Supplementary for Seeing the Trees for the Forest: Rethinking Weakly-Supervised Medical Visual Grounding

A. Self-enhancement with DAP

The interpretability map Φ shows a reasonable localization capability as it achieves 33.6% average dice score on MS-CXR. Furthermore, Φ helps enhance the model by narrowing down the pathological search space for VG. By dampening the influence of background, the model can discover finer pathological signals, which get dominated without the initial interpretability map prompting. Fig. 1 shows that our model further refines the localization ofter Φ narrows down the search space.

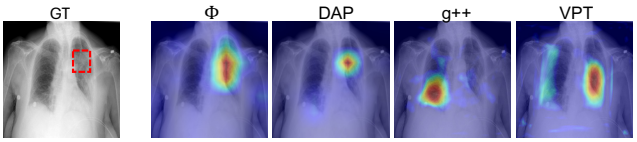

Figure 1. Results of interpretability map, our DAP, and others.

Fig. 2 (Left) plots the Dice score of our model against Φ on RSNA. Most points lie above the red line, showing our model imprives Dice upon the interpretability map, indicating self-enhancement. Fig. 2 (Right) shows DAP surpasses baselines even when Φ fails (Dice < 0.3).

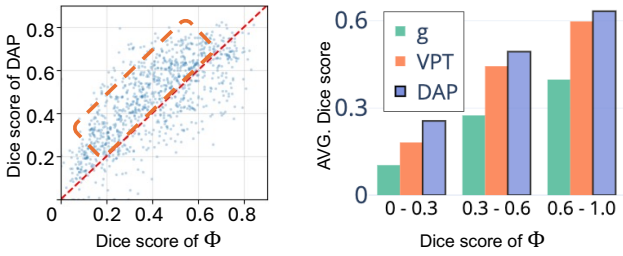

Figure 2. *Left:* Relationship between the performance of the interpretability map and our DAP. Points above the red line (highlighted in orange) indicate *self-enhancement* cases, where our DAP further improves the Dice score of interpretability maps. *Right:* Avg. Dice on samples grouped by the quality of the interpretability map. Dice of other VG are shown for comparisons.

B. Hyperparameters tuning

In this section, we study the impact of different hyperparameter settings, including the batch size, learning rate, the weights of loss objectives, text prompts variety and prompt depth. We conduct the hyperparamters tuning on 20% of the MS-CXR [1] dataset.

B.1. Batch Size and Learning Rate

We vary the batch size of $\{128, 256, 512\}$ and report the result in Table. 1. It is shown that larger batch size has a positive impact on the overall performance.

Table 1. Batch size versus performance.

batch size	CNR↑	PG↑	Dice↑
128	1.027	0.443	0.340
256	1.038	0.440	0.347
512	1.042	0.449	0.350

We vary the learning rate of $\{1e^{-1}, 1e^{-2}, 1e^{-3}\}$ and report the result in Table. 2. The optimal learning rate is $1e^{-3}$. We set batch size to 512 and learning rate to $1e^{-3}$ for other experiments.

Table 2. Learning Rate versus performance.

lr	CNR↑	PG↑	Dice↑
$1e^{-1}$	0.917	0.358	0.288
$1e^{-2}$	1.054	0.425	0.345
$1e^{-3}$	1.042	0.449	0.350

B.2. Loss weights

We vary the weights of the Disease-aware Global contrastive loss $\mathcal{L}_{glb}$ and the local contrastive loss $\mathcal{L}_{lcl}$ to evaluate their significance. We set their weights to $\{0.1, 1, 2\}$ and report the result in Table. 3. It is demonstrated that setting the weights for $\mathcal{L}_{glb}$ to 1 and $\mathcal{L}_{lcl}$ to 0.1 achieves the best overall performance, yielding the highest dice score of

0.350, PG of 0.449, and CNR of 1.042. Notably, increasing the weight for $\mathcal{L}_{lcl}$ to 1 or 2 leads to a reduction in the dice score and slight degradation in PG and CNR, suggesting that overemphasizing localization introduces diminishing returns. $\mathcal{L}_{glb}$ significantly influences segmentation performance, while localization loss has a lesser impact.

Table 3. Loss weights versus performance.

$\mathcal{L}_{glb}$	$\mathcal{L}_{lcl}$	CNR↑	PG↑	Dice↑
1	0.1	1.042	0.449	0.350
1	1	1.030	0.445	0.343
1	2	1.036	0.438	0.344
2	0.1	1.036	0.443	0.341
2	1	1.042	0.450	0.344
0.1	1	1.038	0.436	0.343
0.1	2	1.041	0.446	0.345

B.3. Robustness with other interpretability methods.

