RAU: Reference-based Anatomical Understanding with Vision Language Models

Yiwei Li[1,2]* **, Yikang Liu[1]*, Jiaqi Guo[1,3]**, Lin Zhao[1], Zheyuan Zhang[1], Xiao Chen[1], Boris Mailhe[1], Ankush Mukherjee[1], Terrence Chen[1], and Shanhui Sun[1]***

[1] United Imaging Intelligence, Boston, MA, USA
[2] School of Computing, University of Georgia, Athens, GA, USA
[3] Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA

Abstract. Anatomical understanding, which is the ability to identify, localize, or segment anatomical structures, is critical in medical image analysis; however, its progress is constrained by the scarcity of expert-labeled data. A promising remedy is to leverage an annotated reference image to guide the interpretation of an unlabeled target. Although recent vision–language models (VLMs) exhibit non-trivial visual reasoning, their reference-based understanding and fine-grained localization remain limited. We introduce RAU, a framework for reference-based anatomical understanding with VLMs. We first show that a VLM learns to identify anatomical regions through relative spatial reasoning between reference and target images, trained on a moderately sized dataset. We validate this capability through visual question answering (VQA) and bounding box prediction. Next, we demonstrate that the VLM-derived spatial cues can be seamlessly integrated with the fine-grained segmentation capability of SAM2, enabling localization and pixel-level segmentation of small anatomical regions, such as vessel segments. Across two in-distribution and two out-of-distribution datasets, RAU consistently outperforms a SAM2 fine-tuning baseline using the same memory setup, yielding more accurate segmentations and more reliable localization. More importantly, its generalization ability to unseen modalities makes it scalable to unseen datasets, a property crucial for medical image applications. To the best of our knowledge, RAU is the first to explore the capability of VLMs for reference-based identification, localization, and segmentation of anatomical structures in medical images. Its promising performance highlights the potential of VLM-driven approaches for anatomical understanding in automated clinical workflows.

Keywords: Vision Language Model · Anatomical Understanding · Reinforcement Learning

* Equal contribution.
** This work was carried out during an internship at United Imaging Intelligence, Boston, MA.
*** Corresponding author. Email: shanhui.sun@uii-ai.com

1 Introduction

Anatomical understanding, which is the ability to identify, localize, or segment anatomical structures, is critical in medical image analysis [7]. It underpins a wide range of critical applications, such as automatic quantitative analysis of coronary angiography [37, 51], intraoperative navigation [25] and automated report generation [64], and supports accurate diagnostics [18], effective treatment planning [15], and precise surgical execution [16]. Traditionally, each of these tasks often requires the design and training of dedicated models tailored to their specific objectives [2, 60]. However, the scarcity of high-quality, annotated medical imaging datasets poses a significant challenge [61], since the acquisition of such annotations is resource-intensive and requires domain expertise [24]. This shortage of labeled data substantially hinders the ability to train robust and generalizable models for anatomical understanding [7], underscoring the need for approaches that can mitigate dependence on large-scale manual annotation [14].

Multimodal large language models (MLLMs) offer a promising solution to this challenge due to their strong learning and generalization abilities [19, 45]. Numerous studies have demonstrated that vision–language models (VLMs) can enhance their understanding of visual content through targeted training strategies [28, 56]. For instance, supervised fine-tuning (SFT) can be employed to align the model's outputs with high-quality, task-specific annotations, thereby improving accuracy in domain-relevant recognition tasks [**?**, 28, 56]. Reinforcement learning (RL), particularly methods such as Group Relative Policy Optimization (GRPO) [52], empowers VLMs with stronger reasoning capabilities by encouraging explicit chain-of-thought (CoT) generation, which not only enables complex multimodal inference but also drives superior generalization across both in-domain and out-of-domain data [17, 52, 72]. Furthermore, few-shot learning paradigms exploit a small number of annotated exemplars, in-context learning (ICL) enables VLMs to adapt to new anatomical queries directly from reference demonstrations at inference time, and retrieval-augmented generation (RAG) enriches the model's contextual grounding by incorporating relevant external knowledge [13, 31, 36, 67, 70]. These techniques are particularly beneficial for medical image understanding, where human anatomy is largely shared across individuals, as they collectively mitigate the impact of scarce high-quality annotations while preserving adaptability and robustness across diverse clinical scenarios [71].

Despite recent progress, medical VLMs have so far approached anatomical understanding primarily as a knowledge- and pattern-matching problem [9, 46, 68]: with sufficient supervised data, they can generalize across tasks and output formats (VQA, bounding boxes, or even segmentation when coupled with vision foundation models [26, 38]), but they remain heavily data-hungry and their reasoning is largely constrained by the medical priors accumulated during training [21, 46]. For more complex anatomical understanding tasks—such as identifying a vessel segment [47] or precisely localizing where a stenosis was after treatment [73] — labels are scarce and target structures often lack distinctive semantic cues [77]. As a result, approaches relying primarily on a model's

encoded medical knowledge are insufficient for the pixel-level localization these tasks demand. In real clinical workflows, human annotators typically solve such cases by comparing a target scan against a reference image and reasoning about their spatial relationships [47]. Motivated by this practice, we propose to explicitly steer VLMs toward reference-based spatial relationship reasoning: instead of asking the model to "know" every structure a priori, we provide a small set of high-quality annotated reference images and let the model infer anatomy in the target image by reasoning over their spatial correspondence, thereby enabling more precise and data-efficient anatomical understanding.

Here are the main contributions of this work:

- Inducing reference-based spatial reasoning in VLMs for medical images. We empirically show that targeted training enables a VLM to reason about relative spatial relations with respect to a reference image, reducing the need for large annotated datasets in anatomical understanding.
- Pixel-level anatomical region identification via VLM × SAM2. We show that the VLM's spatial reasoning capability is preserved after integration and co-training with SAM2. Leveraging SAM2's precise segmentation capability, the combined system yields accurate identification of target structures in novel images given a reference image.
- Generalization to Out-of-Distribution (OOD) data with broad applicability. Our method yields consistent improvements on OOD datasets, underscoring its robustness to distribution shifts and suggesting its potential for downstream use cases including automatic reporting, surgical navigation, and image-guided intervention planning.

