CG-Reasoner: Centroid-Guided Positional Reasoning Segmentation for Medical Imaging with a Robust Visual-Text Consistency Metric

Supplementary Material

1. CG-Reasoner Training Algorithm

Alg. 1 summarizes the end-to-end training workflow of the proposed CG-Reasoner framework, while Alg. 2 outlines the computation of the PRScore metric used to evaluate multimodal spatial reasoning performance.

Algorithm 1 Training Procedure for the CG-Reasoner Framework

Input: Training set $\{(I_i, T_i, M_{\text{gt},i}, (x_{c,i}, y_{c,i}))\}_{i=1}^{N}$

Output: Predicted segmentation mask $\hat{M}$, generated reasoning text $\hat{T}$

for $i = 1$ to N **do**
  $\mathbf{V} \leftarrow f_{\text{Encoder}}(I_i)$
  $\mathbf{T}_{\text{emb}}, \hat{T} \leftarrow f_{\text{LLM}}(I_i, T_i)$
  $\mathbf{S} \leftarrow f_{\text{LLM}}(T_i \mid \text{LoRA})$
  $(\hat{x}_c, \hat{y}_c) \leftarrow f_{\text{T2C}}(\hat{T})$ // frozen Text2Centroid
  $\hat{M} \leftarrow f_{\text{Decoder}}(\mathbf{V}, \mathbf{T}_{\text{emb}})$
  $\mathcal{L}_{seg} \leftarrow \text{Dice}(M_{\text{gt},i}, \hat{M}) + \text{BCE}(M_{\text{gt},i}, \hat{M})$
  $\mathcal{L}_{txt} \leftarrow \text{CE}(\hat{T}, T_i)$
  $\mathbf{p}_{\text{gt}} \leftarrow (x_{c,i}, y_{c,i}), \quad \mathbf{p}_{\text{pred}} \leftarrow (\hat{x}_c, \hat{y}_c)$
  $S_{\text{spatial}} \leftarrow \frac{1}{2}\big[\,\text{ncos}(\mathbf{p}_{\text{gt}}, \mathbf{p}_{\text{pred}}) + (1 - d(\mathbf{p}_{\text{gt}}, \mathbf{p}_{\text{pred}}))\big]$
  $\mathcal{L}_{spatial} \leftarrow 1 - S_{\text{spatial}}$
  $\mathcal{L}_{total} \leftarrow \lambda_{seg}\mathcal{L}_{seg} + \lambda_{txt}\mathcal{L}_{txt} + \lambda_{spatial}\mathcal{L}_{spatial}$
  $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{total}$ // update Decoder + LoRA only
end for
return $\hat{M}$, $\hat{T}$

Algorithm 2 Computation of PRScore Metric

Input: Ground-truth text T_{gt}, predicted text T_{pred}, predicted mask M_{pred}

Output: PRScore $\in [0, 1]$

$e_{\text{gt}}^{\text{text}} \leftarrow f_{\text{SBERT}}(T_{\text{gt}})$
$e_{\text{pred}}^{\text{text}} \leftarrow f_{\text{SBERT}}(T_{\text{pred}})$
$S_{\text{sem}} \leftarrow \text{ncos}(e_{\text{gt}}^{\text{text}}, e_{\text{pred}}^{\text{text}})$
$\mathbf{p}_{\text{gt}} \leftarrow f_{\text{T2C}}(T_{\text{gt}})$
$\mathbf{p}_{\text{pred}} \leftarrow f_{\text{T2C}}(T_{\text{pred}})$
$S_{\text{spatial}} \leftarrow \frac{1}{2}\big[\,\text{ncos}(\mathbf{p}_{\text{gt}}, \mathbf{p}_{\text{pred}}) + (1 - d(\mathbf{p}_{\text{gt}}, \mathbf{p}_{\text{pred}}))\big]$
$\mathbf{m} \leftarrow \text{centroid}(M_{\text{pred}})$
$S_{\text{vis}} \leftarrow \frac{1}{2}\big[\,\text{ncos}(\mathbf{p}_{\text{pred}}, \mathbf{m}) + (1 - d(\mathbf{p}_{\text{pred}}, \mathbf{m}))\big]$
PRScore $\leftarrow \alpha\, S_{\text{sem}} + \beta\, S_{\text{spatial}} + \gamma\, S_{\text{vis}}, \quad$ where $\alpha + \beta + \gamma = 1$
return PRScore

2. Comparison of PRScore with other Reasoning Metrics

Table 1 highlights a fundamental limitation of conventional text-similarity metrics—BLEU, ROUGE-L, and METEOR. Although the predicted reasoning texts in our experiment describe *different* spatial positions (*upper right*, *bottom right*, *bottom*), these metrics yield nearly identical scores across all predictions. This behavior arises because such metrics primarily measure lexical overlap rather than geometric or spatial correctness, making them insensitive to positional reasoning errors. In contrast, PRScore exhibits a clear and significant performance drop when the predicted position deviates from the ground-truth location, correctly penalizing spatial inconsistencies. We observe this trend consistently across many additional examples, confirming that PRScore reliably detects spatial reasoning errors that traditional metrics fail to capture.

