$

rather than deriving this probability by taking the maximum of the three individual sign scores. An additional head estimates the four-class overall reaction grade:

$$\hat{\mathbf{y}}_i^q = \text{softmax}\left(W_q \mathbf{h}_i + \mathbf{b}_q\right). \tag{52}$$

The three rare dermal findings are optimized using focal binary cross-entropy with class-dependent positive weighting, the strong-reaction output uses weighted BCE, and the reaction-grade output uses multiclass cross-entropy. Class weights are computed only from the training set and are capped at 25 to avoid unstable gradients from extremely rare categories.

Prediction–evidence consistency. To connect case-level semantic confidence with localized evidence, the spatial responses for sign k are summarized by a Top-q operator:

$$g_{i,k} = \text{TopMean}_q\left(\{\widetilde{G}_{i,v,k}\}_{v\in\{g,r,m\}}\right), \tag{53}$$

where the operator averages the highest-response fraction of the aligned semantic evidence. The default setting uses $q = 10\%$.

Consistency between semantic prediction and localization is optimized by

$$\mathcal{L}_{\text{cons}} = \frac{1}{K_s}\sum_{k=1}^{K_s} \text{SmoothL1}\left(\hat{y}_{i,k}^s, g_{i,k}\right), \qquad K_s = 3. \tag{54}$$

The consistency term does not replace direct spatial annotation. Instead, it connects two independently supervised outputs and discourages high semantic confidence without corresponding localized evidence.

The semantic-stage objective can therefore be written as

$$\mathcal{L}_{\text{sem}} = \lambda_s \mathcal{L}_{\text{sign}} + \lambda_a \mathcal{L}_{\text{strong}} + \lambda_q \mathcal{L}_{\text{grade}} + \lambda_g \mathcal{L}_{\text{ground}} + \lambda_c \mathcal{L}_{\text{cons}}, \tag{55}$$

with only terms relevant to a particular ablation activated during training.

B.4 SEMANTIC MASK REFINEMENT

Separated measurement and segmentation masks. MediPPD maintains two distinct induration masks. The metrology mask P_i^{met} belongs to the frozen physical pathway and determines d_i^{geo} and the final physical estimate $\hat{d}_i^{\mathrm{ind}}$. Semantic Mask Refinement instead produces P_i^{seg}, which is used exclusively as the final segmentation output.

This distinction prevents improvements in semantically refined boundaries from retrospectively changing the calibrated measurement.

Multiview evidence alignment. Only the induration grounding channel from the three views is used for mask refinement:

$$\widetilde{\mathcal{G}}_i^{\mathrm{ind}} = \left[\widetilde{G}_{i,g,\mathrm{ind}}, \widetilde{G}_{i,r,\mathrm{ind}}, \widetilde{G}_{i,m,\mathrm{ind}}\right]. \tag{56}$$

Global and Masked maps can be resized directly into the common image frame. The ROI map is first restored using the inverse crop transformation so that its response occupies only the corresponding region in the original image. This coordinate restoration prevents local ROI coordinates from being incorrectly interpreted as global image coordinates.

Bounded residual refinement. Let L_i^{met} denote the logits underlying the frozen metrology mask,

$$P_i^{\mathrm{met}} = \sigma(L_i^{\mathrm{met}}). \tag{57}$$

The residual network takes the frozen mask together with the aligned semantic evidence:

$$\mathbf{U}_i = [P_i^{\mathrm{met}}; \widetilde{\mathcal{G}}_i^{\mathrm{ind}}]. \tag{58}$$

Rather than generating a new mask independently, it predicts a bounded logit correction:

$$\Delta L_i = \lambda_{\mathrm{mask}} \tanh\left(R_\theta(\mathbf{U}_i)\right), \tag{59}$$

followed by

$$L_i^{\mathrm{seg}} = L_i^{\mathrm{met}} + \Delta L_i, \qquad P_i^{\mathrm{seg}} = \sigma\left(L_i^{\mathrm{seg}}\right). \tag{60}$$

The $\tanh$ function explicitly limits how strongly semantic evidence can modify the geometry-derived mask. The last layer of R_θ is zero-initialized so that

$$P_i^{\mathrm{seg}} = P_i^{\mathrm{met}} \tag{61}$$

at the beginning of refinement training.

The default residual network is lightweight, using a 5×5 convolution followed by two 3×3 convolutional blocks with GroupNorm and GELU activations; the hidden width is 16 channels. This keeps the refinement capacity small relative to the frozen visual representation and encourages semantic evidence to function as a correction rather than an independent segmentation generator.

Geometric preservation. Although the refined mask is not fed back into physical metrology, we additionally regularize it to avoid arbitrarily changing the physical extent of the reaction. Its calibrated area-equivalent diameter is

$$d_i^{\text{seg-geo}} = s_i \sqrt{\frac{4 \sum_u P_i^{\mathrm{seg}}(u)}{\pi}}. \tag{62}$$

We constrain this value around the frozen geometric anchor:

$$\mathcal{L}_{\mathrm{pres}} = \max\left(0, |d_i^{\text{seg-geo}} - \mathrm{sg}\left(d_i^{\mathrm{geo}}\right)| - \epsilon_d\right), \tag{63}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and ϵ_d defines the tolerated geometric deviation.

The segmentation component combines overlap-sensitive and pixel-wise supervision:

$$\mathcal{L}_{\mathrm{seg}} = \mathcal{L}_{\mathrm{Dice}} + \mathcal{L}_{\mathrm{BCE}}^{\mathrm{bal}}, \tag{64}$$

and the complete refinement objective is

$$\mathcal{L}_{\mathrm{SMR}} = \mathcal{L}_{\mathrm{seg}} + \lambda_{\mathrm{pres}} \mathcal{L}_{\mathrm{pres}}. \tag{65}$$

Crucially, the final physical measurement remains

$$\hat{d}_i^{\mathrm{ind}} = f_{\mathrm{met}}\left(P_i^{\mathrm{met}}, I_i, s_i\right), \tag{66}$$

and not $f_{\mathrm{met}}(P_i^{\mathrm{seg}}, I_i, s_i)$. Semantic refinement can therefore improve the reported segmentation without introducing semantic feedback into the physical estimator.

B.5 TRAINING

MediPPD is trained sequentially so that later semantic objectives cannot overwrite the geometry required for physical measurement. The optimization procedure contains three stages.

Stage I: physical metrology. The geometry-sensitive metrology pathway and bounded diameter correction are optimized first. The diameter branch minimizes Eq. 34 together with the corresponding geometry-supervision objectives. The residual estimator is initialized at zero correction, so optimization begins from the explicit reference-calibrated geometric solution.

Once Stage I is complete, the entire physical diameter estimator, including reference calibration, morphometric encoding, and bounded residual correction, is frozen:

$$\nabla_{\theta_{\mathrm{met}}} \mathcal{L}_{\mathrm{sem}} = 0. \tag{67}$$

This constraint is essential because semantic classification losses are not permitted to alter the millimeter estimator.

Stage II: grounding and clinical interpretation. With the metrology pathway frozen, Multiview Dermal Grounding and Task-Routed Clinical Fusion are optimized jointly. Training activates the dermal-sign, strong-reaction, reaction-grade, rare-grounding, and prediction–evidence consistency objectives:

$$\mathcal{L}_{\mathrm{stage2}} = \lambda_s \mathcal{L}_{\mathrm{sign}} + \lambda_a \mathcal{L}_{\mathrm{strong}} + \lambda_q \mathcal{L}_{\mathrm{grade}} + \lambda_g \mathcal{L}_{\mathrm{ground}} + \lambda_c \mathcal{L}_{\mathrm{cons}}. \tag{68}$$

The frozen vision–language backbone remains unchanged; only the lightweight semantic projections, concept-query offsets, fusion modules, and prediction heads are optimized.

To reduce imbalance from rare dermal findings, positive class weights are derived exclusively from the training set and clipped at 25. Training sampling additionally increases the probability of cases containing rare positive findings. No class frequency from the validation or test partition is used to determine these weights.

Stage III: semantic mask refinement. Finally, all previously learned metrology and semantic modules are frozen and only R_θ is optimized according to

$$\mathcal{L}_{\mathrm{stage3}} = \mathcal{L}_{\mathrm{SMR}}. \tag{69}$$

Thus, refinement cannot modify either the physical diameter estimator or the semantic grounding representation from which its auxiliary evidence is obtained.

Optimization settings. For the lightweight MediPPD modules, we use AdamW with an initial learning rate of 10^{-3} and weight decay 10^{-4}. The gradient norm is clipped to 5. The adapter hidden dimension is 128 and the default batch size is 32. The metrology and semantic stages are trained for up to 20 and 40 epochs, respectively, and mask refinement is trained for up to 40 epochs.

Checkpoint selection is task-specific rather than being based on a single indiscriminate sum of all losses. Metrology checkpoints are selected using the corresponding validation objective before freezing. Grounding and interpretation checkpoints are selected using their semantic validation objectives, and Semantic Mask Refinement is selected using segmentation performance on the validation data. All hyperparameters and checkpoints are fixed before final evaluation on the held-out test partition.

The complete training procedure can be summarized as Metrology $\longrightarrow$ Freeze $\longrightarrow$ Grounding + Interpretation $\longrightarrow$ Freeze $\longrightarrow$ Mask Refinement, which operationalizes the proposed selective separation: geometry is learned and protected first, semantic reasoning is subsequently conditioned on geometry, and final segmentation refinement remains downstream of both.

C IMPLEMENTATION DETAILS

This section summarizes the network instantiation, optimization settings, default hyperparameters, and evaluation procedure used for MediPPD. Unless otherwise stated, all hyperparameters are selected using the validation partition and fixed before evaluation on the held-out test set.

C.1 NETWORK CONFIGURATION

Geometry-sensitive visual parser. The geometry-sensitive frontend predicts the circular physical-reference mask and the metrology induration mask used by Physical Scale Metrology. In our implementation, this parser is instantiated with YOLOv8s-Seg (Jocher et al., 2023). The frontend

operates at an input resolution of 768 pixels and is trained using segmentation supervision for the physical reference and clinician-defined induration regions. Importantly, this visual parser serves as the geometric frontend of MediPPD rather than constituting a methodological contribution by itself.

The circular reference object has a known physical diameter of

$$D_c = 30.0 \text{ mm}, \tag{70}$$

which is used for image-specific pixel-to-millimeter calibration. If a valid reference region cannot be recovered, the corresponding physical scale is considered unavailable and the system does not generate a millimeter measurement.