2 Related Work

2.1 Anatomical Understanding with Non-VLM Pipelines

Classical anatomical understanding in medical imaging has been dominated by non-VLM pipelines based on registration and fully supervised segmentation. Atlas-based registration warps a patient scan into a standardized anatomical space so that labels can be propagated from an annotated atlas to the target image [48, 59, 62]. In practice, modern pipelines often combine strong universal segmentation baselines such as nnU-Net [22] or multi-organ foundation models [6, 65, 69] with learning-based deformable registration modules like VoxelMorph [5]. These systems can yield high-quality organ masks in relatively standardized CT or MR protocols and are now widely used as building blocks for downstream tasks.

However, these approaches have important limitations in real-world clinical and workflow scenarios. In surgical navigation (e.g., determining at which vessel segment the wire or catheter tip is located under fluoroscopy), registration to a template mask with ROI labels provides global context but tends to perform

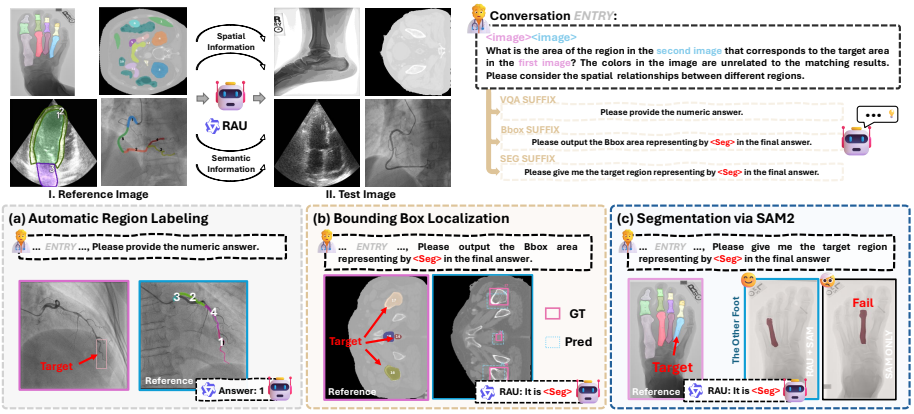

Fig. 1: RAU enables anatomical understanding with minimal reference. We explore three task modes: (a) VQA, (b) Bbox Localization (c) Segmentation with SAM2 integration. Representative cases are shown in the 2nd row. Additional results and examples are provided in later sections.

unstably due to thin and tubular structures of interventional devices and vessels, as well as motion, artifacts, incomplete views, and rapidly changing surgical scenes [43, 79]. In longitudinal studies, classical registration is a natural tool for tracking the same lesion across timepoints, but performance can degrade under large anatomical changes, inconsistent fields-of-view, or tumor shrinkage/growth, which often require expert intervention [57]. As for machine-assisted annotation, automatic masks from nnU-Net [22]/TotalSegmentator or MedSAM [41, 65] can accelerate labeling, yet their data-hungry nature [66] and sensitivity to scanner/protocol shifts [39, 50] limit deployment across institutions.

Given these limitations, our aim is not to design yet another fully supervised segmentor, but a reference-based framework for anatomical understanding. By guiding a VLM to reason about spatial relationships between a labeled reference image and an unlabeled target, our method (i) achieves OOD generalization across modalities and datasets, (ii) remains robust in complex scenes where contrast and texture cues are weak but relative locations are preserved, and (iii) naturally supports multiple output formats (VQA answers, bounding boxes, and segmentation masks) to serve different clinical tasks with minimal extra supervision.

2.2 VLM-based Anatomical Grounding and Its Limitations

VLMs offer a complementary paradigm that demonstrates strong capabilities in free-form image captioning, spatial relational reasoning, and compositional part–whole understanding [4, 8, 53]. In medical domains, such models have been adapted for image captioning and reporting, as well as for multiple-choice VQA

that aligns with structured exams or board-style questions [45,54]. Compared to conventional pipelines, VLMs exhibit superior task-level generalization: by leveraging encoded medical priors, a single model can support diverse downstream tasks and generalize more robustly across imaging modalities and acquisition qualities.

Beyond global captioning and VQA, VLMs have been extended to explicit grounding and localization. Open-vocabulary detectors such as GLIP, OWL-ViT, and Grounding DINO treat text as detection queries and output bounding boxes [1,29,35,75], enabling prompts like "wire tip" or "stenosis" to be mapped to candidate regions. Representative systems such as LISA [26] decode special segmentation tokens through SAM/Mask2Former-style backbones [74], while SEEM and EVF-SAM leverage language-as-queries and promptable segmentation to obtain masks guided by text or points [76, 80]. With ICL [12] and additional SFT and reinforcement learning [11, 32], these VLM-based pipelines can, in principle, support a unified interface for surgical navigation, diagnosis, longitudinal tracking, and annotation.

Yet despite recent progress, current VLM-based methods still fall short of clinical-grade anatomical understanding [30]. Regardless of whether the output format is free-form text, bounding boxes, or segmentation masks, the underlying reasoning primarily relies on priors accumulated during pretraining and subsequent SFT/RL, and thus remains confined to the medical knowledge in training data [7, 45]. Even when SFT and RL are used to strengthen "logical" chains of thought, this logic typically encodes what should happen in canonical disease presentations, and may still fall short when precise, spatially grounded localization is required. As a result, the best medical VLMs still struggle to achieve reliable anatomical understanding in safety-critical settings: their behavior remains data-hungry and brittle, especially for rare configurations, device interactions, or subtle negative findings [24, 33]. This limitation is partly due to the prohibitive cost of collecting large-scale, pixel-accurate annotations that cover the combinatorial complexity of real clinical workflows, and partly due to the intrinsic diversity and open-endedness of medical imaging environments. Consequently, methods that can impose anatomical structure and remain reliable in specialized applications, even under limited supervision, are especially valuable. Motivated by this, we study a reference-based spatial reasoning paradigm in which a VLM compares an unlabeled target image to a sparsely annotated template to achieve data-efficient anatomical understanding across modalities and tasks. Unlike prior work that mainly focuses on training-free SAM2 adaptation, VLM-assisted medical segmentation, or reasoning-oriented MLLM segmentation, RAU studies how the spatial inference ability a VLM acquires through VQA and bounding-box pretraining can be combined with the SAM2 memory framework for reference-conditioned segmentation and anatomical understanding, thereby separating the new task setting from the specific algorithmic contribution.