In addition, we examine the Yes/No evaluation protocol used by the PRS-Med method, where ChatGPT or LLaMA is prompted to answer whether the predicted spatial description matches the ground truth:

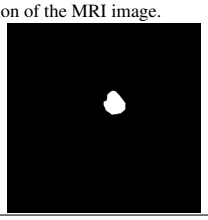
```
``You are a doctor and you want to
see the position of the tumor in the
medical image.  Given the following
question and groundtruth, check if
the prediction has the position word
similar to the groundtruth (ignore
the tissue).  Return only 'Yes' or
'No'.''
```

However, we find that this approach is not robust: ChatGPT and LLaMA almost always output *"Yes"* even when the predicted position is incorrect (e.g., *bottom right* vs. *top right*). This further demonstrates that binary LLM judgments cannot reliably capture fine-grained positional errors. PRScore addresses these shortcomings by jointly evaluating semantic similarity, spatial alignment, and text–mask consistency, offering a substantially more ROBUST and discriminative measure of multimodal reasoning quality.

3. Light-weight Encoder and Decoder design

The overall efficiency of CG-Reasoner depends heavily on the choice of the vision encoder, as its parameters directly determine the total size of the unified multimodal framework. As shown in Fig. 1, the reported parameter counts represent the full CG-Reasoner system when paired with different backbones: Tiny-SAM (31.5M), Med-SAM

Table 1. Comparison of reasoning metrics for spatial description evaluation. Traditional text-similarity measures (BLEU, ROUGE-L, METEOR) fail to distinguish correct and incorrect positional reasoning, producing nearly identical scores across different predictions. In contrast, PRScore captures meaningful spatial discrepancies, while the ChatGPT Yes/No protocol used in PRS-Med remains unreliable, often marking incorrect positions as correct.

Prompt	Ground Truth	Prediction	BLEU	ROUGE-L	METEOR	ChatGPT	PRScore
Describe the tumor location in the MRI scan. 	The tumour is found in the **top right** portion and near the **central** region of the MRI image. 	**Actual Prediction:** The tumor is located in the **upper right** region of the image.	0.076	0.552	0.421	Yes	**0.901**
		Tweaked Prediction-1: The tumor is located in the **bottom right** region of the image.	0.076	0.552	0.421	Yes	**0.652**
		Tweaked Prediction-2: The tumor is located in the **bottom** region of the image.	0.084	0.552	0.421	Yes	**0.54**6

(105M), Swin-UNet (49M), and ConvNeXt-Tiny (51M). Despite being the largest configuration, Med-SAM does not achieve competitive mean Dice and IoU scores and is consistently outperformed by both ConvNeXt-Tiny and Swin-UNet, indicating that increasing model size alone does not guarantee improved segmentation quality. Tiny-SAM, although lightweight, exhibits considerably lower accuracy across all datasets. In contrast, ConvNeXt-Tiny provides the best balance of efficiency and performance, achieving the highest mean Dice (0.84) and IoU (0.77) while maintaining a compact 51M parameter footprint. These results demonstrate that ConvNeXt-Tiny is the optimal backbone for CG-Reasoner, offering superior accuracy–efficiency trade-offs and strong generalization across diverse medical imaging modalities.

Alongside the choice of an efficient encoder, CG-Reasoner further benefits from a transformer-free lightweight decoder designed to minimize computational overhead while preserving strong spatial reasoning capability. Our decoder follows a SegFormer-inspired design built entirely from depthwise–separable convolutions, MixFFN blocks, prompt-conditioned FiLM modulation, and a simple two-stage upsampling head—completely avoiding multi-head self-attention or any transformer layers. This design enables effective fusion of visual features with LLM-derived prompt embeddings while keeping computation strictly localized and inexpensive. Despite its expressive ability, the decoder remains extremely compact, contributing fewer than 5M parameters to the overall model size, and thus significantly lighter than the commonly used transformer-based decoders. The proposed light-weight encoder–decoder framework enables the full 20-epoch training procedure to finish within 48 hours, even with evaluation performed after every epoch.