Vision–language backbone. For Multiview Dermal Grounding, we use the visual tower of LLaVA-1.5-7B (Liu et al., 2024) as the frozen vision–language encoder. Its parameters remain fixed throughout MediPPD training. No autoregressive language generation is performed; instead, intermediate spatial visual features are extracted for the Global, ROI, and Masked views.

The spatial representations from each view are adaptively organized into a

$$12 \times 12 \tag{71}$$

grid, corresponding to 144 patch features per view. Thus, the cached multiview feature tensor for each image is

$$\mathbf{P}_i \in \mathbb{R}^{3\times 144\times D_v}, \tag{72}$$

where the first dimension corresponds to Global, ROI, and Masked views and D_v denotes the frozen visual-feature dimension.

The ROI is obtained from the geometry-derived induration support and enlarged by 15% to preserve contextual information around the predicted boundary. The Masked view retains the reaction region while suppressing the remaining image background.

MediPPD adapters. The trainable MediPPD components are lightweight relative to the frozen visual encoder. Unless otherwise stated, the latent adapter dimension is set to

$$d_h = 128. \tag{73}$$

Separate projection and prediction modules are used for metrology, grounding, and case-level interpretation instead of a single fully shared prediction head.

The metrology correction head takes the pooled Global and ROI features together with encoded morphometric descriptors and produces a scalar bounded residual. The maximum residual magnitude is

$$\delta_{\max} = 3.0 \text{ mm}. \tag{74}$$

Its final layer is zero-initialized so that the initial prediction coincides with the reference-calibrated geometric estimate.

For dermal grounding, four semantic queries are maintained for induration, blistering, necrosis, and double-ring reaction. Patch features and concept queries are projected into a common grounding space, and view-specific embeddings distinguish the three visual views.

Task-Routed Clinical Fusion contains separate heads for the three dermal signs, the composite strong-reaction target, and the four-class reaction grade. The strong-reaction probability is predicted by an independent binary head rather than being obtained by directly taking the maximum of the three dermal-sign probabilities.

Mask-refinement network. Semantic Mask Refinement uses a lightweight convolutional residual network R_θ. The implementation contains an initial 5×5 convolution followed by two 3×3 convolutional blocks with GroupNorm and GELU activations. The hidden width is 16 channels. The final convolution is zero-initialized, causing the initial refined mask to reproduce the frozen metrology mask before semantic residual learning begins.

The network receives the metrology mask together with the aligned induration grounding maps from the Global, ROI, and Masked views. Its output is passed through a $\tanh$ function before being added to the metrology logits, thereby bounding the magnitude of semantic correction.

C.2 Training Strategy and Optimization

MediPPD is optimized sequentially to prevent semantic objectives from modifying the geometry required for physical metrology. All model selection and development decisions use the validation partition; the held-out test set is accessed only after the configuration has been fixed.

Geometry frontend. The geometry-sensitive segmentation frontend is trained for up to 60 epochs with a batch size of 12 at 768-pixel input resolution. Its predictions provide the reference and induration geometry required by the subsequent metrology and multiview modules.

Stage I: metrology optimization. The first MediPPD stage optimizes the physical-metrology branch, including the morphometric encoder and bounded residual estimator. The diameter head is supervised by the clinician-recorded transverse diameter using Smooth L1 loss. Training is performed for up to 20 epochs with a batch size of 32.

After this stage, the complete diameter estimator is frozen. In particular, subsequent semantic or segmentation losses cannot update the reference-calibration pathway, morphometric encoder, or bounded diameter correction.

Stage II: grounding and interpretation. The second stage jointly optimizes Multiview Dermal Grounding and Task-Routed Clinical Fusion for up to 40 epochs with batch size 32. The frozen LLaVA visual tower is not updated.

The optimization objective combines dermal-sign recognition, strong-reaction classification, reaction-grade classification, spatial grounding, and prediction–evidence consistency:

$$\mathcal{L}_{\text{stage2}} = \lambda_s \mathcal{L}_{\text{sign}} + \lambda_a \mathcal{L}_{\text{strong}} + \lambda_q \mathcal{L}_{\text{grade}} + \lambda_g \mathcal{L}_{\text{ground}} + \lambda_c \mathcal{L}_{\text{cons}}. \tag{75}$$

Rare dermal findings are handled using class-aware weighting. Positive weights are estimated exclusively from the training partition and clipped to a maximum value of 25 to prevent excessively large gradients. Sampling additionally increases the probability of cases containing rare positive findings.

Stage III: semantic mask refinement. After the metrology and semantic modules have been fixed, only the Semantic Mask Refinement network is optimized. Training is performed for up to 40 epochs. The refinement objective combines segmentation supervision with the metrology-preservation term introduced in Section B.4:

$$\mathcal{L}_{\text{stage3}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{pres}} \mathcal{L}_{\text{pres}}. \tag{76}$$

Neither the frozen metrology estimator nor the frozen vision–language representation is updated during this stage.

Optimizer. The trainable MediPPD adapters are optimized using AdamW with initial learning rate

$$\eta = 10^{-3} \tag{77}$$

and weight decay

$$\omega = 10^{-4}. \tag{78}$$

Gradient norms are clipped to 5. Checkpoints are selected using validation performance corresponding to the task optimized in each stage rather than using test-set performance.

The overall optimization order is therefore

$$\text{Geometry/Metrology} \rightarrow \text{Freeze} \rightarrow \text{Grounding + Interpretation} \rightarrow \text{Freeze} \rightarrow \text{Mask Refinement}. \tag{79}$$

C.3 Hyperparameter Settings

Table 10 summarizes the default configuration used in the main experiments.

The 15% ROI margin, 12×12 grounding resolution, and Top-10% evidence aggregation are the default settings used for the main configuration. Hyperparameter sensitivity for spatial evidence aggregation and consistency weighting is reported in the main paper, and all selected values are fixed before held-out test evaluation.

The correction bound $\delta_{\max} = 3$ mm is applied only to the learned diameter residual and does not constrain the calibrated geometric anchor itself. Similarly, the bounded residual used by Semantic Mask Refinement modifies only the final segmentation output; it is never propagated back to the physical diameter estimator.

C.4 Inference and Evaluation

Inference procedure. At inference time, each image is first processed by the geometry-sensitive frontend to obtain the reference mask and metrology induration mask. The reference region determines the image-specific millimeter scale, after which the calibrated geometric anchor and morphometric descriptors are computed.

Table 10: Default implementation and hyperparameter settings of MediPPD.

Component	Setting
Physical reference diameter	30.0 mm
Geometry frontend	YOLOv8s-Seg
Geometry input resolution	768 px
Geometry training epochs	60
Geometry batch size	12
Vision–language encoder	LLaVA-1.5-7B visual tower
VLM parameters	Frozen
Visual views	Global / ROI / Masked
ROI expansion margin	15%
Spatial grounding grid	12×12
Number of semantic queries	4
Adapter hidden dimension	128
Maximum diameter residual	± 3.0 mm
Metrology training epochs	20
Semantic training epochs	40
Main adapter batch size	32
Consistency Top-q	10%
Consistency weight	1.0
Maximum positive-class weight	25
Mask-refinement hidden width	16
Mask-refinement epochs	40
Optimizer	AdamW
Initial learning rate	10^{-3}
Weight decay	10^{-4}
Gradient-norm clipping	5

The geometry-derived induration region is then used to construct the Global, ROI, and Masked views. Frozen vision–language features are extracted from the three views and passed to the grounding and case-interpretation modules. Finally, the aligned induration grounding maps are supplied to Semantic Mask Refinement.

The prediction flow can be summarized as

$$I_i \rightarrow \left\{P_i^c, P_i^{\mathrm{met}}\right\} \rightarrow \left\{\hat{d}_i^{\mathrm{ind}}, \widetilde{\mathcal{G}}_i, \hat{\mathbf{y}}_i\right\} \rightarrow P_i^{\mathrm{seg}}. \tag{80}$$

The final diameter $\hat{d}_i^{\mathrm{ind}}$ is computed exclusively from the frozen physical-metrology pathway. The refined segmentation P_i^{seg} is therefore not fed back into diameter estimation.

Coordinate restoration. Predictions generated in ROI coordinates are inversely mapped to the original image coordinate system before spatial evaluation. Global and Masked grounding outputs are likewise resized to the common original-image frame. Consequently, all segmentation and grounding methods are evaluated using the same spatial coordinate system.

Induration segmentation. The final refined mask is evaluated against the clinician-defined induration annotation using Dice and intersection-over-union (IoU):

$$\mathrm{Dice} = \frac{2|P \cap M|}{|P| + |M|}, \qquad \mathrm{IoU} = \frac{|P \cap M|}{|P \cup M|}. \tag{81}$$

Probability masks are binarized using a threshold of 0.5 for mask-based evaluation.

Physical metrology. Physical diameter estimates are evaluated using mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination R^2:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{d}_i^{\mathrm{ind}} - d_i^{\mathrm{ind}}\right|, \tag{82}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{d}_i^{\mathrm{ind}} - d_i^{\mathrm{ind}}\right)^2}. \tag{83}$$

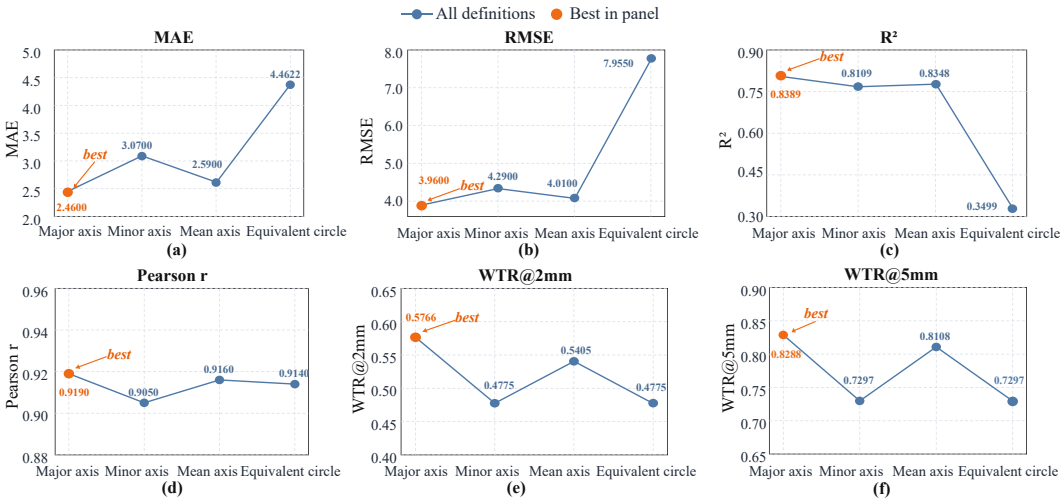


Figure 10: Comparison of reference-calibrated geometric diameter definitions before residual correction.