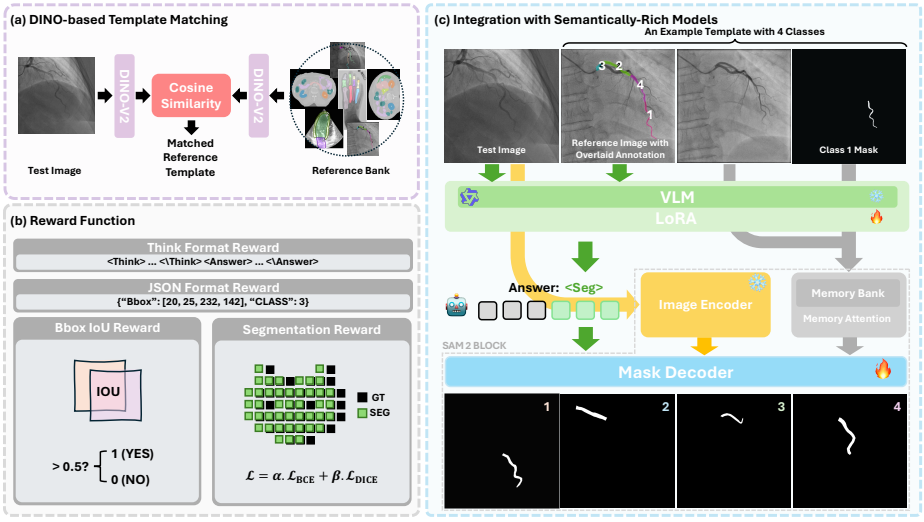

Fig. 2: Method overview. (a) For a target test image, we extract DINOv2 features and retrieve the best-matching annotated template from a reference bank using cosine similarity. (Details of the reference bank are provided in the Supplementary Material.) (b) Reinforcement-learned Qwen VLM reads the paired target–reference and generates a special <Seg>token, which is projected via an MLP adapter and injected into the SAM2 decoder as a soft spatial prompt. Reward functions include format validity and segmentation quality (Dice+BCE). (c) SAM2 leverages both the projected VLM embedding and memory attention from the reference mask to guide segmentation on the unlabeled target without explicit box/point prompts. Note that panel (b) depicts the full implementation-level reward (thinking-format, JSON-format, box, and segmentation), whereas Eqs. (2) and (5) present only the simplified task-level rewards.

3 Enhancing Anatomical Understanding in VLMs

In this section, we systematically explore how to best endow VLMs with spatial anatomical understanding. Section 3.1 formulates reference-based anatomical understanding as a VQA task, where the VLM, conditioned on a labeled reference image and an unlabeled target, answers anatomy-aware questions to localize structures. Section 3.2 then upgrades the output space from text to explicit bounding boxes, providing a more flexible and reusable representation for downstream pipelines than pure VQA-style outputs. However, bounding boxes are inherently suboptimal for thin, curved, or spatially fragmented targets (e.g., vessels or elongated devices). Section 3.3 therefore couples the VLM with a segmentation foundation model and directly predicts dense masks, enabling more precise and fine-grained anatomical understanding.

Across Sections 3.1–3.3, we employ a shared experimental protocol and dataset suite. We train on two labeled datasets (RAOS-CT [40], Arcade-X-Ray [47]) and evaluate out-of-distribution generalization on multiple external datasets (LERA-

X-Ray [58], CAMUS-Ultrasound [27], AMOS2022-MRI, CT-Liver, CT-Lung) that span diverse modalities and anatomical regions; see Supplementary Section 2 for details. In this paper, Vanilla refers to the base Qwen2.5-VL-7B model. Notably, VQA is trained on RAOS only, whereas bbox prediction and segmentation are trained on both RAOS and Arcade; since VQA is a relatively simple task, training on RAOS alone already yields good generalization to OOD datasets. For clarity of presentation, the RL settings, the segmentation reward formulation, the box-stage label-prototype construction, the practical cost matrix, the fallback rule for unmatched or low-confidence matches, and the compute, training, and inference costs are detailed in the supplementary material (Experiment Details), together with representative failure cases.

3.1 Visual Question Answering

We formulate the task in a VQA-style framework by annotating anatomical regions in the reference image and using prompt engineering to guide the VLM in identifying corresponding regions in the target image. An example prompt can be found in Figure 2b. This guides the VLM to perform spatial reasoning by grounding the query image in the labeled reference, thereby facilitating accurate target region identification.

Methods Formally, given a reference image I^{ref} with annotated regions $\{r_1, .., r_m\}$ and their labels $\{y_1, .., y_m\}$, and a target image I^{tgt}, the task is to predict the label $\hat{y}$ corresponding to a queried region r^{tgt} in I^{tgt}. We formulate this VQA interaction [3] as:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P_\theta\big(y \mid I^{\text{ref}}, \{(r_i, y_i)\}_{i=1}^{m}, I^{\text{tgt}}, r^{\text{tgt}}, \text{prompt}\big), \tag{1}$$

where P_θ is the VLM parameterized by θ and $\mathcal{Y}$ is the set of candidate anatomical labels.

We employ a two-stage strategy: first SFT, followed by RL. During SFT, the VLM is trained on labeled reference–target pairs to learn the task formulation and adapt to the medical domain. Later, GRPO is applied to further enhance performance and generalizability . The reward function is defined as a weighted combination of accuracy and format correctness:

$$R = \lambda_{\text{acc}} \cdot \mathbb{1}\{\hat{y} = y\} + \lambda_{\text{fmt}} \cdot \mathbb{1}\{\text{output format is valid}\}, \tag{2}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, and $\lambda_{\text{acc}}, \lambda_{\text{fmt}}$ are weighting coefficients. This reinforcement signal explicitly enforces the VLM to produce correct labels in the expected format, thereby fostering robust spatial reasoning ability.

Experiment Settings We first train our reference-based VQA models on RAOS-CT, aiming to teach the VLM to solve the task via spatial anatomical reasoning rather than surface pattern matching. Each training sample consists of a labeled reference image, an unlabeled target image, and a prompt that specifies

Table 1: Labeling Accuracy via VQA on in-/out-of-domain datasets. Models are trained on RAOS (650k). The number of label classes for each dataset is shown in parentheses next to its name. Mixture is constructed by combining all OOD datasets. Best results are in **bold** and second-best results are underlined.