4. More Segmentation Results

Fig. 2 presents qualitative segmentation results across all six medical imaging datasets used in our study. For each modality, we show the input medical image, its corresponding ground-truth annotation, and segmentation outputs generated by several state-of-the-art models alongside our CG-Reasoner framework. The comparison reveals that our CG-Reasoner consistently produces sharper delineations and more anatomically faithful masks, demonstrating improved robustness across diverse modalities. These qualitative observations align with the quantitative gains reported earlier, reaffirming the effectiveness of our methodology.

5. More Reasoning Results

Fig. 3 presents additional qualitative reasoning outputs generated by CG-Reasoner across all six medical imaging modalities. For each dataset, we include two representative examples illustrating how the model interprets a clinical prompt and produces spatially grounded textual descriptions. These examples highlight CG-Reasoner's ability to correctly identify lesion position, describe anatomical context, and generate coherent reasoning that aligns with visual evidence in the image. These results further validate the effectiveness of our centroid-guided multimodal design, showing that CG-Reasoner provides interpretable visual reasoning alongside accurate segmentation.

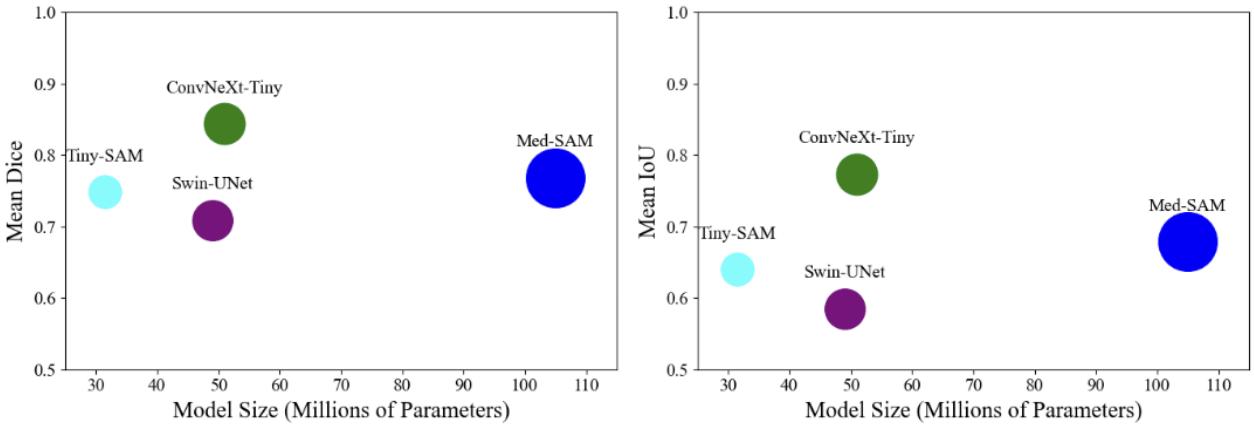


Figure 1. Quantitative comparison of mean Dice and IoU scores vs CG-Reasoner framework with different vision encoder backbones.

