MedLoc-R1: Performance-Aware Curriculum Reward Scheduling for GRPO-Based Medical Visual Grounding

Supplementary Material

A. Datasets Details

We provide detailed statistics of the dataset splits for the three benchmarks used in our experiments: **HAM10000**, **HEEL**, and **TN3K**.

- **HAM10000** is a public dataset containing 10,015 dermoscopic images released in 2018 by the Medical University of Vienna. It is one of the most important benchmark datasets in the field of automatic skin cancer detection and classification. It covers 7 common skin lesion types: actinic keratoses (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (vasc). We derived bounding boxes from the provided lesion masks and randomly split them into training and test sets with 8,012 training and 2,003 testing samples across 7 skin lesion categories.
- **HEEL** is a public dataset of 3,956 lateral foot X-ray images collected at Kirkuk General Hospital, comprising three diagnostic categories: `Normal` (1,842 images), `Heel Spur` (1,316 images), and `Sever's disease` (798 images). All images are labeled by orthopedic specialists and cross-validated by radiologists. Following the original protocol, we use 3,164 images for training and 792 for testing, while preserving the class distribution across the three categories.
- **TN3K** is an open-access thyroid nodule ultrasound dataset comprising 3,493 B-mode images from 2,421 patients, each annotated with pixel-wise nodule masks. The official split contains 2,879 training and 614 test images. In our experiments, we derive a binary classification subset with 2,655 training and 544 testing samples, labeled as `malignant` and `benign` cases.

To offer a clearer overview of their composition, Table 5 summarizes the class-wise training and testing splits for all three datasets.

Table 5. Class-wise number of training and test samples for the three benchmarks.

Dataset	Class	Train	Test
HAM10000	nv	5367	1338
	mel	900	213
	bkl	868	231
	bcc	418	96
	akiec	253	74
	vasc	118	24
	df	88	27
	Total	8012	2003
HEEL	Normal	1473	369
	Heel Spur	1053	263
	Severe	638	160
	Total	3164	792
TN3K	malignant	932	219
	benign	1723	331
	Total	2655	544

B. Experimental Configuration Details

We elaborate here on the configuration details of the threshold scheduling strategies evaluated in Table 3. All strategies manipulate three key scheduling parameters: the step size δ_k, the performance percentile P_k, and the stability margin S_k. These parameters govern how the threshold τ_k is updated during training.

Adaptive (Ours). The adaptive strategy **dynamically** adjusts δ_k, P_k, and S_k according to the current threshold value τ_k throughout training. It operates in three regimes:

- When $\tau_k < 0.60$, a large step size $\delta_k = 0.15$ is used to encourage rapid threshold progression, with $P_k = 0.80$ and $S_k = 0.20$.
- When $0.60 \leq \tau_k < 0.75$, we moderate the update using $\delta_k = 0.10$, $P_k = 0.75$, and $S_k = 0.35$.
- When $\tau_k \geq 0.80$, updates become conservative with $\delta_k = 0.05$, $P_k = 0.55$, and $S_k = 0.40$.

This staged configuration allows the model to explore aggressively in early training while stabilizing and refining predictions in later stages, leading to more robust convergence behavior.

Fixed-Aggressive. This strategy uses fixed values across the entire training process, with a large step size $\delta_k = 0.15$, a relatively low performance percentile $P_k = 0.60$, and a small stability margin $S_k = 0.40$. The configuration encourages fast threshold updates with minimal stability constraints, favoring aggressive adaptation dynamics.

Fixed-Moderate. The moderate variant sets intermediate values: $\delta_k = 0.10$, $P_k = 0.70$, and $S_k = 0.25$. It aims to balance adaptation speed and stability, representing a middle ground between aggressive and conservative strategies.

Table 6. Ablation of **piecewise decay** schedule parameters.

