Retrieval-Augmented Reasoning for Visual Localization

Pan Tao[1*], **Bosen Peng**[2*], **Yu Fang**[3, 4†]

[1]Department for West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu 610041, China
[2]Department of Ophthalmology, West China Hospital, Sichuan University, Chengdu 610041, China
[3]School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China
[4]School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu 610500, China
taopan_mail@stu.scu.edu.cn, bersonpeng@163.com, yufang@my.swjtu.edu.cn

Abstract

Open-vocabulary medical image localization holds significant potential for clinical applications. However, the practical reliability of current Vision-Language Models (VLMs) is constrained by critical limitations. They generate spatial prompts from statistical patterns rather than explicit medical evidence, resulting in unreliable localization. Furthermore, this implicit reasoning process is untraceable, failing to meet the clinical demand for evidence-based decision-making. To address these challenges, we propose RAR-VL (Retrieval-Augmented Reasoning for Visual Localization), a framework that transforms VLMs from implicit guessers into explicit reasoners. RAR-VL achieves this by integrating two key components: Retrieval-Augmented Generation (RAG) to source verifiable evidence from a medical knowledge base, and a Multimodal Chain-of-Thought (MCoT) to construct a structured, traceable reasoning path from evidence to localization. Experiments validate RAR-VL's state-of-the-art performance in zero-shot localization tasks, where it significantly outperforms existing open-vocabulary baselines. These results confirm that our retrieval-augmented reasoning framework effectively enhances both localization reliability and clinical trustworthiness.

Introduction

Prompt-based segmentation models, such as the Segment Anything Model (SAM) (Kirillov et al. 2023), have emerged as a transformative technology in image segmentation (Brasó, Ošep, and Leal-Taixé 2025; Tai et al. 2025; Liu et al. 2025). By accepting flexible prompts from users—in the form of points, bounding boxes, or text—these models can precisely delineate arbitrary objects within an image. However, when this powerful paradigm is applied to specialized domains like medicine, its reliance on manual interaction constitutes a significant application bottleneck. On one hand, effective prompting often demands deep medical expertise from the operator to ensure locational accuracy. On the other hand, faced with the escalating volume of clinical data and diagnostic demands (Ronneberger, Fischer, and Brox 2015; Alom et al. 2018; Isensee et al. 2019; Antonelli et al. 2022; Kim et al. 2019; Perslev et al. 2019; Yu et al. 2020; Xia et al. 2020; He et al. 2021), a model reliant on case-by-case manual interaction is inefficient and lacks the scalability to process data at scale. Therefore, developing automated spatial prompt generation techniques, namely visual localization, is a critical necessity for the large-scale deployment of large models in the medical field.

Two primary pathways exist to achieve such automation, each with distinct challenges. A straightforward approach is to train a front-end object detector for specific medical segmentation tasks to automatically generate bounding box prompts (Liu et al. 2023; Cheng et al. 2024). However, this path confronts the inherent challenges of medical data acquisition. Owing to strict patient privacy regulations and the necessity for annotation by clinical experts—a process that is both time-consuming and prohibitively expensive—large-scale, high-quality annotated datasets are exceptionally scarce. This renders the training of a robust, specialized detector capable of covering all open-vocabulary needs exceedingly difficult. Against this backdrop, leveraging the powerful image-text understanding capabilities of Visual Language Models (VLMs) emerges as a highly promising alternative (Feng et al. 2025; Zhang et al. 2025; Li et al. 2025b, 2023; Farina et al. 2025; Wang et al. 2025a). Pretrained on vast quantities of general-domain image-text data, VLMs have learned a rich set of general-purpose features, giving them the potential to understand open-vocabulary instructions and perform initial localization without extensive domain-specific annotated data.

Despite this potential, the reliability of VLMs in clinical practice is constrained by two critical limitations.

- **Reliance on Implicit Correlations:** When performing localization, VLMs tend to depend on implicit statistical correlations learned from general-domain data, rather than on the explicit medical evidence required for high-stakes, high-precision diagnostic tasks. This reliance on ambiguous statistical priors often leads to deviations in the generated prompts, making them insufficiently reliable.
- **Lack of Interpretability:** More critically, this reasoning process, which relies on implicit correlations, is entirely opaque and untraceable. This runs contrary to the clinical principle that every decision must be based on verifiable evidence, and it constitutes the fundamental reason they are difficult to trust in safety-critical scenarios.

[*]These authors contributed equally.
[†]Corresponding author.

These core issues of reliability and interpretability motivate the need for a new framework.