Chefer et al.'s [2] method was chosen for its strong track record in VG tasks and alignment with the bi-modal nature of VLMs, which is often used in previous VG works in VPT, g, and g++. To assess robustness, we implemented DAP with different methods on RSNA and COVID datasets with results in Tab. 4. It shows that GradCAM and SmoothGrad closely match the original performance, while LRP achieves similar scores to Self-EQ (CVPR2024). As such, DAP generalizes to other interpretability methods but we empirically found that [2] gives optimal performance.

Table 4. CNR scores of different interpretability methods.

	Chefer et al[8]	GradCAM	SmoothGrad	LRP	Self-EQ
RSNA	**1.630**	1.462	1.507	0.973	1.075
COVID	**1.018**	0.903	0.891	0.612	0.659

B.4. Prompt depth

We investigate the impact of prompting at different layers within the vision encoder. Specifically, we evaluate prompting at the pixel space of the original image, the first half of the encoder, the full encoder, the second half, and exclusively at the last layer. We report the result in Table. 5. The "last layer" setup achieves the best trade-off between segmentation accuracy and robustness, offering a strong balance between dice and CNR. In contrast, configurations like "full" or "last half" slightly enhance CNR but compromise dice, underscoring the effectiveness of focusing on deeper features for balanced performance. In contrast, the "first only" and "first half" configurations underperform.

Table 5. Disease aware prompting layer depth versus performance.

layer	CNR↑	PG↑	Dice↑
first layer	1.037	0.443	0.341
first half	1.039	0.442	0.343
full	1.043	0.450	0.343
last half	1.046	0.443	0.344
last layer	1.042	0.449	0.350

B.5. Fixed versus Varied text Prompts

We investigate textual prompting strategies by exploring one fixed prompt, and multiple paraphrases per disease class. The settings range from 1 prompt per class to 20 and 50 prompts per class, with results summarized in Table. 6. The findings demonstrate that increasing the number of prompts per class enhances model performance, likely due to expanded vocabulary exposure, which improves the robustness of the text model.

Table 6. Number of text prompts per class versus performance.

# prompts	CNR↑	PG↑	Dice↑
1	0.991	0.417	0.334
10	1.037	0.443	0.344
20	1.036	0.448	0.349
50	1.042	0.449	0.350

C. Few-shot Supervised finetuning performance

We further evaluate the proposed approach in few-shot settings, starting with weakly-supervised training followed by 20-shot fine-tuning using ground truth dense labels. As shown in Table. 7, our proposed DAP achieves results under weakly-supervised settings that are on par with the 20-shot fine-tuned performance of competing methods, demonstrating its efficiency in learning meaningful representations with minimal supervision for visual grounding.

D. Text prompts construction

We utilize GPT-4o [4] to generate disease-centric descriptions for chest X-ray findings. The model is instructed to produce 50 distinct prompts for each disease, explicitly avoiding anatomical location details. The prompt we used is:

> For the disease `<disease name>`, provide a sentence to describe it on chest X-rays. Exclude any reference to anatomical locations and ensure findings are concise, medically accurate, and reflect a professional radiology reporting style.

Method	Venue	CNR↑		PG↑		Dice↑	
		weak	20-shot	weak	20-shot	weak	20-shot
g [6]	NIPS22	1.461	1.606	0.284	0.782	0.450	0.512
g++ [5]	CVPR23	1.350	1.625	0.467	0.733	0.445	0.456
VPT [3]	ICASSP24	1.173	1.599	0.612	0.785	0.468	0.535
DAP	Proposed	1.630	1.741	0.747	0.816	0.474	0.551

Table 7. Weakly-supervised and Few-shots finetuning performance on RSNA [7] dataset.

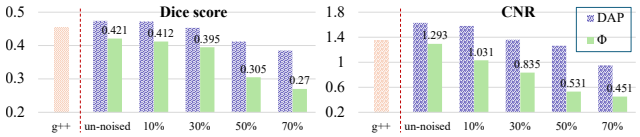

Figure 3. Dice score (*left*) and CNR (*right*) of DAP and Φ against the ground truth of RSNA dataset under different noise levels, compared to the strong baseline g++[44].