VLM	In-Distribution (ID)	Out-of-Distribution (OOD)				
	RAOS-test-CT (20)	Arcade (26)	CT-Liver (3)	CT-Lung (3)	AMOS2022-MRI (17)	Mixture
Qwen2.5VL-7B(Vanilla)	16.62%	13.40%	53.17%	54.69%	19.45%	20.46%
MedFlamingo [44]	21.03%	14.41%	50.20%	57.95%	34.22%	22.09%
Med-VLM-R1 [46]	18.27%	12.11%	58.91%	61.13%	20.06%	21.23%
HuatuoGPT-Vision [9]	23.73%	13.85%	63.23%	60.78%	22.81%	24.33%
Lingshu [68]	22.87%	14.20%	61.81%	59.32%	19.42%	23.95%
MedRegA [63]	15.75%	12.63%	46.72%	41.59%	13.04%	16.80 %
BiRD [20]	16.49%	15.92%	57.19%	51.93%	16.76%	22.94%
MedPLIB [21]	17.28%	14.26%	54.33%	56.47%	27.81%	23.61%
SFT (3 Epochs)	41.62%	26.88%	58.51%	57.70%	45.52%	42.16%
SFT (5 Epochs)	64.11%	33.92%	77.09%	75.73%	67.58%	64.91%
GRPO (800 Steps)	42.25%	38.38%	62.03%	61.81%	41.37%	44.65%
GRPO (1600 Steps)	62.36%	41.03%	75.62%	73.33%	63.75%	62.47%
GRPO (2400 Steps)	**70.68**%	**48.37**%	**83.76**%	**80.10**%	**72.93**%	**70.84**%

the anatomical query on the reference. The model must infer the correspondence and produce an answer in a constrained format that includes only its reasoning trace and the final index of the matching region in the reference image (see Supplementary 3.2). After training on RAOS-CT, we evaluate the same model on the RAOS-CT test split and on a suite of out-of-distribution datasets—Arcade-X-Ray, LERA-X-Ray, CAMUS-Ultrasound, AMOS2022-MRI, CT-Liver, and CT-Lung—under the same VQA-style protocol, to assess generalization and robustness across modalities, anatomical regions, and imaging conditions.

Experiment Results As shown in Tab. 1, scaling up training with either SFT or RL leads to substantial improvements in labeling accuracy over the vanilla baseline. On the ID dataset RAOS-test-CT, accuracy increases from 16.62% to 64.11% with 5 epochs of training, and further to 70.68% with 2400 steps RL-based finetuning, representing a 54.06% absolute gain.

Training also yields strong cross-dataset generalization. Averaged over the five OOD datasets, accuracy improves from 32.23% with the vanilla baseline to 63.85% with SFT, and further to 71.20% with RL. The RL finetuned model achieves the best score on every OOD dataset: 48.37% on Arcade, 83.76% on CT-Liver, 80.10% on CT-Lung, 72.93% on AMOS, and 70.84% on the Mixture set, consistently surpassing SFT by 4.4 to 14.5 percentage points, depending on the dataset. Within the RL regime, accuracy improves monotonically with additional optimization steps across both ID and OOD settings, suggesting that continued policy optimization enhances generalizable decision-making rather than overfitting to the training distribution.

Qualitative inspection of CoT outputs during training reveals a convergence toward reference-guided relational reasoning, wherein the model increasingly grounds its decisions in spatial relationships between anatomical regions rather than in simple visual features. Examples can be seen in Supplementary Figure A3. Notably, the poor performance of Med-VLM-R1 [46] further corroborates this observation: when reinforcement learning is applied directly for reasoning-

based target recognition in medical images, the model often attempts to localize organs without leveraging relational priors, leading to relatively low accuracy. Similarly, MedFlamingo [44], HuatuoGPT-Vision [9], and Lingshu [68]—large-scale foundation models with stronger medical priors—also perform poorly on this task, indicating that even medically specialized models struggle to achieve robust anatomical understanding in complex settings. This further suggests that reasoning from explicit spatial relationships, rather than directly recognizing organs, is a more principled and effective strategy. Together, we show that VLMs can develop non-trivial spatial understanding and robust generalization in reference-based localization tasks, and that reasoning over spatial relations between different regions of the reference image provides a more generalizable strategy.

However, the presence of many densely arranged targets, such as intestinal loops in abdominal CT [40], in the reference image can degrade the accuracy of VQA-based approaches, underscoring the need for more precise and anatomy-aware method.

3.2 Bbox Prediction and Global Matching

In this section, we extend the task to bounding box (bbox) prediction. Rather than classifying a region label, the VLM is prompted to directly output the location of the corresponding region in the target image. However, VLMs often hallucinate numerical coordinates [34], especially under limited supervision.

Methods To address this, we use a two-stage mechanism: 1) the VLM outputs [26] one or more special `<Seg>` tokens in response to the reference-target prompt; 2) the corresponding embeddings $\mathbf{e}_{\texttt{<Seg>}} \in \mathbb{R}^d$ are passed through a lightweight MLP to regress a bounding box $\hat{\mathbf{b}} = (x, y, w, h) \in \mathbb{R}^4$:

$$\hat{\mathbf{b}} = \mathrm{MLP}(\mathbf{e}_{\texttt{<Seg>}}). \tag{3}$$

To further improve labeling accuracy, we extend the VLM's output to generate all bounding boxes corresponding to the set of anatomical labels in a single forward pass. Instead of predicting each label independently, we formulate labeling as a global matching problem: predicted boxes are jointly assigned to categories using Optimal Transport [10]. This design leverages spatial relationships across labels (i.e.,the relative positions between organs)as a soft constraint during assignment. Formally, let $\{\hat{\mathbf{b}}_i\}_{i=1}^N$ denote the predicted boxes and $\{\mathbf{p}_j\}_{j=1}^N$ be the label prototypes (e.g., expected positions or embeddings derived from the reference image). We construct a cost matrix $C \in \mathbb{R}^{N\times N}$ where C_{ij} represents the spatial or semantic distance between $\hat{\mathbf{b}}_i$ and $\mathbf{p}_j$. The optimal transport plan $\pi^* \in \mathbb{R}^{N\times N}$ is then obtained by solving:

$$\pi^* = \arg\min_{\pi\in\Pi} \sum_{i=1}^{N}\sum_{j=1}^{N} \pi_{ij}\cdot C_{ij}, \quad \text{s.t. } \pi\mathbf{1} = \mu,\ \pi^\top\mathbf{1} = \nu, \tag{4}$$

where Π is the set of doubly stochastic matrices (or relaxed transport plans), and μ, ν are marginal distributions (uniform in our case). Once the transport plan π^* determines the optimal label-to-box assignment, we apply a refinement mechanism: for unmatched or low-confidence matches (e.g., those with large cost or low attention scores), we trigger a fallback step that re-generates candidate boxes within the unassigned region. This adaptive recovery strategy improves robustness in hard cases such as small, overlapping, or occluded targets.