To address the foregoing challenges, we propose RAR-VL (Retrieval-Augmented Reasoning for Visual Localization), a novel framework designed to transform a VLM from an implicit guesser into an explicit reasoner. Our framework does not simply fine-tune the model; instead, it fundamentally reconstructs its localization decision-making paradigm. Specifically, the framework first leverages Retrieval-Augmented Generation (RAG) (Zhang et al. 2025) to retrieve verifiable evidence from an external, specialized medical knowledge base. Subsequently, it employs a Multimodal Chain of Thought (MCoT) (Wang et al. 2025b) to construct a structured, traceable reasoning path that guides the VLM to perform logical comparisons and validate the applicability of the evidence, ultimately outputting a high-confidence spatial prompt. This design systematically enhances localization reliability while imbuing the model's decisions with the trustworthiness essential for clinical applications.

The main contributions of this paper are summarized as follows:

- We propose RAR-VL, an innovative framework that fundamentally addresses the reliability problem of VLMs in medical localization tasks by replacing unreliable, implicit statistical guessing with traceable, explicit evidence-based reasoning.
- We design a novel reasoning architecture that integrates Retrieval-Augmented Generation (RAG) and a Multimodal Chain-of-Thought (MCoT) to ensure that localization is based on verifiable evidence.
- Our proposed framework transforms the black-box decision process of a VLM into a transparent and auditable verification flow by generating a clear reasoning path, thereby establishing a critical foundation of trust for AI in high-stakes clinical scenarios.
- Extensive experiments on multiple public medical datasets demonstrate the effectiveness of our method, which achieves state-of-the-art zero-shot localization performance.

Related Work

Segment Anything Model 2

Prompt-based segmentation has recently emerged as a flexible and powerful interaction paradigm, achieving significant breakthroughs in computer vision. The initial milestone in this area, the Segment Anything Model (SAM), demonstrated unprecedented zero-shot image segmentation capabilities after being trained on a massive dataset (Kirillov et al. 2023). Its successor, the Segment Anything Model 2 (SAM2), further extends this zero-shot capacity from static images to the video domain, proposing a unified, promptable foundational model for visual segmentation (Ravi et al. 2024). Notably, SAM2 not only introduces mechanisms such as streaming memory to process temporal information but, more importantly, it also surpasses the original SAM in image segmentation tasks, achieving higher precision and a several-fold increase in speed (Ravi et al. 2024; Xiong et al. 2024; Guo et al. 2025; Bai et al. 2025). However, the full potential of both SAM and the more powerful SAM2 is contingent upon the quality of the input prompts. Therefore, the reliable and automated generation of these precise prompts to overcome the bottleneck of manual interaction constitutes the core challenge that our research aims to address.

Visual Language Models

To address the bottleneck of automated prompt generation, Visual Language Models (VLMs) offer a highly promising technical pathway (Xie et al. 2025; Shen et al. 2025; Jang, Lee, and Sohn 2025; Yamaguchi et al. 2025). Surpassing earlier models like CLIP, which focused primarily on image-text alignment through contrastive learning, the new generation of Large Multimodal Models (LMMs) exhibit deeper levels of vision-language fusion and reasoning (Vaswani et al. 2017; Feng et al. 2025; Zhang et al. 2025; Li et al. 2025b). Among these advanced models, Qwen-2.5 (Team 2024) distinguishes itself with its exceptional performance. It employs an advanced large language model as its core, deeply integrated with a powerful visual encoder. This architecture enables it not only to perform basic image-text matching but also to execute complex tasks, including detailed image description generation, multi-turn visual dialogue, and precise referential comprehension (Li et al. 2025a). Consequently, its powerful general-purpose capabilities make it an ideal candidate for generating spatial prompts from natural language instructions (Feng et al. 2025; Zhang et al. 2025). However, despite the power of such LMMs, a fundamental challenge remains: their decision-making process inherently relies on implicit statistical correlations learned from vast general-domain data, rather than on the explicit evidence-based reasoning required for medical diagnostics (Zhang et al. 2025; Li et al. 2025b). This inherent limitation is precisely the target that our RAR-VL framework is designed to systemically address through the introduction of external evidence.

Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) (Fan et al. 2024; Du et al. 2024) is a pivotal paradigm originating from the field of Natural Language Processing (NLP) (Zhang et al. 2025). Its core idea is to retrieve relevant information from a large-scale, trusted external knowledge base to serve as context before the model proceeds with generation or reasoning. This mechanism has been proven to effectively reduce hallucinations in large models and significantly enhance the factual accuracy and consistency of the generated content by incorporating external, verifiable knowledge (Zhang et al. 2025; Qi et al. 2024). In recent years, this concept has also begun to permeate the multimodal domain. In this work, we adapt the RAG paradigm to the task of visual localization, aiming to provide a VLM with the explicit visual evidence it inherently lacks. This approach is analogous to providing the model with a reliable external reference, thereby systematically addressing its predicament of relying solely on vague internal knowledge for implicit guesswork.