Dataset	δ_1	δ_2	β_1	β_2	A@0.5
HAM (0.3–0.8)	0.10	0.05	0.55	0.70	92.23
	0.15	0.10	0.55	0.75	**94.26**
	0.20	0.15	0.60	0.75	93.97
	0.25	0.20	0.60	0.75	93.10
HEEL (0.3–0.8)	0.10	0.05	0.50	0.70	93.38
	0.15	0.10	0.50	0.75	**94.19**
	0.20	0.15	0.60	0.75	94.11
	0.25	0.20	0.60	0.75	93.87
TN3K (0.3–0.8)	0.10	0.05	0.50	0.70	65.97
	0.15	0.10	0.50	0.75	66.18
	0.20	0.15	0.60	0.75	**66.31**
	0.25	0.20	0.60	0.75	65.56

Fixed-Conservative. This configuration employs a small step size $\delta_k = 0.05$, high performance requirement $P_k = 0.80$, and a tight stability margin $S_k = 0.15$. These constraints slow down the threshold adaptation process, ensuring greater caution and smoother updates throughout training.

All fixed strategies retain constant values for all three parameters, while the adaptive strategy transitions between configurations in a stage-wise manner depending on τ_k. This dynamic scheduling is a key factor contributing to the superior performance of our method.

C. More results on Piecewise Decay Schedule

To assess how the hyperparameters of the piecewise decay schedule influence RL-based localization, we perform an ablation over the stage-wise step sizes $(\delta^{(1)}, \delta^{(2)}, \delta^{(3)})$ and the boundary thresholds $(\beta^{(1)}, \beta^{(2)})$. In all settings, the adaptive IoU threshold is initialized at τ_0=0.3 and driven toward τ_{target}=0.8, while the schedule parameters control *how fast and in what shape* this transition occurs. Our main results adopt the configuration $\delta^{(1)}$=0.15, $\delta^{(2)}$=0.10, $\delta^{(3)}$=0.05 and $\beta^{(1)}$=0.55, $\beta^{(2)}$=0.75; Table 6 reports additional piecewise configurations evaluated on HAM10000, HEEL, and TN3K. The default choice consistently achieves the best or near-best A@0.5, whereas both more aggressive and flatter step-size patterns lead to noticeable drops in performance. These results indicate that the schedule is not overly sensitive within a reasonable range, but benefits from *moderate early-stage increments followed by gentler late-stage refinement*, which provides a stable progression of τ_k and alleviates reward sparsity during training.

D. Additional baseline results

To provide a stronger empirical comparison, we additionally included three external baselines from recent literature.

Table 7. Additional baseline results (**left**) and main ablation study on group size G (**right**).

Method	HAM10000	HEEL	TN3K
GroundingDINO-L	84.27	83.61	32.70
BoxMed-RL	92.36	74.51	56.39
MedGround-R1	88.23	89.52	51.43
MedLoc-R1 (Ours)	**94.46**	**94.19**	**66.18**

Group size	HAM10k
4	89.10
6	93.51
8	**94.46**
10	94.36

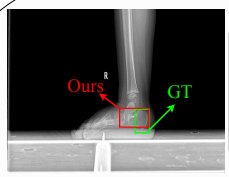
GroundingDINO-L [14] is a state-of-the-art open-set object detector, which we fine-tuned on our training set and used the highest-confidence predicted box as the final prediction. **BoxMed-RL** [10], originally developed for radiology report generation, adopts GRPO-based Spatially Verifiable Reinforcement (SVR) to align medical findings with bounding boxes on sentence-box aligned datasets. Its IoU-based reward, defined as IoU when IoU $>$ 0 and 0 otherwise, is effectively equivalent to our Raw-IoU baseline. We therefore reproduced its SVR framework in our bbox-only setting using instruction prompts of the form "Provide the bounding box for {target}." **MedGround-R1** [34] is a recent GRPO-based medical grounding method that combines spatial accuracy and semantic consistency in the reward design, together with a Chain-of-Box reasoning template. As shown in Table 7 (left), MedLoc-R1 consistently outperforms all three external baselines across the three datasets in terms of A@0.5, further validating the effectiveness of our performance-aware curriculum reward scheduling. In addition, Table 7 (right) reports a grid search over the key GRPO hyperparameter, namely the group size G.

E. Failure Case Visualizations

We present several representative failure cases of MedLoc-R1 in Figure 8. Although the predicted bounding boxes are not always perfectly aligned with the ground truth, they are generally centered on the correct target regions, indicating that the model captures the key visual and semantic cues required for medical grounding.