Training proceeds in two steps: first, the MLP is optimized with the VLM frozen; then, end-to-end fine-tuning is performed with GRPO, guided by two manually designed reward functions:

$$R = \lambda_{\text{det}} \cdot \text{AP}(\hat{\mathbf{b}}, \mathbf{b}) + \lambda_{\text{fmt}} \cdot \mathbb{1}\{\text{output format is valid}\}. \tag{5}$$

Table 2: Labeling Accuracy across Methods. Top: ID accuracy across methods. Bottom: OOD accuracy across methods. VLM+SAM2 refers to the approach proposed in Section 3.3.

Dataset	SFT-VQA(Baseline)	RL-VQA	RL-Bbox	**VLM+SAM2(Our Method)**
In-Distribution (ID): labeling accuracy (best score)				
RAOS (Whole Body CT)	64.11%	74.68%	78.16%	89.38%
Arcade (Vessel X-Ray)	53.92%	64.37%	41.09%	81.62%
Out-of-Distribution (OOD): labeling accuracy (best score)				
LERA (Bone X-Ray)	30.41%	54.57%	55.92%	61.87%
CAMUS (Heart Ultrasound)	40.29%	88.33%	55.06%	95.41%

Experiment Settings For the bbox variant, we jointly train on RAOS-CT and Arcade-X-Ray, where Arcade's challenging vessel and device configurations encourage learning non-trivial spatial layouts. As in Section 3.1, the model receives a labeled reference image, an unlabeled target image, and a reference-based instruction, and must output, in a fixed format, the coordinates of the bounding box in the reference image that corresponds to the queried region in the target (see Supplementary 3.3 for details). We then test generalization on two out-of-distribution datasets, LERA-X-Ray and CAMUS-Ultrasound, under the same reference-based VQA protocol, omitting CT/MRI datasets that are too similar to RAOS-CT. Since our aim is to probe anatomical understanding rather than low-level detection heuristics, we report labeling accuracy of the predicted bounding boxes as the primary metric.

Experiment Results As shown in Tab. 2. On the RAOS dataset, incorporating global bounding-box matching improves labeling accuracy, from 74.68% to 78.16%, indicating that spatial grounding via box-level alignment helps guide more accurate label assignment. However, this improvement does not generalize well to structurally complex datasets such as Arcade, where the bounding-box–based method yields only 41.09% accuracy. The drop highlights a key limitation: elongated or topology-fragile structures like vessels are poorly captured

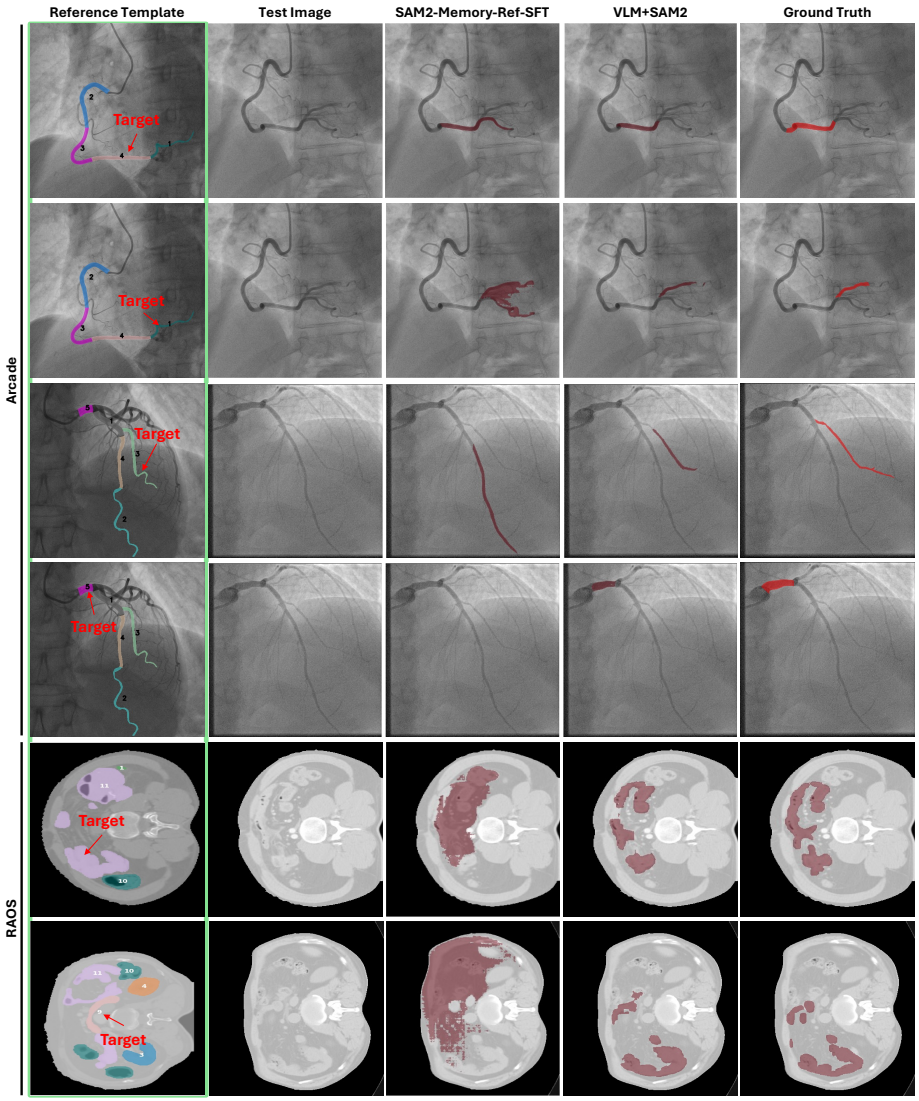

Fig. 3: In-distribution qualitative results (RAOS & Arcade). Columns: reference template, test image, SAM2-SFT w/ Memory (fine-tuned on the same memory for both datasets), VLM-SAM2, and GT. Our VLM-guided SAM2 yields tighter boundaries, better continuity on elongated/fragmented vessels, and fewer leaks/misses under mild target–atlas misalignment—resulting in visibly more accurate masks than the SAM2-SFT baseline.

by rectangular boxes, which constrains model generalization across anatomical types. Importantly, these results also suggest that reinforcement learning

enables the VLM to attend to the correct regions, yet the accuracy remains limited by the coarse nature of bounding boxes, motivating the adoption of finer-grained strategies such as segmentation-based matching for improved precision in reference-based labeling tasks.