Multimodal Chain of Thought

The Chain of Thought (CoT) and its extension into the multimodal domain, Multimodal Chain of Thought (MCoT), represent another key technology aimed at unlocking the complex reasoning capabilities of large models (Wang et al. 2025b; Lai and Nissim 2024). Its core mechanism guides the model to generate a series of intermediate, step-by-step logical reasoning processes before arriving at a final answer (Liang et al. 2025). Research has demonstrated that this structured approach not only boosts model performance on complex tasks but also that the generated reasoning chain greatly enhances model interpretability (Wang et al. 2025b). In our framework, MCoT plays a crucial role; it is not employed for general-purpose tasks but is specifically designed to construct an explicit and traceable validation path. Through this path, the VLM's process of adopting external evidence becomes rigorous and auditable. This not only provides a logical guarantee for outputting high-reliability prompts but also directly addresses the fundamental demand in clinical applications for trustworthy, evidence-based decision-making.

Method

The application of large-scale Vision-Language Models (VLMs) to high-stakes domains like medical imaging is fundamentally constrained by their reliance on opaque, implicit statistical priors. This chapter introduces RAR-VL (Retrieval-Augmented Reasoning for Visual Localization), a novel reasoning framework that directly confronts this challenge by proposing a new paradigm of case-based explicit cognitive reasoning. This framework transforms the model from an intuitive guesser into a deliberative, evidence-backed reasoner by dynamically coupling a general-purpose, parametric Large Multimodal Model (LMM) with an external, non-parametric medical knowledge base. Notably, RAR-VL is a fully zero-shot reasoning framework that requires no task-specific training or fine-tuning, instead unlocking and constraining the potential of pre-trained models by constructing a structured reasoning process.

Problem Formulation and Paradigm Reconstruction

In the task of open-vocabulary medical image localization, given a medical image I and a natural language instruction C, the objective is to generate a precise spatial prompt B. Traditional end-to-end methods implicitly model this task as a direct conditional probability problem. This can be characterized as an unconstrained, end-to-end parametric approach, where the decision-making process depends entirely on the model's internal, parameterized knowledge θ, formalized as:

$$B_{\text{implicit}} = \text{VLM}(I, C; \theta) \tag{1}$$

The inherent opacity of this model poses significant credibility challenges in high-risk applications.

Our work challenges this paradigm by introducing an explicit, non-parametric knowledge constraint, thereby reformulating the problem as a constrained, evidence-guided conditional inference task:

$$P(B \mid I, C, E; \theta) \tag{2}$$

Our RAR-VL framework is designed to realize this paradigm. It is composed of two core modules: a Non-parametric Knowledge Integration module to ground the reasoning in real-world evidence, and a Deliberative Reasoning module to ensure that evidence is used in a logical and verifiable manner. The complete end-to-end inference pipeline of the RAR-VL framework is formally detailed in Algorithm 1.

Non-parametric Knowledge Integration (RAG)

The cornerstone of explicit reasoning is to provide the LMM with specific, verifiable real-world knowledge that is absent from its own parameterized knowledge. This module aims to dynamically integrate an external, non-parametric knowledge base with the reasoning process.

Construction of the Visual Knowledge Base At the core of our retrieval mechanism is a comprehensive, structured medical knowledge base $\mathcal{K}$. This knowledge base is meticulously curated by aggregating a large-scale collection of public medical imaging datasets, creating a repository of visual exemplars. Each entry in $\mathcal{K}$ links a visual pattern to a consistent semantic anchor (text) and a precise geometric location (box) derived from its ground-truth segmentation mask, structured as $\{$*image_path, text, box*$\}$.

Feature Space Mapping for Evidence Retrieval To enable efficient retrieval, the knowledge base is indexed offline. A pre-trained vision-language model, BLIP2 (Li et al. 2023) is employed as a visual encoder to transform each image in $\mathcal{K}$ into a high-dimensional feature vector. This normalization projects all visual concepts onto a unified hypersphere, allowing for efficient and meaningful similarity computation via the inner product, which becomes equivalent to cosine similarity.