We additionally report within-tolerance rates at 2 and 5 mm:

$$\mathrm{WTR}\,@\tau = \frac{1}{N}\sum_{i=1}^{N} 1\left[\left|\hat{d}_i^{\mathrm{ind}} - d_i^{\mathrm{ind}}\right| \le \tau\right], \qquad \tau \in \{2,5\}\,\mathrm{mm}. \tag{84}$$

Dermal-sign recognition. Blistering, necrosis, and double-ring reaction are evaluated independently using Average Precision (AP), and their arithmetic mean is reported as Macro AP:

$$\mathrm{MacroAP} = \frac{1}{3}\sum_{k=1}^{3} \mathrm{AP}_k\,. \tag{85}$$

The composite strong-reaction prediction is evaluated using AUROC and AUPRC. Because strong dermal findings are relatively infrequent, AUPRC is reported alongside AUROC to reflect performance under class imbalance.

Spatial grounding. Grounding performance is evaluated separately from case-level recognition. We report Rare G-Dice for spatial overlap of the rare dermal concepts and localized mAP@0.5 for region-level localization. Localized predictions are evaluated only after all view-specific outputs have been restored to the original-image coordinate system.

Prediction–evidence consistency. For each dermal finding, the Top-10% spatial evidence score is compared with the corresponding case-level prediction. Spearman rank correlation,

$$\rho_{\text{P-E}} = \rho_{\text{Spearman}}\left(\hat{y}_{i,k}, g_{i,k}\right), \tag{86}$$

is reported to quantify whether higher semantic confidence is accompanied by stronger localized evidence.

Classification AP and localization AP are computed from different prediction objects and are therefore reported separately. A method can achieve high case-level AP without producing spatially faithful evidence; this distinction is central to the Spatial Grounding Failure examined in the benchmark.

Test-set protocol. All preprocessing rules, architecture choices, checkpoints, hyperparameters, evidence-aggregation settings, and decision rules are fixed using the training and validation partitions. The held-out test partition is used only once these choices have been finalized and is evaluated using the same protocol for MediPPD, competing baselines, and ablation variants.

D Additional Ablation Studies

We provide additional ablations that complement the main-paper analysis by examining the geometric definition used for physical measurement, the contribution of calibrated morphometric descriptors, the complete combination of multiview evidence, and the design choices underlying semantic

Table 11: Contribution of calibrated morphometric descriptors to physical metrology.

Morphometric Input	**MAE** ↓	**RMSE** ↓	R^2 ↑
None	2.6210	4.1320	0.8246
Diameter	2.4420	3.9810	0.8372
Diameter + Area	2.3910	3.9120	0.8428
Diameter + Axes	2.3520	3.8420	0.8483
Full Seven	**2.3299**	**3.7901**	**0.8524**

grounding. Unless otherwise specified, only the component under study is changed while the remaining MediPPD configuration is kept fixed.

D.1 Physical Metrology Design Analysis

The physical extent of an irregular induration region can be summarized using different geometric definitions. Given the major and minor axes a_i^{px} and b_i^{px}, mask area A_i^{met}, and image-specific scale s_i, we compare

$$\begin{aligned} d_i^{\mathrm{maj}} &= s_i a_i^{\mathrm{px}}, \\ d_i^{\mathrm{min}} &= s_i b_i^{\mathrm{px}}, \\ d_i^{\mathrm{mean}} &= \frac{s_i}{2}\left(a_i^{\mathrm{px}} + b_i^{\mathrm{px}}\right), \\ d_i^{\mathrm{eq}} &= s_i \sqrt{\frac{4A_i^{\mathrm{met}}}{\pi}}. \end{aligned} \tag{87}$$

Here d_i^{eq} corresponds exactly to the area-equivalent geometric anchor d_i^{geo} used in the main formulation. To isolate the behavior of different geometric summaries, all definitions are evaluated before applying the learned bounded residual correction.

As shown in Figure 10, the major-axis definition provides the strongest uncorrected agreement with the clinician-recorded transverse diameter, achieving an MAE of 2.4600 mm, an RMSE of 3.9600 mm, and $R^2 = 0.8389$. This behavior is expected because the clinical target is directional, whereas the minor axis, mean axis, and area-equivalent diameter summarize different aspects of lesion geometry. The area-equivalent definition retains a strong linear association with the clinical measurement but exhibits larger absolute error, reflecting the mismatch between an isotropic area-based summary and a directional transverse extent.

Importantly, MediPPD uses d_i^{eq} as a calibrated geometric anchor rather than as the final diameter prediction. Unlike a single principal-axis measurement, the area-equivalent anchor is directly determined by the complete soft induration mask and preserves an explicit connection between predicted lesion extent, image-specific reference calibration, and physical scale. The bounded residual estimator subsequently incorporates calibrated morphometric descriptors, including major and minor axes, together with visual evidence to recover task-specific directional information. With this correction, the final model reduces MAE from 4.4622 to 2.3299 mm, RMSE from 7.9550 to 3.7901 mm, and increases R^2 from 0.3499 to 0.8524, while WTR@2mm and WTR@5mm improve from 0.4775 and 0.7297 to 0.5946 and 0.8378, respectively. These results support using a stable reference-calibrated area anchor together with constrained task-specific correction rather than relying on a single uncorrected geometric definition.

D.2 Morphometric Feature Contribution

We next examine which calibrated geometric descriptors are useful for the bounded diameter correction. Starting from visual features alone, we progressively provide the geometric diameter, physical area, calibrated axes, and finally the complete seven-dimensional morphometric representation described in Section B.1. These features are supplied only to the metrology pathway; the case-interpretation pathway remains isolated from explicit lesion-size variables.

Table 11 shows a consistent improvement as increasingly informative physical descriptors are introduced. Using the geometric diameter alone reduces MAE from 2.6210 to 2.4420 mm. Adding physical area further lowers MAE to 2.3910 mm, while incorporating the calibrated axes improves it to 2.3520 mm. The complete descriptor set achieves the strongest performance, with 2.3299 mm MAE, 3.7901 mm RMSE, and $R^2 = 0.8524$.

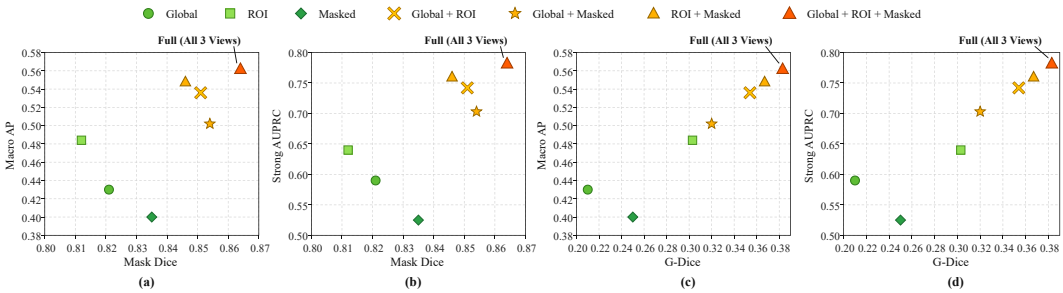


Figure 11: Complete ablation of Global, ROI, and Masked views.

These results indicate that the bounded correction benefits from more than a single diameter estimate. Area and axis geometry provide complementary information about lesion extent and anisotropy, while shape ratio, compactness, and mask confidence further characterize the reliability and morphology of the geometric anchor. The gain is progressive rather than being driven by one additional variable, supporting the compact multi-descriptor representation used by Physical Scale Metrology.

D.3 Multiview Configuration Analysis

The main method combines three complementary representations: the *Global* view preserves the complete visual context, the *ROI* view concentrates representation capacity around the geometry-derived reaction region, and the *Masked* view suppresses irrelevant background. To determine whether these views provide genuinely complementary evidence, we evaluate all seven non-empty combinations.

Several complementary behaviors are visible in Figure 11. Among single-view variants, ROI provides the strongest grounding and recognition results, reaching 0.3017 G-Dice and 0.4862 Macro AP, indicating the importance of concentrating semantic reasoning around the geometry-derived reaction region. In contrast, the Masked view produces the strongest single-view segmentation result with 0.8351 Dice, suggesting that background suppression is particularly useful for recovering the induration boundary. The Global view alone is weaker in localization but retains contextual information unavailable to the cropped alternatives.

Combining views consistently improves over their individual counterparts. Global + ROI increases G-Dice to 0.3512 and Strong AUPRC to 0.7442, while ROI + Masked obtains 0.3641 G-Dice and 0.7584 AUPRC. Using all three views yields the strongest result on every reported metric, including 0.8631 Mask Dice, 0.3768 G-Dice, 0.5644 Macro AP, and 0.7837 Strong AUPRC.

The results support the intended functional decomposition of the multiview design: Global preserves contextual appearance, ROI enhances local dermal detail, and Masked reduces background interference. No individual view simultaneously provides the strongest segmentation, grounding, and recognition performance, whereas their combination improves all three objectives.

D.4 Semantic Grounding Design Analysis

We further analyze three components of Multiview Dermal Grounding: semantic query initialization, sparse spatial supervision, and rare-class learning. Together, these experiments evaluate how language priors and imbalance-aware optimization affect the correspondence between case-level dermal interpretation and localized visual evidence.

Semantic query initialization. The grounding queries are initialized using clinically descriptive phrases rather than arbitrary learned vectors. We compare the default medical phrases with class names alone, a generic skin description, and random initialization. The quantitative comparison in Table 12 evaluates how different initialization strategies affect both case-level interpretation and spatial grounding performance.

Medical-phrase initialization performs best across all metrics, improving Macro AP from 0.4617 with random initialization to 0.5644 and G-Dice from 0.2867 to 0.3768. Class-name initialization already provides a meaningful advantage over random queries, but the richer descriptive phrases produce a further improvement. This suggests that semantically informative initialization provides a

Table 12: Effect of semantic query initialization.