3.3 Integration with Semantically-Rich Models

VQA- and bbox-style prompting methods perform poorly when anatomical targets are scattered, overlapping, or thin and elongated, as bounding boxes provide only coarse localization. In addition, the autoregressive nature of language modeling limits detail preservation, hindering accurate region prediction in cluttered or ambiguous scenes.

While VLMs provide strong global reasoning and instruction-following abilities, they lack explicit memory mechanisms for fine-grained semantic recall. In contrast, SAM2 introduces a learnable Memory Bank [49] architecture, where each slot encodes semantic and spatial cues from reference masks [78]. These embeddings serve as long-term anchors, supporting accurate segmentation even under ambiguous or noisy conditions. Therefore, to combine the complementary strengths of both modules, we design a fusion interface where the VLM guides the attention toward relevant regions via language reasoning and spatial reference, while SAM2 executes the segmentation grounded on learned semantic memory.

Methods Specifically, as shown in Fig. 2, the VLM generates one or more `<Seg>` tokens, whose embeddings $\{\mathbf{h}_i^{\texttt{<Seg>}}\}_{i=1}^{K}$ encode the linguistic-semantic information of the target region. These embeddings are then projected into the space of SAM2's memory slots via a shared MLP:

$$\mathbf{q}_i = \mathrm{MLP}(\mathbf{h}_i^{\texttt{<Seg>}}), \quad \mathbf{q}_i \in \mathbb{R}^d, \tag{6}$$

where d is the dimensionality of SAM2's memory bank. Next, during segmentation, these projected queries $\mathbf{q}_i$ are used to retrieve relevant memory slots $\{\mathbf{m}_j\}$ from the bank via dot-product attention:

$$\alpha_{ij} = \frac{\exp(\mathbf{q}_i^\top \mathbf{m}_j)}{\sum_k \exp(\mathbf{q}_i^\top \mathbf{m}_k)}, \quad \mathbf{z}_i = \sum_j \alpha_{ij} \cdot \mathbf{m}_j, \tag{7}$$

where $\mathbf{z}_i$ is the fused representation fed into SAM2's decoder to generate the final segmentation mask.

The VLM is initialized with the weights of RL-VQA (Sec. 3.1). In the SFT phase, we jointly train the embeddings of the `<Seg>` tokens generated by the VLM, the MLP projection layer, and the SAM2 decoder under segmentation supervision, while freezing the other VLM weights. The loss is a weighted sum of Dice [55] and binary cross-entropy (BCE) [23] losses. In the RL phase, we unfreeze the VLM and optimize it with GRPO. The reward directly reflects segmentation quality, defined as a weighted sum of Dice and BCE losses.

Table 3: Quantitative analysis of segmentation performance (Dice/gIoU; higher is better). SAM2/MedSAM2-Memory denotes using the original SAM2/MedSAM2 weights while providing the reference image as the memory input. SAM2/MedSAM2-Memory-Ref-SFT follows the same strategy but further applies SFT. Our Method corresponds to the VLM+SAM2 approach in Figure 2c.

	In-Distribution (ID)		Out-of-Distribution (OOD)	
Method (Dice/gIoU)	**Arcade**	**RAOS**	**CAMUS**	**LERA**
SAM2-Memory	0.0346 / 0.0211	0.1084 / 0.0573	0.2592 / 0.1389	0.1493 / 0.0807
MedSAM2-Memory [42]	0.0674 / 0.0486	0.1807 / 0.0925	0.3522 / 0.2101	0.1814 / 0.0996
SAM2-Memory-Ref-SFT	0.1914 / 0.1149	0.2509 / 0.1434	0.3384 / 0.2037	0.1843 / 0.1015
MedSAM2-Memory-Ref-SFT	0.2435 / 0.1852	0.2965 / 0.2094	0.4290 / 0.2772	0.2603 / 0.1723
LISA-SFT [26]	0.1278 / 0.0794	0.2070 / 0.1130	0.2778 / 0.1481	0.1578 / 0.0785
SegZero-SFT [38]	0.1326 / 0.0821	0.2145 / 0.1318	0.2580 / 0.1260	0.1799 / 0.0930
Our Method	0.6754 / 0.5099	0.7151 / 0.4320	0.7503 / 0.5133	0.7010 / 0.5396

Experiment Settings For the segmentation variant, we train on RAOS-CT and Arcade-X-Ray, and evaluate OOD performance on LERA-X-Ray and CAMUS-Ultrasound. Each training sample consists of a paired reference image and target image: the reference image is fully annotated, so a corresponding organ mask is available for every reference. The VLM configuration follows the settings in Section 3.1 and 3.2. For the SAM2 component, the reference image and its mask are encoded as memory tokens, and SAM2 uses this memory together with the VLM-provided conditioning on the target image to generate a segmentation mask. As in previous sections, our primary metric is labeling accuracy, while segmentation metrics are additionally reported to compare against traditional baselines and to verify the effectiveness of the proposed training scheme. (Details can be found in Supplementary 3.1)

Experiment Results As shown in Tab. 3, Our Method (VLM+SAM2), although fine-tuned only on RAOS and Arcade, consistently outperforms all SAM2 or MedSAM2-based baselines as well as LISA-SFT and SegZero-SFT on both ID (Arcade, RAOS) and OOD (CAMUS, LERA) datasets. Compared to the strongest SAM2 family baseline, MedSAM2-Memory-Ref-SFT, our approach improves the average Dice from 0.31 to 0.71 and the average gIoU from 0.21 to 0.50 across the four datasets, indicating substantially more reliable and fine-grained anatomical masks.

Interestingly, MedSAM2 variants, despite being pretrained on large-scale medical data, do not close this gap: when multiple candidate structures exhibit similar shapes or lack distinctive semantic cues, MedSAM2 alone cannot reliably select the correct target from the memory, leading to low Dice/gIoU. Likewise, LISA-SFT and SegZero-SFT, which possess stronger reasoning capabilities, achieve reasonable performance on ID data after SFT but degrade sharply on OOD data. This suggests that in our setting SFT largely encourages memorization of specific training images rather than true spatial generalization: anatomical correspondence in complex clinical scenarios cannot be solved by textual reasoning alone. In contrast, explicitly combining language-guided spa-

tial reasoning with a segmentation foundation model enables robust anatomical grounding that transfers across modalities and datasets.