During inference, a given query (I, C) is encoded into a normalized query vector using the same frozen visual encoder. The similarity to all entries in the knowledge base is then efficiently calculated via inner product with the precomputed feature matrix. The top-k most similar entries are returned as the candidate evidence set E_{cand}. This process can be summarized as:

$$E_{\text{cand}} = \text{Retrieve}(I, C; \mathcal{K}) \tag{3}$$

Deliberative Reasoning Module (MCoT)

While sourcing relevant evidence is a critical first step, the core novelty of RAR-VL lies in its deliberative process for utilizing that evidence. Upon obtaining the candidate evidence E_{cand}, the framework enters the core reasoning stage, which is designed not to blindly adopt the evidence but to validate its effectiveness through a structured logical process.

Parametric Reasoning Core The reasoning core of our framework is a state-of-the-art, unfine-tuned Large Multimodal Model (LMM), such as Qwen2.5-VL. We leverage

its capabilities in a zero-shot inference setting as a general-purpose parametric reasoning engine. For efficiency, inference is optimized using techniques like 8-bit quantization and Flash Attention 2.

Multimodal Chain-of-Thought To constrain the LMM's reasoning process and prevent hallucinations or logical shortcuts, we designed a Multimodal Chain-of-Thought (MCoT) prompting strategy. This strategy is essentially a Cognitive-Computational Scaffold that compels the LMM to follow a hierarchical, hypothesis-testing reasoning path, mimicking the deliberative thought process of humans when faced with evidence. This path includes the following three logical steps:

- **Step 1: Conceptual Congruence Assessment.** The LMM first assesses the conceptual congruence between the retrieved exemplar E^* and the current task. It judges whether the semantic information of E^* is conducive to understanding the target in instruction C and image I, thereby outputting a boolean judgment v_c.
- **Step 2: Geometric Hypothesis Testing.** Only if semantic relevance is confirmed (v_c = True) does the LMM proceed to this step. It treats the exemplar's localization data B_e as a geometric hypothesis and tests its validity in the context of the current image I, yielding a boolean judgment v_p.
- **Step 3: Meta-cognitive Policy Synthesis.** Finally, the framework executes a meta-level reasoning policy $A = f_{\text{policy}}(v_c, v_p)$ based on the outcomes of the validation stages: if both checks pass, it executes Prior Adoption; if only the semantic check passes, it performs Concept-guided Global Search; if the semantic check fails, it falls back to the Zero-shot Independent Reasoning mode.

Ultimately, the structured text output by the LMM contains the complete reasoning chain and the decision result. By parsing this text, we can obtain a high-confidence spatial prompt B^*. The entire deliberative reasoning process is formally detailed in Algorithm 1.

Algorithm 1: The **RAR-VL** Framework Inference Pipeline

Input: Query Image I, Natural Language Instruction C
Parameters: Knowledge Base $\mathcal{K}$, LMM Reasoner Φ
Output: High-Confidence Spatial Prompt B^*

function RAR-VL_INFERENCE(I, C)
  $E_{\text{cand}} \leftarrow \text{Retrieve}(I, C; \mathcal{K})$
  if E_{cand} is empty **then**
    $B^* \leftarrow \Phi_{\text{ZeroShot}}(I, C)$
    return B^*
  end if
  $E^* \leftarrow \text{get\_best\_candidate}(E_{\text{cand}})$
  $v_c, v_p, B_{\text{decision}} \leftarrow \Phi_{\text{MCoT}}(I, C, E^*)$
  if v_c = True and v_p = True **then**
    $B^* \leftarrow E^*.box$
  else if v_c = True and v_p = False **then**
    $B^* \leftarrow B_{\text{decision}}$
  else
    $B^* \leftarrow \Phi_{\text{ZeroShot}}(I, C)$
  end if
  return B^*
end function

Experiments

To comprehensively evaluate the efficacy, robustness, and internal mechanisms of our proposed RAR-VL framework, we designed a series of rigorous experiments. This chapter first introduces the datasets, evaluation metrics, and baselines used for our experiments. We then present the quantitative comparison of RAR-VL against state-of-the-art methods on several benchmark tasks. Finally, through a detailed set of ablation studies, we provide an in-depth analysis of the contributions and synergistic effects of the framework's core components.

Experimental Setup

Datasets Our experiments are conducted on a diverse set of medical imaging datasets, with each chosen to evaluate different aspects of our framework's capabilities:

- **ISIC 2018:** As a widely recognized benchmark for skin lesion segmentation, this dataset serves to evaluate the model's baseline performance on tasks involving quasi-natural images.
- **Medical Segmentation Decathlon (MSD) & BraTS 2021:** We selected four representative tasks from the MSD challenge—Heart, Hippocampus, Prostate, and Spleen—and the BraTS 2021 brain tumor dataset. These datasets include various imaging modalities such as MRI and CT and complex anatomical structures, used to rigorously test the model's generalization and reliability in highly-specialized domains.