Initialization	Blister AP	Necrosis AP	Ring AP	Macro AP	G-Dice	AUPRC
Medical phrase	**0.9167**	**0.2677**	**0.5089**	**0.5644**	**0.3768**	**0.7837**
Class name	0.8740	0.2210	0.4620	0.5190	0.3416	0.7412
Generic skin	0.8430	0.1980	0.4410	0.4940	0.3148	0.7084
Random	0.7910	0.1810	0.4130	0.4617	0.2867	0.6811

Table 13: Ablation of the sparse grounding objective.

Grounding Loss	G-Dice ↑	Loc. mAP ↑	Pos. Recall ↑	Empty Rate ↓	Macro AP ↑
BCE	0.2418	0.1712	0.3333	0.6486	0.4582
Balanced BCE	0.3507	0.2305	0.7778	0.5315	0.5358
Focal	0.3294	0.2376	0.6667	0.5766	0.5471
Positive Dice	0.3421	0.2113	**0.8889**	**0.4775**	0.5016
Dice + Balanced BCE	**0.3768**	**0.2471**	0.7778	0.4865	**0.5644**

useful prior for aligning sparse dermal findings with frozen visual patches, while the learnable query offset still permits task-specific adaptation.

Sparse grounding objective. We next compare alternative losses for sparse spatial evidence. Since rare dermal findings occupy only a small fraction of spatial patches, ordinary BCE is strongly dominated by negatives. Table 13 compares standard BCE, balanced BCE, focal loss, positive-map Dice, and the proposed combination.

Standard BCE produces the weakest result and a high empty-prediction rate of 0.6486, consistent with domination by negative patches. Positive Dice substantially increases positive-region recall to 0.8889, but its localized mAP and Macro AP remain lower than those of the combined objective. Balanced BCE provides stronger discrimination under sparse supervision, whereas Dice directly encourages overlap with positive evidence. Combining the two achieves the highest G-Dice (0.3768), localized mAP (0.2471), and Macro AP (0.5644), supporting their complementary roles in the grounding objective.

Rare-class learning. Finally, we evaluate alternative strategies for learning the three low-frequency dermal findings. Figure 12 shows that the default strategy combines imbalance-aware sampling with class-weighted focal optimization, allowing the model to increase exposure to rare positive cases without relying solely on large positive weights.

Weighted BCE achieves the highest positive recall (0.8889) and the highest necrosis AP, but it also increases the predicted-positive rate to 0.1081 and substantially reduces Strong AUPRC. In contrast, the combined strategy provides the strongest overall class balance: it achieves the highest blistering AP (0.9167), double-ring AP (0.5089), Macro AP (0.5644), and Strong AUPRC (0.7837), while maintaining a positive recall of 0.7778.

Taken together, the three analyses show that effective dermal grounding depends on complementary design choices at different levels. Descriptive semantic queries provide an informative concept prior, Dice plus balanced BCE prevents sparse spatial supervision from collapsing toward empty predictions, and the combined rare-class strategy improves recognition under severe class imbalance. These components jointly support the spatially grounded dermal interpretation used by MediPPD.

E Reliability and Sensitivity Analysis

We further evaluate the reliability of MediPPD under stochastic optimization and analyze the sensitivity of three design choices that are fixed in the main experiments: the maximum diameter-correction range, the spatial configuration of multiview grounding, and the operating threshold of the strong-reaction predictor. These experiments complement the robustness analyses in the main paper by focusing on optimization stability and configuration sensitivity rather than input perturbations.

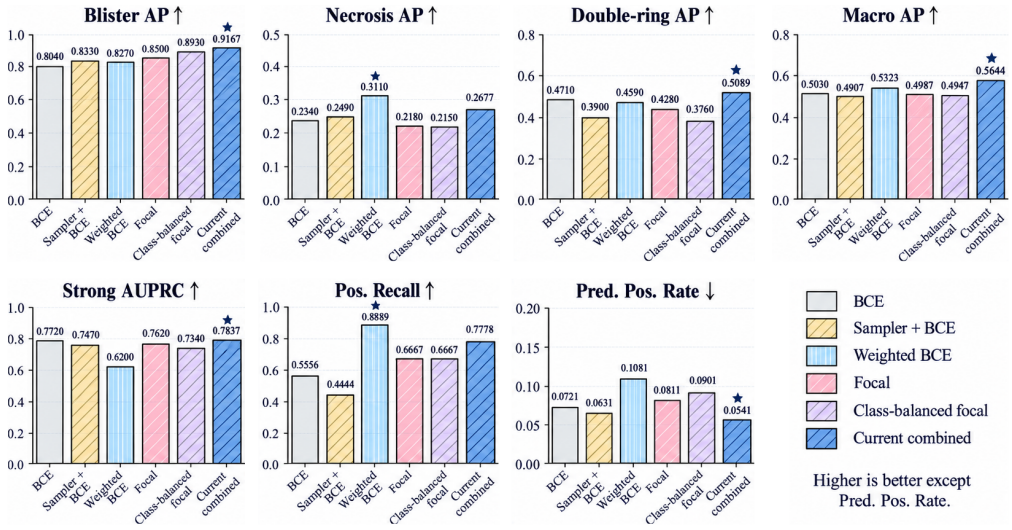


Figure 12: Comparison of rare-class learning strategies for dermal-sign interpretation.

Table 14: Multi-seed stability of MediPPD under a fixed data partition. The final row reports mean ± sample standard deviation across five runs.

Seed	Mask Dice ↑	MAE ↓	RMSE ↓	AUROC ↑	AUPRC ↑	Macro AP ↑	Loc. mAP@0.5 ↑
42	0.8631	2.3299	3.7901	0.9719	0.7837	0.5644	0.2471
3407	0.8586	2.4106	3.9214	0.9641	0.7568	0.5482	0.2354
2025	0.8672	2.2841	3.7488	0.9688	0.7905	0.5710	0.2510
2026	0.8610	2.3617	3.8426	0.9730	0.7781	0.5602	0.2438
2027	0.8594	2.4389	3.9542	0.9607	0.7489	0.5397	0.2316
Mean ± Std.	0.8619 ± 0.0034	2.3650 ± 0.0619	3.8514 ± 0.0864	0.9677 ± 0.0052	0.7716 ± 0.0179	0.5567 ± 0.0126	0.2418 ± 0.0081

E.1 Multi-Seed Stability

To determine whether the reported improvements depend on a particular optimization trajectory, we repeat the complete training procedure using five random seeds while keeping the data partition, architecture, hyperparameters, and evaluation protocol fixed. Table 14 reports segmentation, metrology, recognition, and grounding performance for each run.

The variation across runs is small relative to the performance differences observed against the competing methods in the main experiments. Mask Dice varies by only 0.0034 standard deviation, while diameter MAE is 2.3650 ± 0.0619 mm. The semantic pathway is similarly stable, with Strong AUPRC of 0.7716 ± 0.0179, Macro AP of 0.5567 ± 0.0126, and localized mAP@0.5 of 0.2418 ± 0.0081.

The default seed used in the main experiments lies close to the multi-seed mean rather than representing an isolated favorable run. These results indicate that the improvements in metrology, recognition, and grounding are preserved across stochastic initialization and optimization.

E.2 Diameter Correction Sensitivity

Physical Scale Metrology restricts the learned correction around the calibrated geometric anchor using

$$\hat{d}_i^{\text{ind}} = d_i^{\text{geo}} + \delta_{\max} \tanh(\cdot), \tag{88}$$

where $\delta_{\max}$ determines how far the neural correction may deviate from explicit geometry. We vary $\delta_{\max}$ from 0 to 10 mm while keeping the remaining metrology configuration unchanged. The resulting sensitivity analysis is summarized in Table 15.

Setting $\delta_{\max} = 0$ removes neural correction entirely and increases MAE to 2.4600 mm, confirming that a limited appearance-conditioned residual is useful for compensating for discrepancies between the calibrated geometric anchor and clinician-recorded transverse diameter.

The sensitivity analysis also reveals a trade-off between typical-case accuracy and correction flexibility. Tight bounds of 1–2 mm obtain the lowest MAE and highest within-tolerance rates, whereas a substantially larger 10 mm range reduces RMSE and increases R^2 by correcting a small number

Table 15: Sensitivity to the maximum bounded diameter correction. The default configuration uses $\delta_{\max} = 3$ mm.

$\delta_{\max}$ (mm)	MAE ↓	RMSE ↓	R^2 ↑	WTR@2mm ↑	WTR@5mm ↑	Mean Corr. (mm)
0	2.4600	3.9600	0.8389	0.5766	0.8288	0.0000
1	**2.2310**	3.8100	0.8508	**0.6036**	**0.8468**	0.6800
2	2.2510	3.8300	0.8493	**0.6036**	**0.8468**	1.3700
3	**2.3299**	3.7901	0.8524	0.5946	0.8378	1.8200
5	2.2810	3.7600	0.8547	0.5766	0.8378	1.4800
10	2.4100	**3.6200**	**0.8654**	0.5495	0.8018	1.9600

Table 16: Sensitivity to the ROI expansion margin.

Margin	Mask Dice ↑	G-Dice ↑	Macro AP ↑	AUPRC ↑	Loc. mAP ↑
0%	0.8420	0.3204	0.5321	0.7418	0.2014
5%	0.8501	0.3418	0.5457	0.7546	0.2182
10%	0.8577	0.3609	0.5568	0.7724	0.2346
15%	**0.8631**	**0.3768**	**0.5644**	**0.7837**	**0.2471**
20%	0.8612	0.3706	0.5601	0.7788	0.2405
30%	0.8524	0.3487	0.5482	0.7596	0.2191

of larger errors more aggressively. However, the latter simultaneously decreases WTR@2mm from 0.5946 to 0.5495 and WTR@5mm from 0.8378 to 0.8018, indicating less consistent errors around the clinically relevant tolerance ranges.

We therefore use $\delta_{\max} = 3$ mm as a moderate correction range. It retains a tight physical constraint around the reference-calibrated anchor while achieving 2.3299 mm MAE, 3.7901 mm RMSE, $R^2 = 0.8524$, and strong within-tolerance performance. More importantly, the correction remains explicitly bounded, preserving the role of physical calibration rather than allowing the neural estimator to degenerate into unconstrained direct regression.