E. Robustness to flawed Φ

We inject noise into Φ and investigate *how DAP degrades as noise increases*. We first consider pixels of Φ with value $>$ 0.3 as important. Then, we flip top-$k \in \{10, 30, 50, 70\}\%$ of important pixels to 0, and retrain the model to find *how good should* Φ *be to benefit DAP*. We conducted experiments on RSNA and plot the result in Fig. 3. DAP's performance is indeed correlated with the quality of Φ, which is lower than g++ when $\text{dice}(\Phi, \text{GT}) \leq 0.3$ at $k = 50\%$. Yet, we note that this is artificial scenario. In practice, Φ exhibits a reasonable localization capability with 0.34/0.42/0.33 dice score on MS-CXR/RSNA/Covid.

F. Qualitative results

We present more qualitative results of our proposed DAP in Fig. 4 and Fig. 5.

References

[1] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *ECCV*, pages 1–21. Springer, 2022. 1, 4, 5

[2] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 2

[3] Pengyue Lin, Zhihan Yu, Mingcong Lu, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. Visual prompt tuning for weakly supervised phrase grounding. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7895–7899. IEEE, 2024. 3

[4] OpenAI. Gpt-4 technical report, 2023. 2

[5] Tal Shaharabany and Lior Wolf. Similarity maps for self-training weakly-supervised phrase grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6925–6934, 2023. 3

[6] Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. *Advances in Neural Information Processing Systems*, 35:28222–28237, 2022. 3

[7] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019. 3

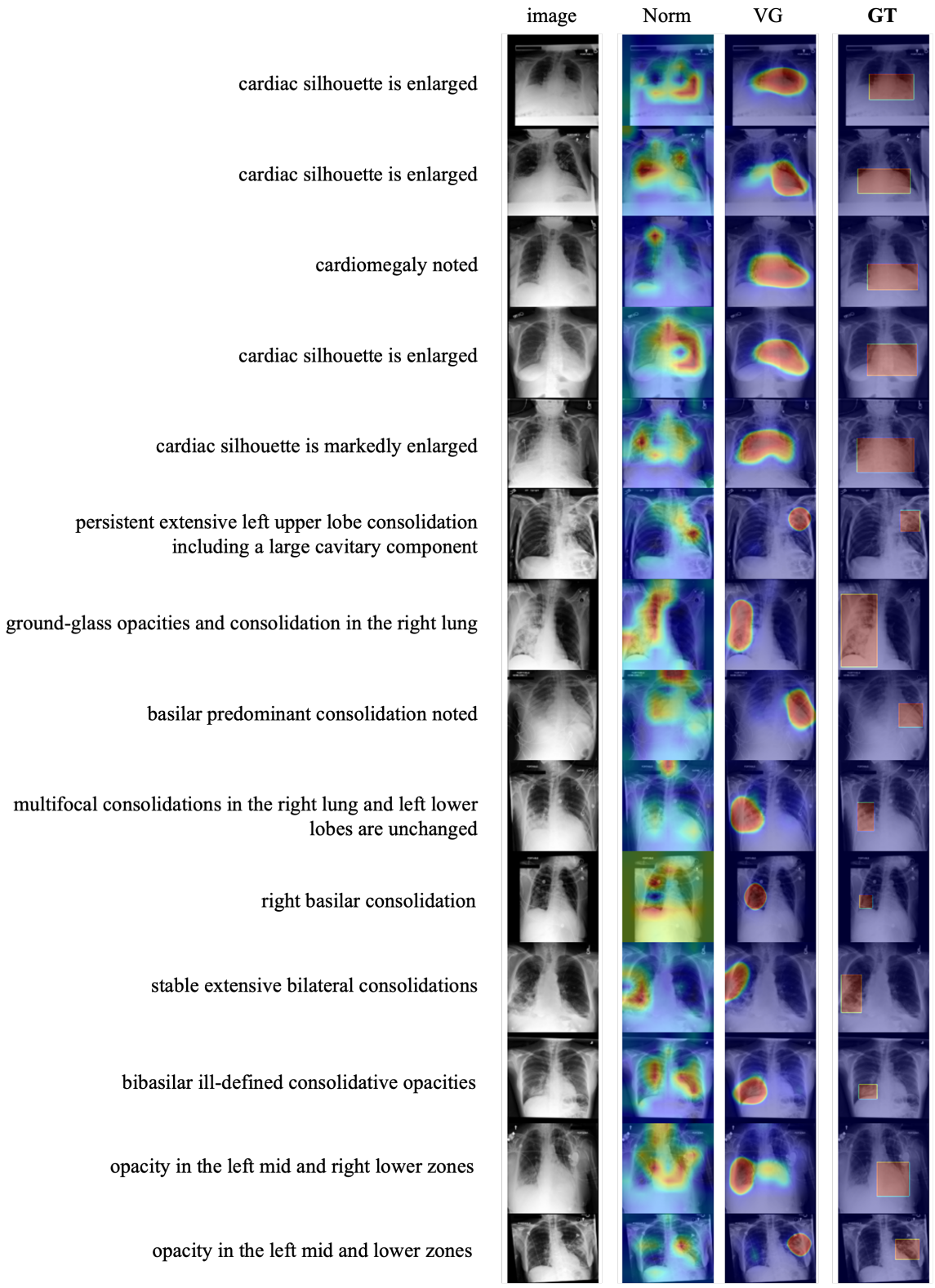

Figure 4. Qualitative results of DAP on MS-CXR [1] dataset.