We implemented rigorous measures to ensure a fair evaluation and prevent data leakage across all benchmarks.

Baselines and Metrics We compare the RAR-VL framework against two main categories of methods: (1) Supervised Specialist Models, such as the U-Net series (Ronneberger, Fischer, and Brox 2015; Alom et al. 2018) and nnU-Net (Isensee et al. 2019), which represent the performance upper-bound, and (2) Zero-shot Generalist Models, including YOLO-World (Cheng et al. 2024), Grounding DINO (Liu et al. 2023), and FG-CLIP (Xie et al. 2025), which are the most direct competitors to our paradigm.

Performance is evaluated using metrics appropriate for our new focus on localization and its downstream effects. For localization, we primarily report Intersection over Union (IoU) and Precision (Prec.). For the downstream 2D segmentation task (ISIC 2018), we report Sensitivity (SE), Specificity (SP), F1-Score, Accuracy (AC), and Dice Coefficient (DC). For 3D volumetric segmentation (MSD & BraTS), we report the Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD). All inference times are benchmarked on the same hardware platform for fair comparison.

Implementation Details Our RAR-VL framework is an inference-only pipeline that requires no training. The knowledge base is constructed from a variety of publicly available

medical datasets. During inference, the final spatial prompt B^* generated by RAR-VL is evaluated in two ways: (1) Directly for its localization accuracy using IoU, and (2) Indirectly by using it as input to downstream segmentation predictors, including both SAM2 (Ravi et al. 2024) and the domain-specific MedSAM, to produce the final pixel-level mask. This allows us to evaluate both the prompt's intrinsic quality and its utility for segmentation pipelines.

Quantitative Results

Localization Performance We first evaluate RAR-VL on the primary task of visual localization, with results shown in the consolidated Table 1. The data clearly shows that generalist models fail to reliably locate targets in specialized domains. On the BraTS dataset, all baselines (FGCLIP, Grounding DINO, YOLO-World) achieve an IoU below 0.1, indicating a complete failure. In contrast, RAR-VL achieves a robust IoU of 0.4820. This pattern is consistent across the MSD tasks; for example, on Prostate, RAR-VL achieves an IoU of 0.8347, whereas YOLO-World scores only 0.0150. On the ISIC 2018 task, RAR-VL also demonstrates the best balance of IoU (0.8424) and precision, surpassing all other zero-shot methods. These results directly validate RAR-VL as a superior zero-shot localization framework.

Downstream Segmentation Performance The high quality of our localization prompts translates directly into exceptional downstream segmentation performance, as detailed in Table 2 and Table 3. On the ISIC 2018 task (Table 2), while YOLO-World achieves a high DC (0.9021), its clinically unusable Specificity (SP) of 0.0817 reveals its poor localization. RAR-VL paired with SAM2 achieves a balanced 0.8701 DC with a 0.9851 SP. Furthermore, when paired with a domain-specific segmenter (MedSAM), our framework's performance leaps to 0.9460 DC, rivaling fully-supervised models. The true power of this paradigm is shown in the MSD results (Table 3), where baselines suffer a catastrophic performance collapse (near-zero DSC scores). RAR-VL is the only zero-shot framework to maintain robust performance. Most impressively, on the Prostate dataset, RAR-VL + MedSAM achieves a DSC of 0.9568, dramatically outperforming even the fully-supervised SOTA nnUNet (0.8311). This demonstrates that our evidence-based localization can drive downstream models to surpass even specialized, fully-supervised methods.

Ablation Studies

Our ablation studies, summarized in Table 4, confirm that this performance is not from a single component but from the indispensable synergy between retrieval and reasoning. The effectiveness of our hybrid paradigm is confirmed by two critical ablations. First, when the retrieval module is removed (w/o Retrieval configuration), the model's performance collapses on all specialized tasks; for instance, Heart DSC plummets from 0.68 to 0.06. This precisely demonstrates that the VLM's implicit knowledge is insufficient for specialized domains. Conversely, when the deliberative reasoning module is ablated (w/o Reasoning configuration), performance is significantly degraded, with Spleen DSC dropping from 0.89 to 0.76. This proves that simply retrieving relevant evidence is not enough; the structured reasoning provided by the LMM is crucial for correctly validating and applying the evidence. This synergy is the foundational mechanism of RAR-VL.