E.3 Spatial Configuration Sensitivity

Multiview Dermal Grounding depends on two spatial configuration parameters: the margin used to expand the geometry-derived ROI and the resolution of the grounding lattice. We analyze both parameters independently while keeping all other settings fixed.

ROI expansion margin. A tightly cropped ROI may exclude reaction-boundary evidence, whereas an excessively large crop reintroduces irrelevant surrounding skin and background. Table 16 evaluates margins between 0% and 30%.

Performance increases consistently as the margin grows from 0% to 15%. For example, G-Dice improves from 0.3204 to 0.3768 and localized mAP@0.5 from 0.2014 to 0.2471. Increasing the margin further produces a gradual decline. At 30%, G-Dice decreases to 0.3487 and localized mAP@0.5 to 0.2191.

This approximately unimodal behavior supports a moderate amount of context around the predicted induration: a 15% margin preserves boundary and nearby dermal evidence without substantially diluting the reaction-focused representation.

Spatial grounding resolution. We next vary the grounding lattice from 6×6 to 24×24. In addition to grounding and segmentation accuracy, we report peak memory consumption and inference time measured under the same evaluation environment.

As Figure 13 shows, very coarse grids lose fine spatial structure: the 6×6 configuration achieves only 0.2982 G-Dice and 0.1826 localized mAP@0.5. Increasing the resolution to 12×12 substantially improves both metrics to 0.3768 and 0.2471, respectively.

Further increasing the spatial resolution does not improve localization. At 24×24, localized mAP@0.5 decreases to 0.2206 while peak memory increases from 265 MB to 1218 MB and inference time from 78.5 ms to 131.1 ms. This suggests that excessively fine grids increase computational cost and expose the model to noisier sparse supervision without providing additional useful spatial information. We therefore use 12×12 as the default resolution.

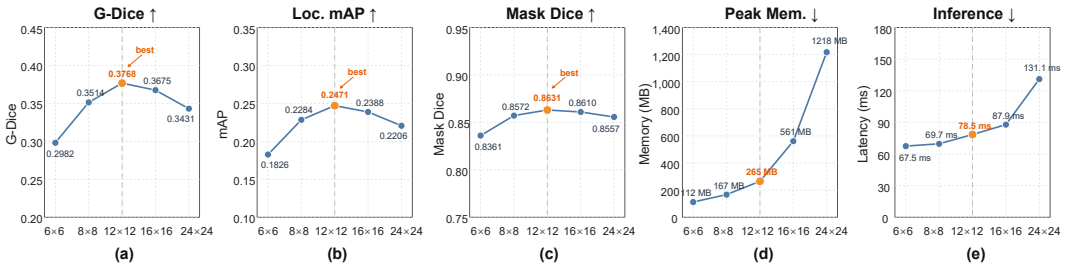


Figure 13: Sensitivity to the spatial grounding resolution. Computational measurements are reported under the same evaluation environment.

Table 17: Strong-reaction performance at representative operating points. Threshold selection changes the sensitivity–specificity trade-off but does not affect the underlying ranking metrics.

Operating Point	Threshold	Sensitivity ↑	Specificity ↑	Precision ↑	NPV ↑	F1 ↑	Youden
Max Youden	0.30	**1.0000**	0.8627	0.3913	**1.0000**	0.5625	**0.8627**
Default	0.50	0.7778	0.9118	0.4375	0.9789	0.5600	0.6895
Balanced F1	0.65	0.7778	0.9510	0.5833	0.9798	**0.6667**	0.7288
High-specificity F1	0.90	0.6667	**0.9706**	**0.6667**	0.9706	**0.6667**	0.6373

Together, the ROI-margin and grid-resolution experiments indicate that the default spatial configuration is not sensitive to a narrowly tuned point: nearby settings retain competitive performance, while the selected 15% margin and 12×12 lattice provide the strongest overall grounding and segmentation results.

E.4 Probability Calibration and Operating Points

Case-level discrimination metrics such as AUROC and AUPRC evaluate ranking quality but do not characterize whether predicted probabilities are well calibrated or how performance changes under different decision thresholds. We therefore additionally evaluate the strong-reaction output using expected calibration error (ECE), Brier score, and several representative operating points.

For B confidence bins, ECE is computed as

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left|\mathrm{acc}(S_b) - \mathrm{conf}(S_b)\right|, \tag{89}$$

where S_b contains samples assigned to bin b. The Brier score is

$$\mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_i^{\mathrm{strong}} - y_i^{\mathrm{strong}}\right)^2. \tag{90}$$

For the default model, the threshold-independent results are

$$\mathrm{AUROC} = 0.9719, \qquad \mathrm{AUPRC} = 0.7837, \tag{91}$$

with

$$\mathrm{ECE} = 0.1085, \qquad \mathrm{Brier} = 0.0690. \tag{92}$$

These results complement the ranking-based metrics by quantifying the agreement between the predicted strong-reaction probabilities and observed binary outcomes.

Table 17 reports four representative decision thresholds. Importantly, changing the threshold affects the binary operating characteristics but not AUROC, AUPRC, ECE, or Brier score computed from the underlying continuous prediction scores.

The operating-point analysis illustrates the expected trade-off between sensitivity and specificity. Lowering the threshold to 0.30 increases sensitivity to 1.0000 and produces the highest Youden index, but reduces specificity to 0.8627 and precision to 0.3913. In contrast, increasing the threshold to 0.90 raises specificity to 0.9706 and precision to 0.6667 at the cost of reducing sensitivity to 0.6667.

The intermediate threshold of 0.65 retains the same sensitivity as the default 0.50 threshold while increasing specificity from 0.9118 to 0.9510 and F1 from 0.5600 to 0.6667. These results show that the continuous strong-reaction score supports different sensitivity–specificity trade-offs without retraining the model. The default threshold is retained for standardized benchmark reporting, while alternative operating points are reported only to characterize the behavior of the predictor under different decision thresholds.

F Additional Qualitative Results

We provide extended qualitative and diagnostic results to complement the quantitative evaluation in the main paper. All visualizations are generated directly from held-out test-set inference without manual modification of model outputs. The selected examples expose the complete processing chain from physical-reference calibration and geometry-sensitive metrology to multiview semantic grounding and bounded mask refinement. All grounding maps use a fixed normalization protocol determined before test evaluation, and all mask thresholds, contour styles, transparency settings, coordinate transformations, and rendering parameters are shared across cases. The purpose of these visualizations is not merely to provide representative examples, but to make the internal information flow of MediPPD directly inspectable and to qualitatively verify the proposed principle of selective separation with controlled interaction.

F.1 Physical Metrology Visualization

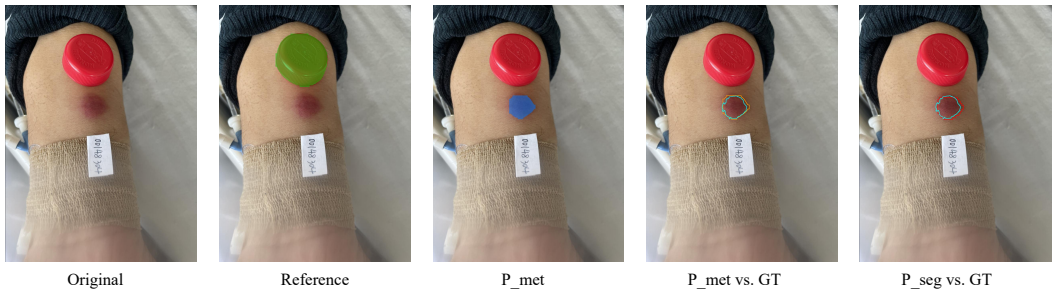
We first examine the geometry-sensitive physical-metrology pathway using four representative test cases with complementary visual characteristics. Case 001 corresponds to a typical clear reaction, Case 002 contains irregular induration geometry, Case 003 exhibits low contrast and an ambiguous boundary, and Case 004 contains stronger background or acquisition interference. These cases are intended to expose how reference calibration, metrology segmentation, and semantic refinement behave under different sources of visual and geometric difficulty.

Reference-calibrated metrology across representative conditions. Figure 14 illustrates how MediPPD transforms image-space appearance into an explicitly calibrated physical representation before semantic refinement. Reference-object localization establishes an image-specific physical scale, avoiding the unreliable use of pixel size as a proxy for real-world diameter under varying acquisition conditions. The metrology mask P^{met} then captures measurement-critical geometry and is compared with the clinician-defined induration region. Clear and irregular cases demonstrate both accurate boundary recovery and the need for bounded residual correction when raw geometry differs from the recorded transverse extent, while low-contrast and challenging-background examples show that physical calibration can remain stable despite difficult visual conditions. The final column presents the semantic-refined mask P^{seg}, which is deliberately excluded from diameter estimation. This separation ensures that P^{met} preserves metric fidelity, whereas P^{seg} can improve spatial agreement without redefining the physical measurement, supporting the design of separating calibrated geometry from semantics-sensitive refinement.

Bounded semantic refinement of frozen geometry. Figure 15 illustrates how semantic information refines the frozen metrology representation without replacing its physical geometry. The first column shows P^{met} as the geometry-sensitive prior, while the aligned multiview features provide semantic evidence for refinement. Rather than predicting a new mask from scratch, Semantic Mask Refinement learns bounded positive and negative logit residuals that locally expand or suppress the metrology mask. This residual formulation enables case-specific corrections, ranging from minor adjustments in clear reactions to broader but spatially constrained modifications for irregular or low-contrast boundaries. A geometric-preservation objective further limits excessive deviation from the original geometry. Importantly, although the refined mask P^{seg} can better match the clinician-defined boundary, the physical diameter is never recomputed from it. Semantic refinement therefore improves segmentation fidelity while preserving the established metrology pathway, realizing controlled interaction between geometry-sensitive measurement and semantics-sensitive interpretation.