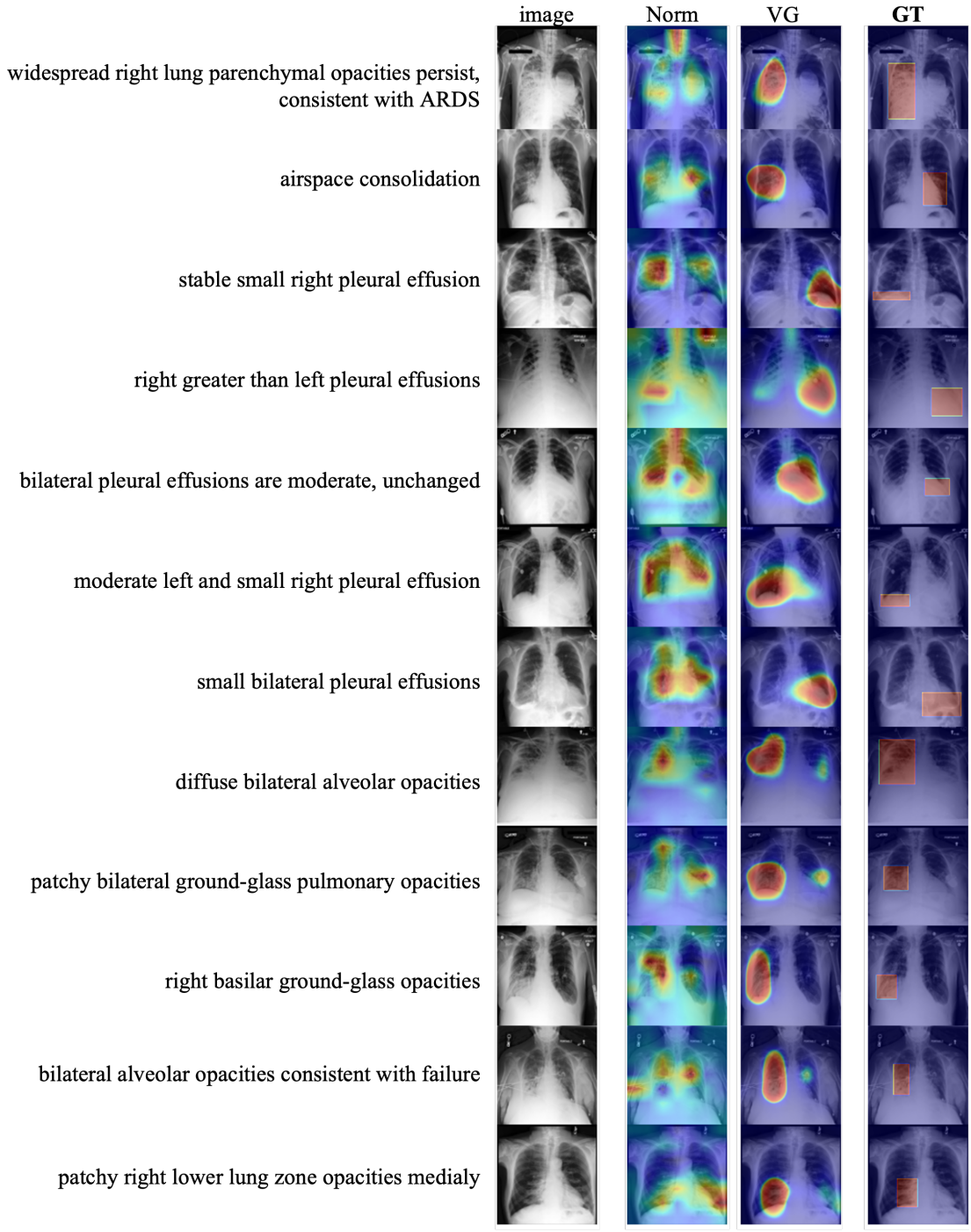

Figure 5. More qualitative results of DAP on MS-CXR [1] dataset.