Beyond performance, efficiency is a key measure of practical utility, with detailed time analysis deferred to the appendix. The data reveals that the deliberative reasoning process introduces a necessary computational cost. However, this cost is justified by the massive leap in performance from complete failure to robust reliability. Counter-intuitively, the retrieval module also acts as an efficiency booster. The w/o Retrieval configuration is by far the most computationally expensive, demonstrating that providing focused evidence critically prunes the search space, making subsequent deliberation far more efficient than an unguided, brute-force VLM reasoning approach.

Conclusion

The application of large vision-language models to medical imaging is hampered by their reliance on opaque, implicit knowledge, leading to unpredictable failures in specialized domains. This paper introduced RAR-VL, a zero-shot reasoning framework that addresses this limitation by grounding a pre-trained LMM in an external, non-parametric knowledge base. This approach recasts the localization task from implicit pattern matching to an explicit, evidence-based reasoning process. Experiments demonstrate that RAR-VL maintains high reliability and robustness on specialized medical datasets where generalist, zero-shot models catastrophically fail, validating the effectiveness of the evidence-based paradigm. By shifting the focus from end-to-end training to structured, evidence-guided reasoning, this work offers a promising direction for developing more trustworthy, data-efficient, and generalizable AI systems for medicine and other safety-critical fields.

References

Alom, M. Z.; Hasan, M.; Yakopcic, C.; Taha, T. M.; and Asari, V. K. 2018. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*.

Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R. M.; et al. 2022. The medical segmentation decathlon. *Nature Communications*, 13(1): 4128.

Bai, Y.; Xu, Y.; Chen, S.; Zhu, X.; Wang, S.; Huang, S.; Song, Y.; Zheng, Y.; Liu, Z.; Tan, S.; et al. 2025. TOPS-speed complex-valued convolutional accelerator for feature extraction and inference. *Nature Communications*, 16(1): 292.

Brasó, G.; Ošep, A.; and Leal-Taixé, L. 2025. Native Segmentation Vision Transformers. *arXiv preprint arXiv:2505.16993*.

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Model	ISIC 2018				BraTS			
	IoU↑	Prec.↑	Sens.↑	Area Ratio	IoU↑	Prec.↑	Sens.↑	Area Ratio
FGCLIP	0.3633	0.3682	0.9665	4.8251	0.0561	0.0598	0.6962	87.0999
Grounding DINO	0.6982	**0.8429**	0.8549	1.9155	0.0985	0.0990	0.9824	74.3120
YOLO-World	0.8217	0.7286	0.9418	5.1335	0.0968	0.0968	**1.0000**	78.9254
RAR-VL (Ours)	**0.8424**	0.8424	**1.0000**	**1.1886**	**0.4820**	**0.4866**	0.9866	**11.4947**

Model	Heart				Hippocampus			
	IoU↑	Prec.↑	Sens.↑	Area Ratio	IoU↑	Prec.↑	Sens.↑	Area Ratio
FGCLIP	0.0709	0.0716	**0.9309**	30.0336	0.4258	0.6312	**0.7036**	2.4984
Grounding DINO	**0.6605**	**0.8558**	0.7050	**1.1317**	**0.5079**	**0.9648**	0.5184	0.5953
YOLO-World	0.0186	0.0191	0.0355	35.1234	0.0041	0.0041	0.0081	0.8991
RAR-VL (Ours)	0.6186	0.8434	0.7191	1.5739	0.3901	0.9429	0.4005	**0.4846**

Model	Prostate				Spleen			
	IoU↑	Prec.↑	Sens.↑	Area Ratio	IoU↑	Prec.↑	Sens.↑	Area Ratio
FGCLIP	0.1626	0.1684	**0.9711**	14.0018	0.1385	0.1393	0.9436	52.4592
Grounding DINO	**0.8868**	0.9484	0.9092	1.3190	**0.8894**	0.9402	**0.9423**	1.6592
YOLO-World	0.0150	0.0153	0.0291	16.5432	0.0060	0.0062	0.0118	60.1121
RAR-VL (Ours)	0.8347	**0.9584**	0.8697	**1.2676**	0.8627	**0.9451**	0.9110	**1.6347**

Table 1: Consolidated core localization performance on ISIC 2018, BraTS, and all four MSD datasets. This compact table shows RAR-VL's robust IoU on specialized tasks where baselines fail. (Note: Grounding DINO IoU for Prostate corrected from 8.8668 to 0.8868).

Du, X.; Zheng, G.; Wang, K.; Zou, Y.; Wang, Y.; Deng, W.; Feng, J.; Liu, M.; Chen, B.; Peng, X.; et al. 2024. Vul-rag: Enhancing llm-based vulnerability detection via knowledge-level rag. *arXiv preprint arXiv:2406.11147*.

Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 6491–6501.

Farina, M.; Mancini, M.; Iacca, G.; and Ricci, E. 2025. Rethinking Few-Shot Adaptation of Vision-Language Models in Two Stages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 29989–29998.

Feng, Y.; Liu, Y.; Yang, S.; Cai, W.; Zhang, J.; Zhan, Q.; Huang, Z.; Yan, H.; Wan, Q.; Liu, C.; et al. 2025. Vision-language model for object detection and segmentation: A review and evaluation. *arXiv preprint arXiv:2504.09480*.

Guo, G.; Guo, Y.; Yu, X.; Li, W.; Wang, Y.; and Gao, S. 2025. Segment Any-Quality Images with Generative Latent Space Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2366–2376.

He, Y.; Yang, D.; Roth, H.; Zhao, C.; and Xu, D. 2021. DINTS: Differentiable neural network topology search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5841–5850.

Isensee, F.; Jäger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2019. Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*.

Jang, J.; Lee, J.; and Sohn, K. 2025. Descriptive Image-Text Matching with Graded Contextual Similarity. *arXiv preprint arXiv:2505.09997*.

Kim, S.; Kim, I.; Lim, S.; Baek, W.; Kim, C.; Cho, H.; Yoon, B.; and Kim, T. 2019. Scalable neural architecture search for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 220–228. Springer.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. *arXiv:2304.02643*.

Lai, H.; and Nissim, M. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models. *arXiv preprint arXiv:2406.02301*.

Li, G.; Xu, J.; Zhao, Y.; and Peng, Y. 2025a. Dyfo: A training-free dynamic focus visual search for enhancing LMMS in fine-grained visual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9098–9108.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Y.; Liu, Z.; Li, Z.; Zhang, X.; Xu, Z.; Chen, X.; Shi, H.; Jiang, S.; Wang, X.; Wang, J.; et al. 2025b. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.

Liang, X.; Lin, M.; Ruan, W.; Liu, Y.; Zhuang, Y.; and Liang, X. 2025. Memory-driven multimodal chain of thought for embodied long-horizon task planning.

Method	Segmenter	SE↑	SP↑	F1↑	AC↑	DC↑
Supervised Specialist Models						
U-Net (t=2)	(Supervised)	0.9479	0.9263	0.8682	0.9314	0.8476
R2U-Net (t=3)	(Supervised)	0.9414	0.9425	0.8920	0.9424	0.8616
Zero-shot Generalist Models						
YOLO-World	SAM2	0.9418	0.0817	0.8216	0.8236	0.9021
Grounding DINO	SAM2	0.7825	0.2595	0.1385	0.3313	0.2433
FG-CLIP	SAM2	0.3523	0.6621	0.3343	0.3948	0.5011
SAM2	(text-prompt only)	0.0258	**0.9968**	0.0493	0.8634	0.0493
MedSAM	(text-prompt only)	0.8679	0.1472	0.2347	0.2436	0.2347
RAR-VL (Ours)	**SAM2**	0.8306	0.9851	0.8701	0.9639	0.8701
RAR-VL (Ours)	**MedSAM**	**0.9657**	0.9883	**0.9460**	**0.9852**	**0.9460**

Table 2: Downstream segmentation performance on the ISIC 2018 task. This unified table shows RAR-VL provides balanced prompts (high SP) and achieves SOTA performance when paired with MedSAM.

Method	Heart		Hippocampus		Prostate		Spleen	
	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑
Supervised Specialist Models								
nnUNet	**93.30**	**96.74**	**89.46**	**97.66**	83.11	97.56	**97.43**	**99.89**
Zero-shot Generalist Models (with SAM2)								
YOLO-World + SAM2	3.66	13.97	0.81	7.76	2.96	9.56	1.19	4.07
Grounding DINO + SAM2	2.62	50.02	17.71	51.60	8.51	49.15	5.85	41.71
FG-CLIP + SAM2	3.33	47.99	18.21	49.23	9.13	48.20	1.50	14.28
SAM2 (text-prompt only)	0.0031	0.0772	0.0000	0.0051	0.0128	0.0654	0.0010	0.0066
RAR-VL + SAM2	67.87	15.08	56.94	43.21	84.62	62.42	88.64	71.03
Zero-shot Generalist Models (with MedSAM)								
MedSAM (text-prompt only)	0.0137	0.0012	0.1535	0.1212	0.0704	0.0474	0.0254	0.0527
RAR-VL + MedSAM	88.73	86.56	79.48	94.70	**95.68**	**99.76**	96.04	97.68

Table 3: Downstream segmentation performance on MSD. Baselines catastrophically fail (DSC ≈ 0). **RAR-VL + MedSAM** achieves SOTA performance, even **beating the fully-supervised nnUNet** on the Prostate task. (Note: Data for text-prompt only rows corrected).