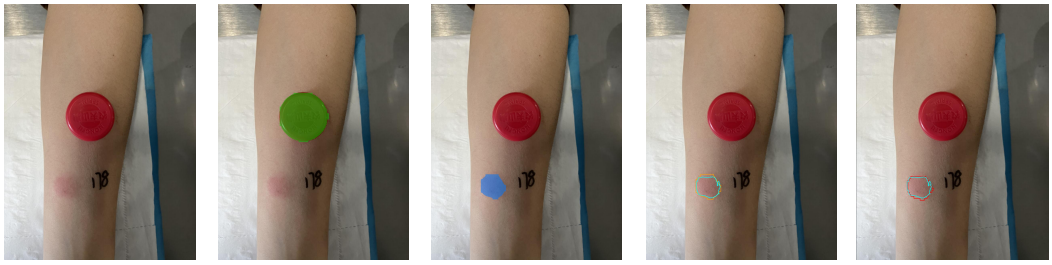
F.2 Dermal Grounding Visualization

We next examine the semantic-grounding pathway using three dedicated strong-positive test cases: Case 005 for blistering, Case 006 for necrosis, and Case 007 for double-ring reaction. Each example is selected from target-positive test samples with available spatial annotation using a fixed reproducible procedure. These cases are used to determine whether high case-level semantic confidence



Original Reference P_met P_met vs. GT P_seg vs. GT

(a) Typical Clear



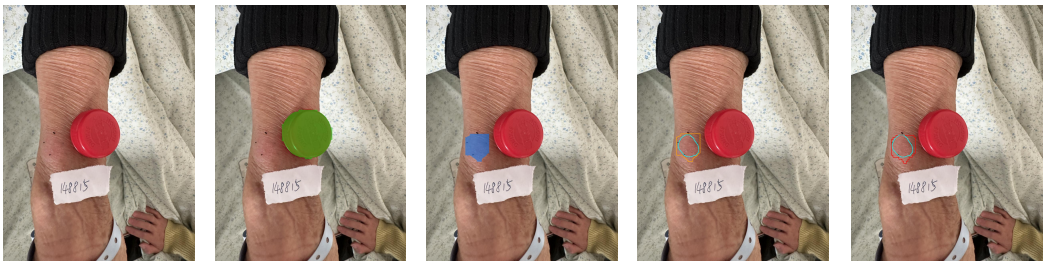
Original Reference P_met P_met vs. GT P_seg vs. GT

(b) Irregular Geometry



Original Reference P_met P_met vs. GT P_seg vs. GT

(c) Low Contrast

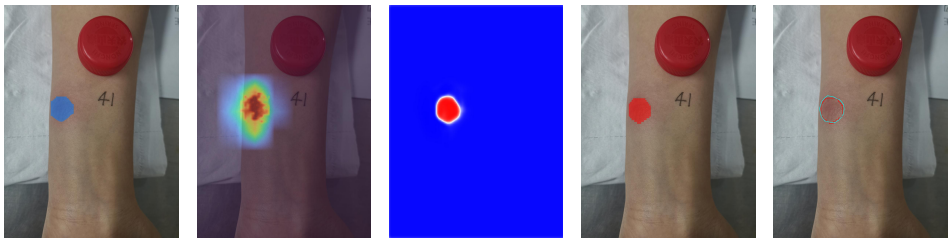


Original Reference P_met P_met vs. GT P_seg vs. GT

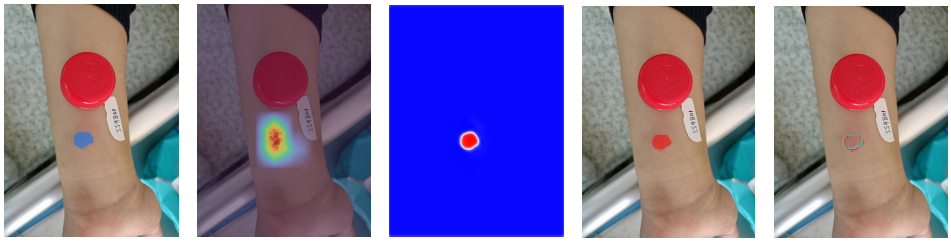
(d) Challenging Distractor

Figure 14: Physical-metrology visualizations for four representative test cases. Rows correspond to a typical clear reaction, irregular induration geometry, low-contrast boundary, and challenging background condition. From left to right: original image, physical-reference localization, geometry-sensitive metrology-mask overlay, metrology mask compared with the clinician-defined induration annotation, and semantic-refined segmentation compared with the same annotation.

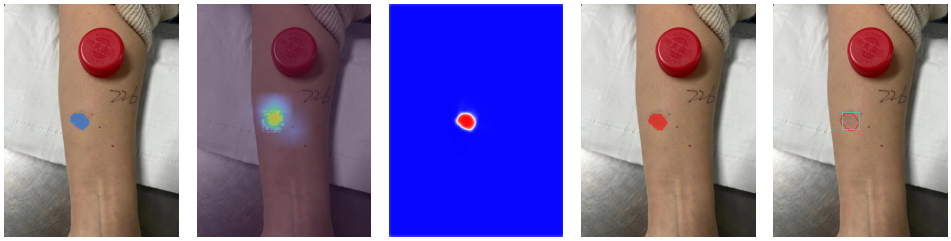
is supported by localized evidence that corresponds to the intended clinical finding rather than to a generic reaction region or irrelevant image content.



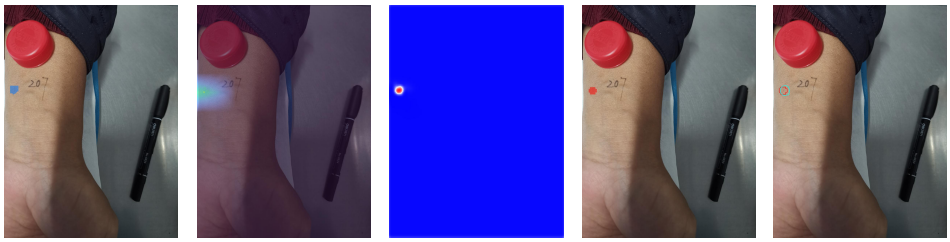
(a) Typical Clear



(b) Irregular Geometry



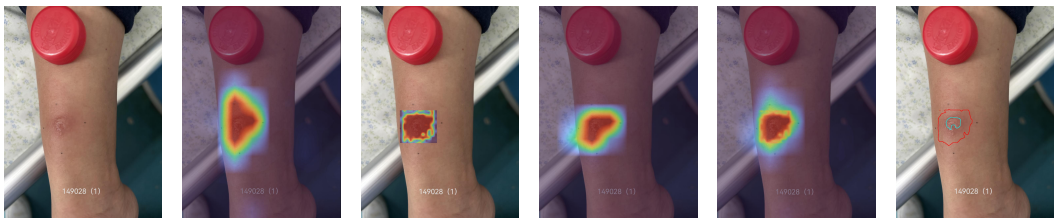
(c) Low Contrast



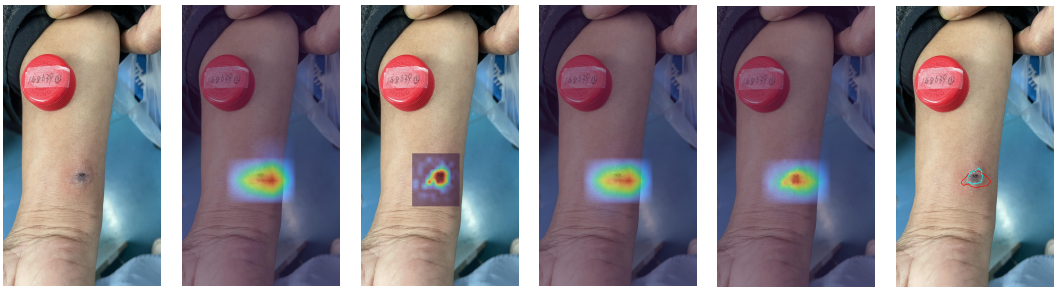
(d) Challenging Distractor

Figure 15: Qualitative visualization of Semantic Mask Refinement for the same four representative cases. From left to right: frozen metrology mask, aligned multiview grounding evidence, signed bounded refinement residual, final refined-mask overlay, and refined segmentation compared with ground truth. The residual modifies uncertain regions while preserving the geometry inherited from the metrology pathway.

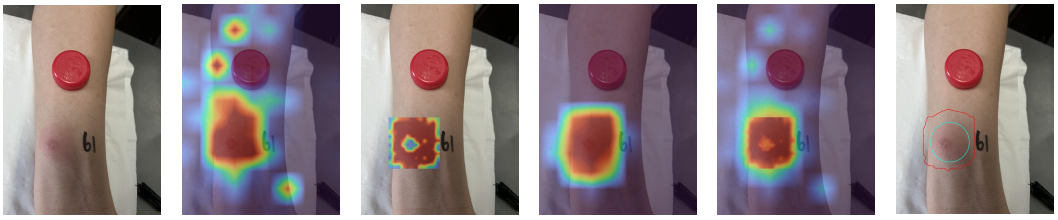
Complementary multiview evidence for strong dermal findings. Figure 16 exposes how Global, ROI, and Masked representations contribute different forms of evidence to the same semantic concept before aggregation. The Global branch preserves the complete image context and therefore retains information about the relation between a local dermal structure and the surrounding reaction, although this broader field of view also introduces more opportunities for irrelevant background structures to influence the response. The ROI branch concentrates the finite spatial representation of the frozen visual encoder around the geometry-derived reaction region, increasing the effective detail available for small blistering or necrotic structures without directly specifying where



(a) Blistering



(b) Necrosis

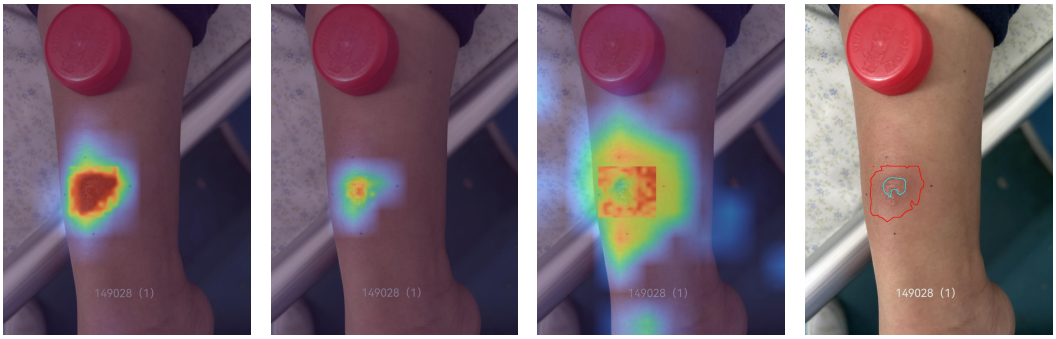


(b) Double-ring

Figure 16: Concept-specific multiview grounding for the three strong dermal findings. Rows correspond to blistering, necrosis, and double-ring reaction. From left to right: original image, target-concept grounding from the Global view, ROI view, and Masked view, multiview aggregated target evidence, and the final grounding prediction compared with the spatial annotation. All grounding maps use the same normalization protocol and are produced directly from their corresponding concept queries.