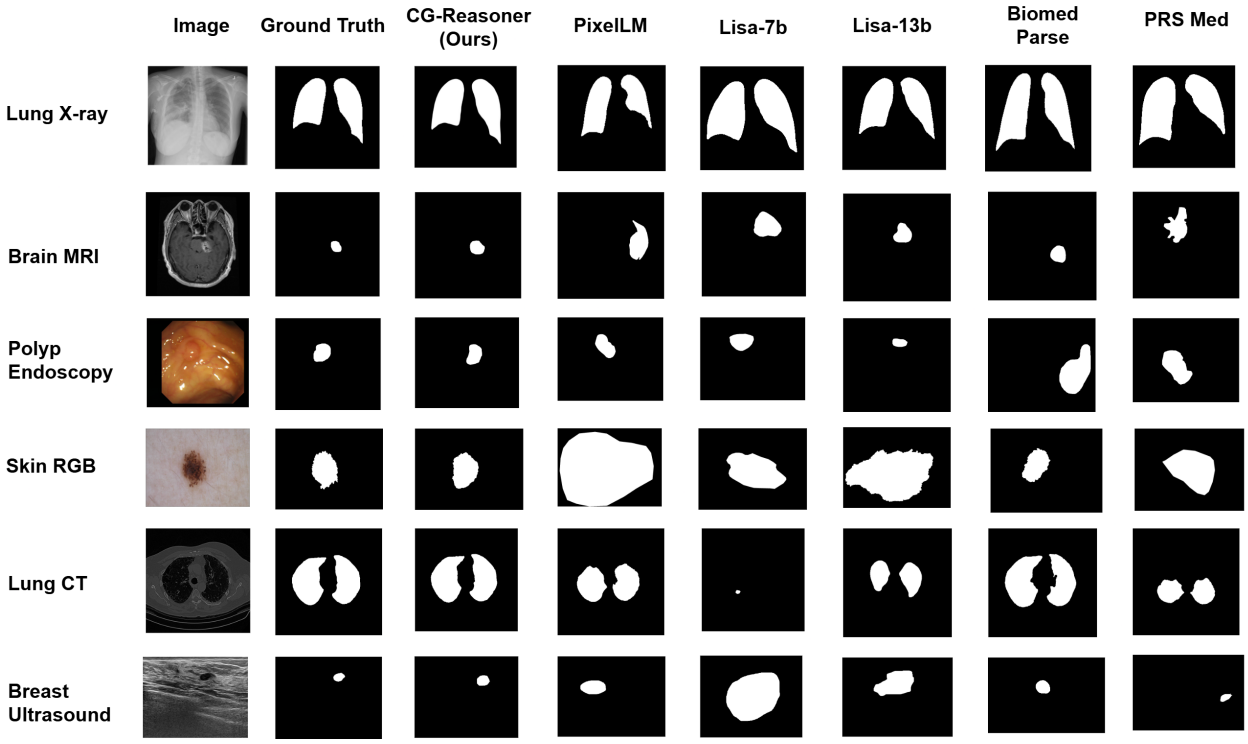
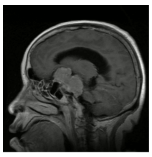
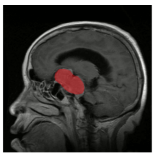


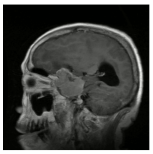
Figure 2. Qualitative comparison of CG-Reasoner segmentation results in comparison to the other state-of-the-art methods



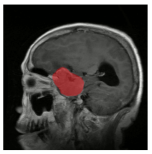
Prompt: Where is the tumour located in the MRI image



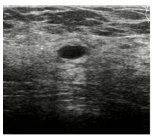
Response: The tumor is located in the left and near the center of the MRI image.



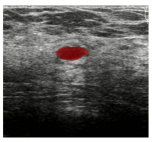
Prompt: Can you describe the position of the tumor in the MRI scan?



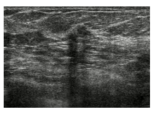
Response: The tumor is located at the center of the image



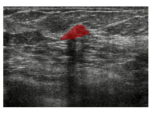
Prompt: In which part of the image is the tumor found?



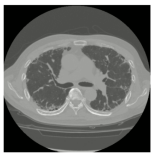
Response: The benign tumor is found near the center in the top region of the image



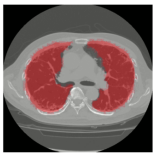
Prompt: Can you describe the location of the tumor in the image?



Response: The benign tumor is positioned in the top and central part of the image.



Prompt: Where are the two anatomies situated within the lung?



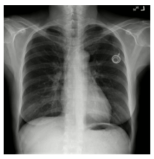
Response: The first anatomy is found in the left lobe of the lung. The second anatomy is observed at right lobe of the lung

Prompt: Where is the second anatomical abnormality situated in the CT scan?

Response: The second anatomical abnormality is located in the lower left portion of the lung.

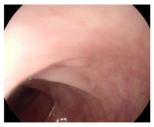
Prompt: In which areas of the image can the anatomies be found?

Response: The covid anatomies are situated around the center left of the image



Prompt: What is the general location of the anatomies in the image?,

Response: The covid anatomies are found in the central and near-central areas of the X-ray image

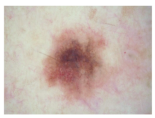


Prompt: Can you describe the location of the polyp in the image?

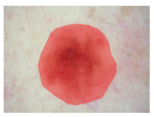
Response: The polyp is found near the central area of the image

Prompt: In which part of the image is the polyp found?

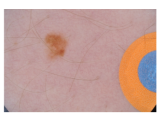
Response: The polyp is positioned towards the bottom left of the image.



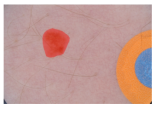
Prompt: Can you describe the location of the malignant skin lesion?



Response: The malignant is located in the central area and is of a benign type.



Prompt: Can you describe the location of the skin tumour in the image?



Response: The skin tumour is located in the top left part of the image.

Figure 3. More examples of CG-Reasoner reasoning results