In addition, to better connect with the previous tasks in Secs. 3.1 and 3.2, we report the labeling outcomes side-by-side in Tab. 2. On ID datasets, our method achieves the best maximum labeling accuracy—89.38% on RAOS and 81.62% on Arcade, exceeding RL-VQA and RL-Bbox by large margins. More importantly, Tab. 2 demonstrates strong OOD transfer with 61.87% on LERA and 95.41% on CAMUS using the same model.

Taken together, the evidence shows that relative spatial understanding is preserved after co-training with SAM2 integration and anatomical structure localization is enhanced by SAM2's granular, semantically rich features and memory mechanism. More importantly, this capability generalizes beyond the training distribution, enabling accurate structure segmentation without large-scale per-dataset fine-tuning.

4 Ablation Study

We conduct an ablation study to examine the effect of different VLM initialization strategies within our SAM2-enhanced pipeline, as summarized in Tab. 2. Three variants are compared: (1) Vanilla, where the VLM is directly initialized from Qwen2.5VL-7B without any task-specific tuning; (2) SFT-VQA, where the VLM is initialized from a Qwen2.5VL-7B supervisedly fine-tuned on the VQA task (Sec. 3.1); and (3) RL-VQA (used in VLM+SAM2), where the VLM is initialized from the model optimized via reinforcement learning on the VQA task.

Initialization with SFT-VQA significantly improves performance on ID datasets. On RAOS and Arcade, it improves over the instruct-only baseline by large margins. However, this improvement does not generalize well to OOD datasets. For instance, while the SFT model achieves 76.36% on CAMUS, its performance on LERA remains poor (33.95%), suggesting it may overfit to dataset-specific patterns. In contrast, initialization with RL-VQA consistently improves both ID and OOD performance. On LERA, it improves labeling accuracy to 61.87%, and similarly boosts CAMUS accuracy to 95.41%. These results indicate that reinforcement fine-tuning with the end-task reward not only leads to better generalization in the trained VQA task, but also improves OOD generalization when integrated with SAM2. We also conduct an ablation study on the quality of the reference images, with results reported in Supplementary Table S2.

To further isolate the contribution of the proposed training recipe under the segmentation metric (Dice/gIoU), we report a complementary ablation in Tab. 5. RAU (VLM–Vanilla) and RAU (VLM–SFT) keep the overall reference-conditioned pipeline unchanged while replacing the VLM with off-the-shelf Qwen weights and an SFT-only variant, respectively; the former establishes the reference-conditioned prompting baseline, whereas the latter ablates the RL stage. To contextualize the upper bound of prompt-based segmentation, we additionally prompt SAM2 with ground-truth bounding boxes; this oracle-localized baseline

Table 4: Ablation on VLM initialization in the SAM2-enhanced pipeline. Top: maximum (best) labeling accuracy across methods on ID datasets. Bottom: maximum (best) labeling accuracy across methods on OOD datasets.

Dataset	Vanilla	SFT-VQA	RL-VQA
In-Distribution (ID): labeling accuracy (best score)			
RAOS (Whole Body CT)	18.23%	85.07%	89.38%
Arcade (Vessel X-Ray)	15.41%	76.93%	81.62%
Out-of-Distribution (OOD): labeling accuracy (best score)			
LERA (Bone X-Ray)	30.12%	33.95%	61.87%
CAMUS (Heart Ultrasound)	69.18%	76.36%	95.41%

still falls short of the full RAU model, indicating that box prompts alone are insufficient for thin and topology-fragile structures. Finally, we split each reference image into upper and lower halves and shuffle them, which preserves the semantic content of the reference anatomy but disrupts its spatial arrangement. Performance drops sharply under this up–down split-and-shuffled reference, showing that the model relies on the spatial correspondence between the reference and the target rather than merely exploiting semantic category cues.

Table 5: Segmentation-metric ablation (Dice/gIoU; higher is better). RAU (Full) is our complete model. The up–down split-and-shuffled reference is only applicable to ID datasets.

	In-Distribution (ID)		**Out-of-Distribution (OOD)**	
Method (Dice/gIoU)	**Arcade**	**RAOS**	**CAMUS**	**LERA**
RAU (VLM–Vanilla)	0.1072 / 0.0713	0.1968 / 0.1105	0.2644 / 0.1196	0.1231 / 0.0650
RAU (VLM–SFT)	0.4370 / 0.3042	0.5093 / 0.3371	0.5506 / 0.3737	0.5041 / 0.3995
SAM2 (with GT Bbox)	0.2168 / 0.1749	0.2972 / 0.2631	0.4480 / 0.3008	0.3216 / 0.2637
Shuffled ref. image	0.0966 / 0.0642	0.2088 / 0.1369	- / -	- / -
RAU (Full)	0.6754 / 0.5099	0.7151 / 0.4320	0.7503 / 0.5133	0.7010 / 0.5396

5 Conclusion and Discussion

We introduced RAU, a reference-based framework that uses a VLM to localize anatomy via relative relationships between a labeled reference and an unlabeled target, and integrates SAM2 to extend region-level reasoning to pixel-level segmentation. Across modalities and anatomical targets, both in- and out-of-distribution, RAU performs consistently in challenging settings such as vessel segment labeling and ultrasound interpretation.

Looking ahead, we will: (i) incorporate structural priors (e.g., part–whole hierarchies or organ trees) to enforce consistency; (ii) develop adaptive memory to better handle viewpoint and shape variation; and (iii) extend RAU to temporal data or 3D volumes. RAU offers a scalable basis for anatomical understanding under limited supervision and a step toward clinically practical systems.