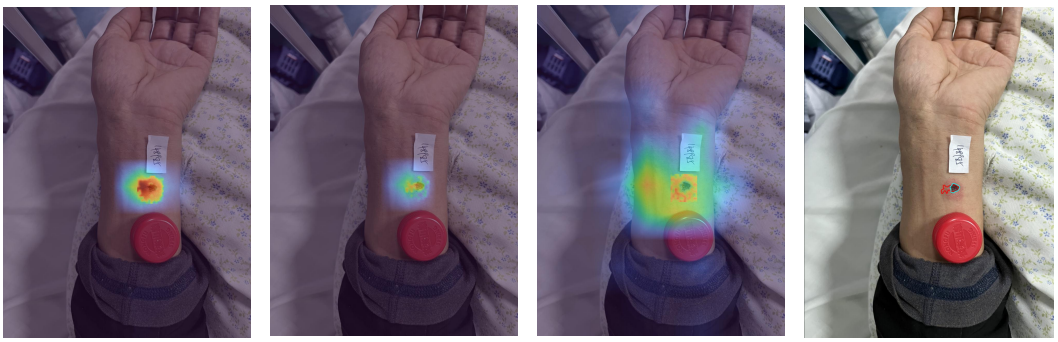
the semantic response should occur. The Masked branch introduces a complementary bias by suppressing background appearance while preserving the lesion in its original spatial context, which can reduce activation associated with labels, reference objects, borders, or other non-lesion structures. After view-specific grounding, the responses are restored to original-image coordinates and aggregated, allowing the final semantic evidence to combine coarse context, localized detail, and background suppression. The three target concepts also illustrate why this multiview construction is useful for PPD interpretation: blistering often occupies a compact region, necrosis can be concentrated in a smaller central tissue-abnormality area, whereas double-ring reaction may exhibit a more distributed annular pattern. A single generic lesion representation would therefore be unlikely to provide equally appropriate spatial support for all three. The final prediction-versus-ground-truth column directly evaluates whether the aggregated evidence overlaps the clinically annotated manifestation, addressing the Spatial Grounding Failure discussed in the main paper. Because the displayed maps are produced from actual query–patch similarity without metrology-mask clipping or ground-truth-based reshaping, spatial agreement in these examples reflects the learned grounding mechanism rather than a post-processing artifact.

Concept specificity beyond generic lesion saliency. Figure 17 evaluates semantic grounding by fixing the image while varying the queried concept. The first three columns compare blistering, necrosis, and double-ring responses under a shared normalization protocol, allowing target and non-



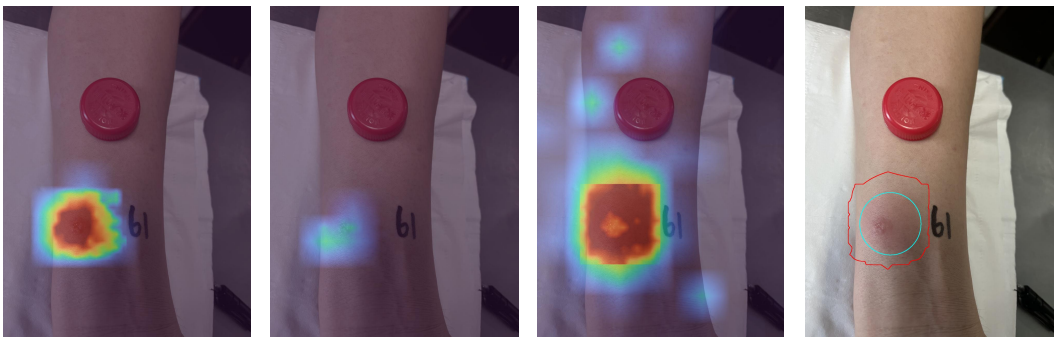
Blistering map **(Target)** | Necrosis map | Double-ring map | Blistering vs. GT

(a) Blistering



Blistering map | Necrosis map **(Target)** | Double-ring map | Necrosis vs. GT

(a) Necrosis

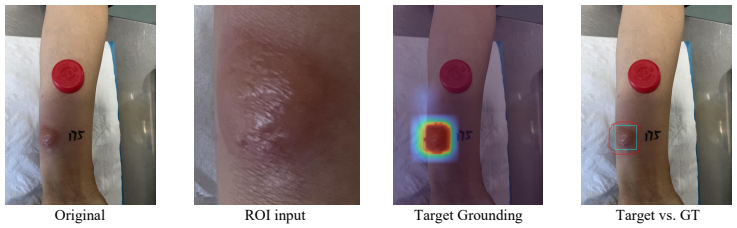


Blistering map | Necrosis map | Double-ring map **(Target)** | Double-ring vs. GT

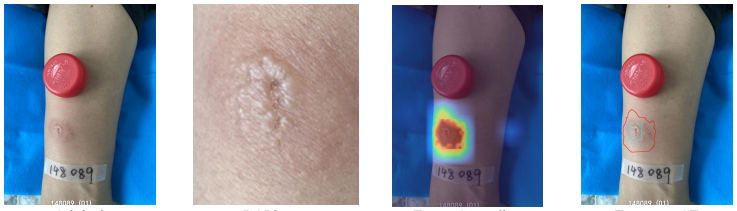
(a) Double-ring

Figure 17: Qualitative analysis of concept-conditioned grounding specificity. Rows correspond to blistering-, necrosis-, and double-ring-positive cases, while the first three columns show grounding responses to the blistering, necrosis, and double-ring queries under a shared normalization protocol. The diagonal panels correspond to the target concept for each case, whereas off-diagonal panels show non-target concept responses. The final column compares target-concept grounding with its spatial annotation.

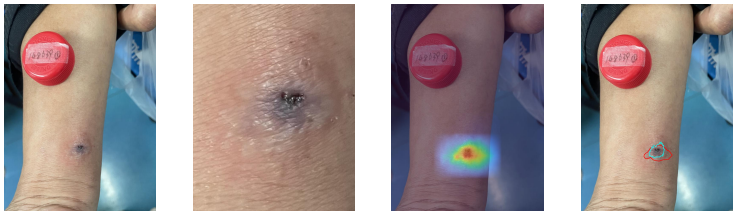
target activations to be assessed directly. If the learned queries acted only as generic lesion detectors, their spatial responses would remain largely similar across concepts. Instead, differences in response concentration, extent, and organization indicate that localization is conditioned on concept identity. Non-target maps are not expected to be completely inactive because dermal findings may share visual characteristics or coexist; the key criterion is whether the target query produces spatial evidence



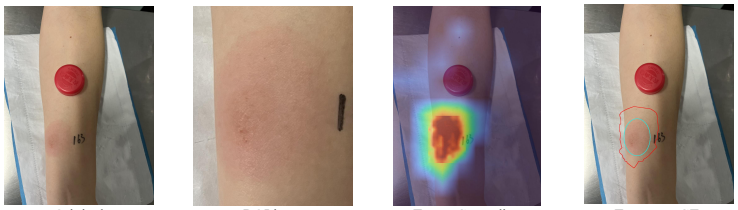
(a) Additional Blistering case



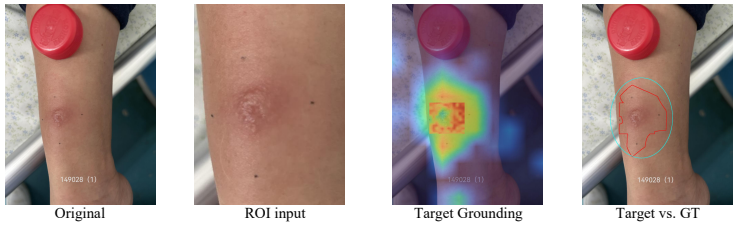
(b) Additional Blistering case



(c) Additional Necrosis case



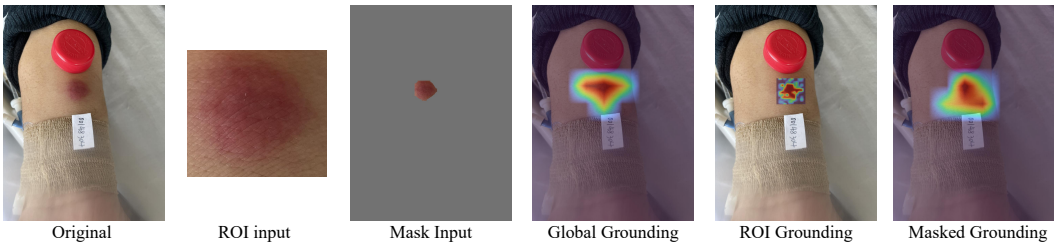
(c) Additional Double-ring case



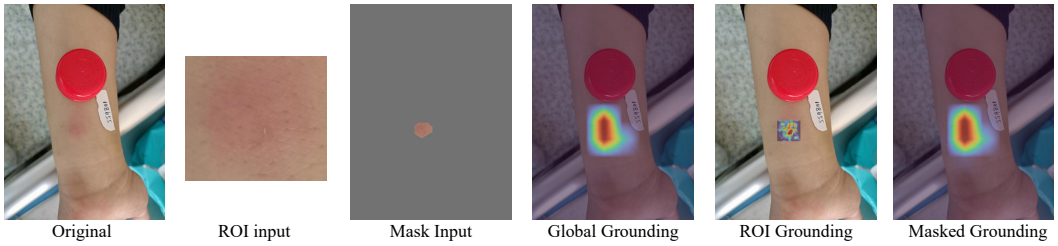
(c) Additional Double-ring case

Figure 18: Additional concept-specific grounding examples from the held-out test set. Each row corresponds to one target-positive case and shows the original image, geometry-derived ROI, target-concept spatial evidence, and grounding prediction compared with the corresponding spatial annotation. The examples are selected using the same reproducible criteria as the primary qualitative cases.