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, Z.; Saseendran, A.; Tong, L.; He, X.; Yousefi, F.; Burlutskiy, N.; Oglic, D.; Diethe, T.; Teare, P. A.; Zhou, H.; et al. 2025. Segment Anyword: Mask Prompt Inversion for Open-Set Grounded Segmentation. In *Forty-second International Conference on Machine Learning*.

Perslev, M.; Dam, E. B.; Pai, A.; and Igel, C. 2019. One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 30–38. Springer.

Qi, J.; Xu, Z.; Shao, R.; Chen, Y.; Di, J.; Cheng, Y.; Wang, Q.; and Huang, L. 2024. Rora-vlm: Robust retrieval-augmented vision language models. *arXiv preprint arXiv:2410.08876*.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241. Springer.

Shen, L.; Gong, G.; Hao, T.; He, T.; Zhang, Y.; Liu, P.; Zhao, S.; Han, J.; and Ding, G. 2025. DiscoVLA: Discrepancy Reduction in Vision, Language, and Alignment for Parameter-Efficient Video-Text Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19702–19712.

Tai, W.-E.; Shih, Y.-L.; Sun, C.; Wang, Y.-C. F.; and Chen, H.-T. 2025. Segment Anything, Even Occluded. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 29385–29394.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Configuration	ISIC 2018			Heart		Hippo.		Prostate		Spleen		BraTS 2021	
	SE↑	SP↑	DC↑	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑	Dice↑	mIoU↑
w/o Reasoning	0.60	0.91	0.55	0.67	0.14	0.49	0.40	0.79	0.52	0.76	0.64	0.76	0.65
w/o Retrieval	**0.83**	0.87	0.84	0.06	0.00	0.14	0.25	0.07	0.03	0.04	0.05	0.27	0.17
w/o Tier-2	0.57	0.89	0.51	0.57	0.13	0.50	0.41	0.80	0.56	0.76	0.68	**0.78**	**0.66**
Unguided SAM2	0.03	**1.00**	0.05	0.00	0.08	0.00	0.01	0.01	0.07	0.00	0.01	0.02	0.01
RAR-VL (Full)	**0.83**	0.99	**0.87**	**0.68**	**0.15**	**0.57**	**0.43**	**0.85**	**0.62**	**0.89**	**0.71**	**0.78**	**0.66**

Table 4: Ablation study on downstream segmentation performance. Removing either reasoning ('w/o Reasoning') or retrieval ('w/o Retrieval') causes a significant or catastrophic performance collapse, proving their synergy.

Wang, L.; Wang, M.; Fu, H.; and Zhang, D. 2025a. Vision-Language Model IP Protection via Prompt-based Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9497–9506.

Wang, Y.; Wu, S.; Zhang, Y.; Yan, S.; Liu, Z.; Luo, J.; and Fei, H. 2025b. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Xia, Y.; Liu, F.; Yang, D.; Cai, J.; Yu, L.; Zhu, Z.; Xu, D.; Yuille, A.; and Roth, H. 2020. 3d semi-supervised learning with uncertainty-aware multi-view co-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3646–3655.

Xie, C.; Wang, B.; Kong, F.; Li, J.; Liang, D.; Zhang, G.; Leng, D.; and Yin, Y. 2025. FG-CLIP: Fine-Grained Visual and Textual Alignment. *arXiv preprint arXiv:2505.05071*.

Xiong, Y.; Varadarajan, B.; Wu, L.; Xiang, X.; Xiao, F.; Zhu, C.; Dai, X.; Wang, D.; Sun, F.; Iandola, F.; et al. 2024. EfficientSAM: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16111–16121.

Yamaguchi, S.; Feng, D.; Kanai, S.; Adachi, K.; and Chijiwa, D. 2025. Post-pre-training for modality alignment in vision-language foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4256–4266.

Yu, Q.; Yang, D.; Roth, H.; Bai, Y.; Zhang, Y.; Yuille, A. L.; and Xu, D. 2020. C2FNAS: Coarse-to-fine neural architecture search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4126–4135.

Zhang, X.; Guo, J.; Zhao, S.; Fu, M.; Duan, L.; Wang, G.-H.; Chen, Q.-G.; Xu, Z.; Luo, W.; and Zhang, K. 2025. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*.