On HEEL, MedLoc-R1 localizes the calcaneus region with good anatomical consistency. In failure cases, the predicted boxes may slightly under-cover or over-cover the annotated region, but they still focus on the correct bone structure while excluding most irrelevant areas. On TN3K, some predictions extend beyond the nodule boundary or miss subtle margins, yet the boxes remain centered on the relevant thyroid nodule region, suggesting that the model effectively exploits both spatial and intensity cues despite the ambiguity of ultrasound images. On HAM10000, although the predicted boxes may omit faint peripheral areas or include limited surrounding healthy skin, they generally cover the diagnostically important part of the lesion.

Overall, these examples suggest that the primary failure mode of MedLoc-R1 lies in spatial imprecision rather than incorrect target identification. Even when the localization is not exact, the model usually attends to the appropriate

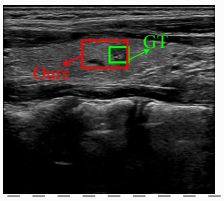


Given task:Identify the X-ray image and provide the bounding box for the calcaneus region. Ensure the bounding box accurately covers the calcaneus area and does not include too much unrelated areas. Output the bounding box coordinate in the format [xmin, ymin, xmax, ymax]. Also generate your thinking process on how you determined the bounding box.

Model Response:

To determine the bounding box for the calcaneus region in the X-ray image, I focused on identifying the anatomical structure that corresponds to the calcaneus. The calcaneus is the bone at the heel of the foot, which appears as a dense, vertically oriented structure in the X-ray. Here's the thought process: 1. **Identify the Calcaneus**: The calcaneus is the dense, vertical bone at the heel of the foot. 2. **Locate the Bounding Box**: I carefully examined the X-ray to find the calcaneus. 3. **Ensure Accuracy**: I made sure the bounding box included only the calcaneus and excluded any other parts of the foot or unrelated areas. The calcaneus is located near the bottom of the X-ray, spanning from the middle to the lower part of the image. Therefore, the bounding box coordinates for the calcaneus region are approximately: [370, 325, 470, 390]. This coordinates ensure the bounding box accurately covers the calcaneus area without including unnecessary parts of the foot.

Ground_truth: [421, 328, 471, 409]

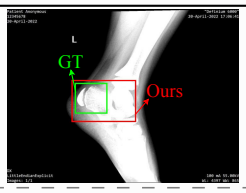


Given task: Identify the thyroid ultrasound image and provide the bounding box for any thyroid nodule present. Ensure the bounding box accurately covers the nodule area and does not include too much unrelated tissue. Output the bounding box coordinates in the format [x_{min}, y_{min}, x_{max}, y_{max}]. Generate your thinking process on how you determined the bounding box.

Model Response:

To determine the bounding box for the thyroid nodule in the ultrasound image, I carefully examined the grayscale ultrasound image to identify the nodule's location. The nodule appears as an area of increased echogenicity (brighter region) within the thyroid gland. I ensured that the bounding box was centered around this bright area without including any surrounding tissue that might be unrelated to the nodule. The thyroid nodule is located centrally within the thyroid gland, appearing as a distinct, slightly elevated area with higher echogenicity compared to the surrounding tissue. To ensure accuracy, I focused on identifying the nodule's boundaries clearly. Here are the bounding box coordinates for the thyroid nodule:[148, 65, 237, 120]. This bounding box covers the central area where the nodule is visible, ensuring it includes only the relevant tissue and excludes unnecessary parts of the image.

Ground_truth: [200, 78, 239, 110]

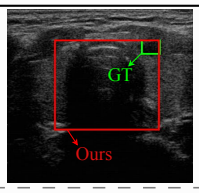


Given task:Identify the X-ray image and provide the bounding box for the calcaneus region. Ensure the bounding box accurately covers the calcaneus area and does not include too much unrelated areas. Output the bounding box coordinate in the format [xmin, ymin, xmax, ymax]. Also generate your thinking process on how you determined the bounding box.