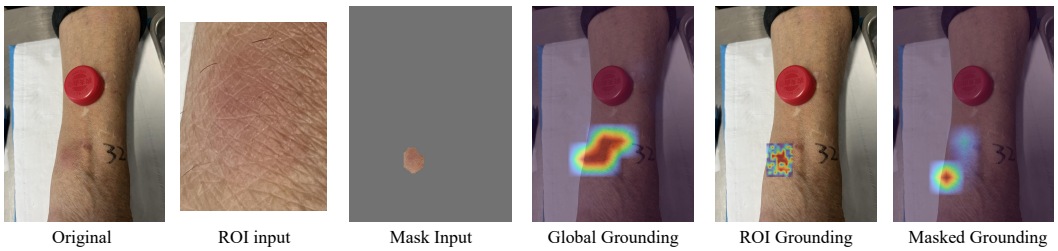
more consistent with the annotated manifestation. Shared normalization is essential because independent rescaling could exaggerate weak off-target responses. The final comparison with ground truth further links semantic specificity to spatial correctness, supporting the interpretation of $\mathbf{q}_k$ as concept-conditioned queries rather than class-agnostic saliency mechanisms.



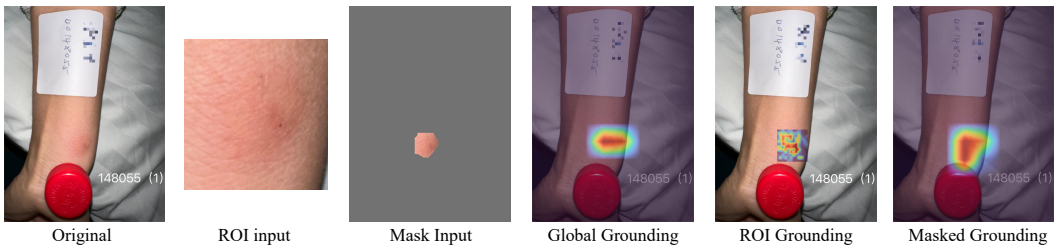
(a) Typical Clear



(b) Irregular Geometry



(c) Low Contrast



(d) Challenging Distractor

Figure 19: Additional diagnostics of multiview inputs and view-specific grounding responses. Rows correspond to Cases 001–004, covering a typical clear reaction, irregular geometry, low contrast, and a challenging background condition. From left to right: original image, geometry-derived ROI input, background-suppressed Masked input, Global-view grounding, ROI-view grounding, and Masked-view grounding for the same selected semantic concept. All three grounding maps within each case use the same normalization protocol.

Generalization of localized evidence across additional positive cases. Figure 18 broadens the qualitative analysis beyond the three canonical strong-positive examples and reduces dependence on a small number of visually favorable cases. The additional samples expose variation in lesion size, target morphology, local contrast, position within the reaction, and surrounding skin appearance while retaining the same inference and rendering protocol. Showing the geometry-derived ROI beside the final target grounding is particularly informative because it distinguishes spatial guidance from semantic localization: the ROI determines the region in which the model receives higher-resolution visual evidence, but it does not prescribe the location or shape of the final semantic response. A meaningful grounding map should therefore remain substantially more selective than the crop itself and should respond preferentially to structures associated with the queried find-

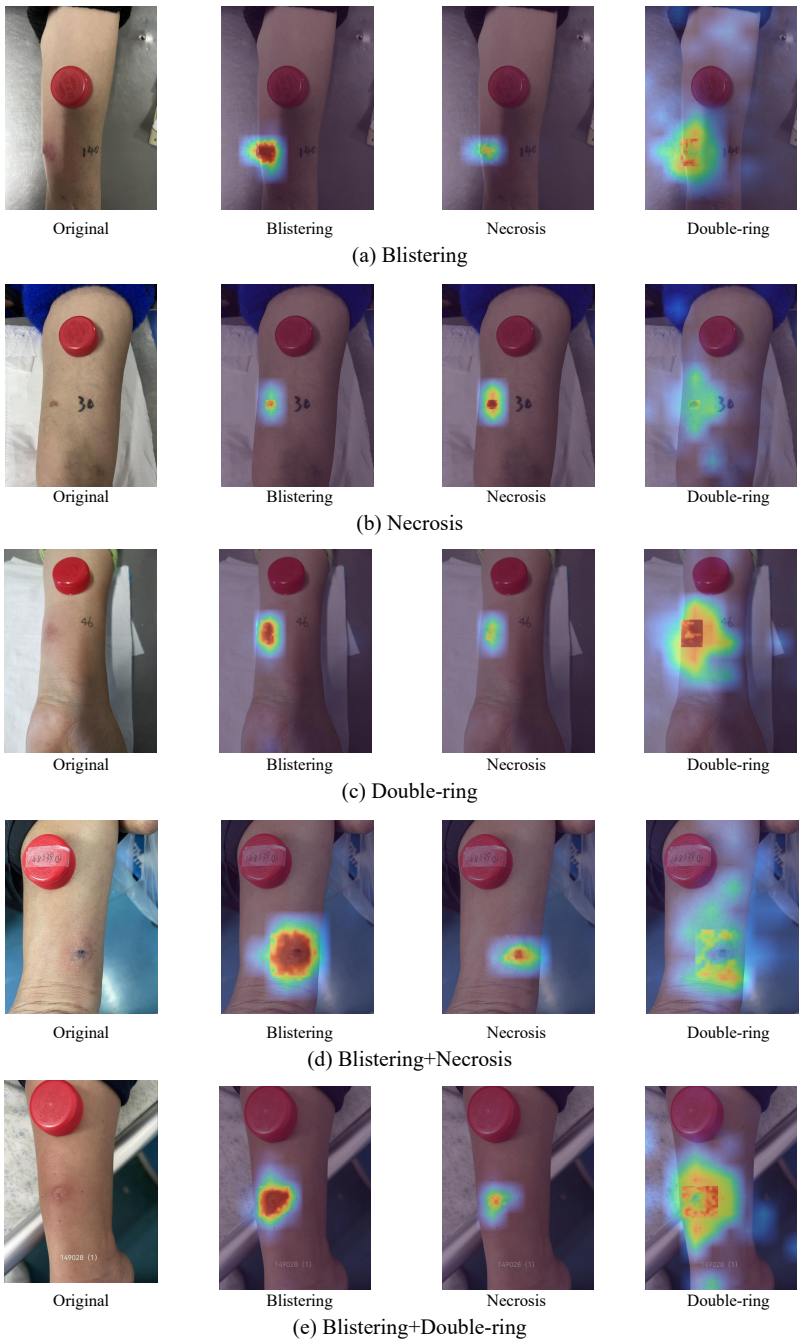


Figure 20: Cross-concept grounding responses in additional held-out test cases. From left to right: original image and grounding responses produced by the blistering, necrosis, and double-ring queries. Rows include single-finding and, where available, co-occurring strong-finding cases under diverse visual conditions. All three concept maps within each case use identical normalization and opacity, allowing their spatial response patterns to be compared directly.

ing rather than uniformly covering the entire reaction region. The final column further evaluates whether this selective evidence corresponds to the spatial annotation under different visual conditions. Importantly, the same normalization, opacity, interpolation, and restoration procedure is used for every example, with no per-case threshold adjustment or manual masking of the heatmap. This fixed protocol prevents difficult cases from being visually enhanced independently and makes apparent changes in grounding strength or compactness more interpretable. These examples should be

viewed together with the quantitative Rare G-Dice and localized mAP@0.5 results: the metrics summarize the average behavior across the test set, while the qualitative panels show what successful, partial, or spatially diffuse localization actually looks like in individual images. The consistency of the grounding pattern across heterogeneous positive cases therefore supports the claim that localized semantic evidence is not restricted to a single curated example or a single visual morphology.

F.3 Additional Visual Diagnostics

The previous visualizations emphasize final metrology, refinement, and target grounding outputs. We additionally expose intermediate multiview representations and cross-concept responses to inspect whether the observed performance is consistent with the intended architecture rather than being produced by hidden shortcut behavior.

Diagnostic interpretation of view construction and view-specific grounding. Figure 19 visualizes both the inputs to the ROI and Masked branches and the grounding responses from Global, ROI, and Masked representations, making their complementarity directly inspectable. The Global view preserves full contextual information but may also attend to irrelevant structures such as reference objects or background regions. The ROI view enlarges the reaction area and emphasizes finer local patterns, while the Masked view suppresses foreground–background competition by attenuating non-lesion appearance. Their heatmaps therefore exhibit different spatial concentration and sensitivity to background activation, reflecting distinct representational biases rather than duplicated signals. The figure also serves as a qualitative shortcut audit by revealing undesirable attention to reference objects or diffuse responses. Conversely, consistent localization across ROI and Masked views with reduced irrelevant activation supports the use of geometry as spatial guidance rather than semantic supervision. These observations provide a mechanistic explanation for the quantitative gains of combining Global, ROI, and Masked representations.

Cross-concept behavior under single and co-occurring findings. Figure 20 complements the multiview analysis by fixing the image while varying the semantic concept. Unlike Fig. 19, which compares visual representations for the same concept, this figure examines how identical image evidence is interpreted by blistering, necrosis, and double-ring queries. Single-positive cases test whether the annotated concept produces the most spatially compatible response, while co-occurring cases assess whether multiple valid findings retain distinct spatial organization rather than collapsing into a generic reaction map. Shared normalization and opacity are essential for preserving relative response magnitude across concepts. Together with Fig. 17, these results show that grounding varies with both visual representation and clinical concept, supporting the interpretation that MediPPD learns localized, concept-dependent semantic evidence rather than a class-agnostic saliency map.

Overall, the qualitative results provide a direct visual counterpart to the quantitative findings in the main paper. The physical-metrology examples show that measurement remains traceable to an explicit image-specific reference and a dedicated geometry-sensitive representation; the Semantic Mask Refinement examples show that semantic information is introduced through a bounded correction rather than by replacing this physical pathway; the target-specific grounding examples demonstrate how Global, ROI, and Masked representations contribute complementary localized evidence; and the concept-comparison diagnostics show that the learned spatial responses change with the queried dermal finding rather than collapsing to a generic reaction map. Taken together, these observations support the central MediPPD principle of selective separation with controlled interaction: measurement-critical geometry is explicitly protected, geometry is allowed to guide semantic reasoning where spatial context is useful, and semantic evidence can improve interpretation and segmentation without modifying the reported physical measurement.

AI use statement

In this work, we used generative AI tools for language polishing and for identifying potential formatting and presentation issues in the manuscript. We have not used generative AI tools for research design, experimental execution, data generation, or result interpretation, and other required disclosure tasks are not applicable to this work. We reviewed and verified all AI-assisted revisions to ensure that they preserved the intended technical meaning and complied with the manuscript formatting requirements. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.