Model Response:

To determine the bounding box for the calcaneus region in the X-ray image, I focused on identifying the anatomical structure that corresponds to the calcaneus. The calcaneus is the bone at the heel of the foot, which appears as a dense, vertically oriented structure in the X-ray. Here's the thought process:
1. **Identify the Bone**: The calcaneus is the dense, vertical bone structure located at the heel. 2. **Locate the Calcaneus**: In the X-ray, the calcaneus is identifiable by its position and density. 3.**Determine the Bounding Box**: The bounding box should encompass the entire calcaneus, ensuring it includes all parts of the bone without including unnecessary areas like the surrounding soft tissues or other bones. Based on these steps, the bounding box coordinates for the calcaneus region are approximately:[200, 220, 400, 350]. This coordinate range captures the main body of the calcaneus while excluding any irrelevant areas such as the toes or adjacent bones.

Ground_truth: [210, 229, 311, 324]

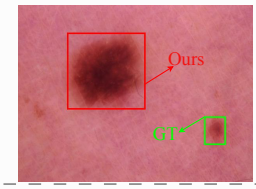


Given task: Identify the thyroid ultrasound image and provide the bounding box for any thyroid nodule present. Ensure the bounding box accurately covers the nodule area and does not include too much unrelated tissue. Output the bounding box coordinates in the format [x_{min}, y_{min}, x_{max}, y_{max}]. Generate your thinking process on how you determined the bounding box.

Model Response:

To determine the bounding box for the thyroid nodule in the ultrasound image, I carefully examined the grayscale ultrasound image to identify the nodule's location. The nodule appears as an area of increased echogenicity (brighter region) within the thyroid gland. I ensured that the bounding box was centered around this bright area without including any surrounding tissue that might be unrelated to the nodule.After careful analysis, I identified the following bounding box coordinates for the thyroid nodule:[105, 68, 340, 270]. This bounding box covers the central area of the bright region, which is likely the thyroid nodule. It excludes the surrounding tissue to ensure accuracy.

Ground_truth: [299, 68, 342, 101]

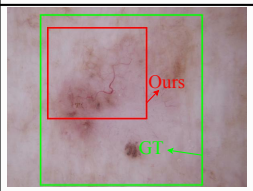


Given task:Identify the image and provide the bounding box for the skin lesion area. Ensure the bounding box accurately covers the lesion area and does not include excessive healthy skin. The output should be four numbers representing the coordinates of the top-left and bottom-right corners of the lesion area, in the format [x_{min}, y_{min}, x_{max}, y_{max}]. Also, generate your thinking process on how you determined the bounding box.

Model Response:

To determine the bounding box for the skin lesion area, I carefully examined the image to identify the region that appears different from the surrounding skin. The lesion is characterized by its irregular shape and the presence of a central dark brown spot with lighter surrounding areas. I identified the top-left corner of the lesion as the point where the darker color begins to appear. The bottom-right corner was determined by the edge of the lesion where it transitions into the surrounding skin. The coordinates for the bounding box were calculated based on these observations:- Top-left corner: The darkest part of the lesion. - Bottom-right corner: The transition between the lesion and the surrounding skin. Therefore, the bounding box coordinates for the skin lesion area are approximately [100, 150, 350, 300]. This approach ensures that the bounding box accurately captures the lesion while excluding any excessive healthy skin around it.

Ground_truth: [475, 283, 530, 355]



Given task:Identify the image and provide the bounding box for the skin lesion area. Ensure the bounding box accurately covers the lesion area and does not include excessive healthy skin. The output should be four numbers representing the coordinates of the top-left and bottom-right corners of the lesion area, in the format [x_{min}, y_{min}, x_{max}, y_{max}]. Also, generate your thinking process on how you determined the bounding box.

Model Response:

To determine the bounding box for the skin lesion area, I carefully examined the image to identify the most prominent and distinct part of the lesion. The lesion appears as a darker, irregularly shaped area with some lighter spots around it. I ensured that the bounding box captures this distinct feature without including any excessive healthy skin. The coordinates for the bounding box are as follows: - Top-left corner: [100, 50].- Bottom-right corner: [350, 280].Therefore, the bounding box coordinates for the skin lesion area are approximately [100, 50, 350, 280]. This bounding box accurately encompasses the lesion while excluding the surrounding healthy skin.

Ground_truth: [82, 18, 489, 446]

Figure 8. Visualization of representative failure cases from our method on the HEEL, TN3K, and HAM10000 datasets. Red boxes denote the predicted bounding boxes by our MedLoc-R1 method, while green boxes indicate the ground truth.

anatomical or pathological region, further supporting its effectiveness in medical visual grounding.