References

1. Alhadidi, T., Jaber, A., Jaradat, S., Ashqar, H., Elhenawy, M.: Object detection using oriented window learning vision transformer: Roadway assets recognition. In: International Conference on Intelligent Systems, Blockchain, and Communication Technologies. pp. 506–522. Springer (2024)
2. Alozai, M.I., Elassra, O.A.Y., Alkhazendar, A.H., Ibrahim, A.S., Gatta, A.S., Raza, S.M.B., Sahnon, A.S.A., Alkhazendar, J.H., Oriko, D.O., Mushtaq, S., et al.: The impact of intraoperative imaging on outcomes in combined neurosurgical and reconstructive procedures: A systematic review. Cureus **17**(6) (2025)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
4. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
5. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
6. Berrezueta, S., Baldeon-Calisto, M., Navarrete, D., Pérez-Pérez, N., Flores-Moyano, R., Riofrío, D., Benítez, D.: Foundation models for medical image segmentation: A literature review. In: 2025 13th International Symposium on Digital Forensics and Security (ISDFS). pp. 1–7. IEEE (2025)
7. Bian, Y., Li, J., Ye, C., Jia, X., Yang, Q.: Artificial intelligence in medical imaging: From task-specific models to large-scale foundation models. Chinese Medical Journal **138**(06), 651–663 (2025)
8. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
9. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., et al.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280 (2024)
10. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
11. Dong, G., Yuan, H., Lu, K., Li, C., Xue, M., Liu, D., Wang, W., Yuan, Z., Zhou, C., Zhou, J.: How abilities in large language models are affected by supervised fine-tuning data composition. arXiv preprint arXiv:2310.05492 (2023)
12. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., et al.: A survey on in-context learning. arXiv preprint arXiv:2301.00234 (2022)
13. Dutt, R., Ericsson, L., Sanchez, P., Tsaftaris, S.A., Hospedales, T.M.: Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. In: Proceedings of Medical Imaging with Deep Learning (MIDL). pp. 1–20. Paris, France (2024)
14. Fan, K., Liang, L., Li, H., Situ, W., Zhao, W., Li, G.: Research on medical image segmentation based on sam and its future prospects. Bioengineering **12**(6), 608 (2025)
15. Gao, Y., Jiang, Y., Peng, Y., Yuan, F., Zhang, X., Wang, J.: Medical image segmentation: A comprehensive review of deep learning-based methods. Tomography **11**(5), 52 (2025)

16. Gaudioso, P., Contro, G., Taboni, S., Costantino, P., Visconti, F., Sozzi, M., Borsetto, D., Sharma, R., De Almeida, J., Verillaud, B., et al.: Intraoperative surgical navigation as a precision medicine tool in sinonasal and craniofacial oncologic surgery. Oral Oncology **157**, 106979 (2024)
17. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
18. Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: A review. Frontiers in artificial intelligence **7**, 1430984 (2024)
19. Hu, M., Qian, J., et al.: Advancing medical imaging with language models. Patterns (2024), review; PMC11075180
20. Huang, X., Huang, H., Shen, L., Yang, Y., Shang, F., Liu, J., Liu, J.: A refer-and-ground multimodal large language model for biomedicine. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 399–409. Springer (2024)
21. Huang, X., Shen, L., Liu, J., Shang, F., Li, H., Huang, H., Yang, Y.: Towards a multimodal large language model with pixel-level insight for biomedicine. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3779–3787 (2025)
22. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
23. Jadon, S.: A survey of loss functions for semantic segmentation. In: 2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB). pp. 1–7. IEEE (2020)
24. Jin, C., Guo, Z., Lin, Y., Luo, L., Chen, H.: Label-efficient deep learning in medical image analysis: Challenges and future directions. arXiv preprint arXiv:2303.12484 (2023)
25. Khan, D.Z., Valetopoulou, A., Das, A., Hanrahan, J.G., Williams, S.C., Bano, S., Borg, A., Dorward, N.L., Barbarisi, S., Culshaw, L., et al.: Artificial intelligence assisted operative anatomy recognition in endoscopic pituitary surgery. NPJ Digital Medicine **7**(1), 314 (2024)
26. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)
27. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging **38**(9), 2198–2210 (2019)
28. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
29. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10975 (2022)
30. Li, Y., Kim, S., Wu, Z., Jiang, H., Pan, Y., Jin, P., Song, S., Shi, Y., Liu, T., Li, Q., et al.: Echopulse: Ecg controlled echocardio-grams video generation. arXiv preprint arXiv:2410.03143 (2024)

31. Lian, C., Zhou, H.Y., Yu, Y., Wang, L.: Less could be better: Parameter-efficient fine-tuning advances medical vision foundation models. arXiv preprint arXiv:2401.12215 (2024). https://doi.org/10.48550/arXiv.2401.12215
32. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
33. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
35. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
36. Liu, S., McCoy, A.B., Wright, A.: Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. Journal of the American Medical Informatics Association **32**(4), 605–615 (2025). https://doi.org/10.1093/jamia/ocaf008
37. Liu, X., Wang, X., Chen, D., Zhang, H.: Automatic quantitative coronary analysis based on deep learning. Applied Sciences **13**(5) (2023). https://doi.org/10.3390/app13052975, https://www.mdpi.com/2076-3417/13/5/2975
38. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
39. Liu, Z., Li, Y., Shu, P., Zhong, A., Jiang, H., Pan, Y., Yang, L., Ju, C., Wu, Z., Ma, C., et al.: Radiology-gpt: a large language model for radiology. Meta-Radiology p. 100153 (2025)
40. Luo, X., Li, Z., Zhang, S., Liao, W., Wang, G.: Rethinking abdominal organ segmentation (raos) in the clinical scenario: A robustness evaluation benchmark with challenging cases. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 531–541. Springer (2024)
41. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
42. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025)
43. Matl, S., Brosig, R., Baust, M., Navab, N., Demirci, S.: Vascular image registration techniques: A living review. Medical image analysis **35**, 1–17 (2017)
44. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). pp. 353–367. PMLR (2023)
45. Nam, Y., Kim, D.Y., Kyung, S., Seo, J., Song, J.M., Kwon, J., Kim, J., Jo, W., Park, H., Sung, J., et al.: Multimodal large language models in medical imaging: Current state and future directions. Korean Journal of Radiology **26**(10), 900–923 (2025)
46. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language

models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347. Springer (2025)
47. Popov, M., Amanturdieva, A., Zhaksylyk, N., Alkanov, A., Saniyazbekov, A., Aimyshev, T., Ismailov, E., Bulegenov, A., Kuzhukeyev, A., Kulanbayeva, A., et al.: Dataset for automatic region-based coronary artery disease diagnostics using x-ray angiography images. Scientific data **11**(1), 20 (2024)
48. Ramadan, H., El Bourakadi, D., Yahyaouy, A., Tairi, H.: Medical image registration in the era of transformers: A recent review. Informatics in Medicine Unlocked **49**, 101540 (2024)
49. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
50. Rayed, M.E., Islam, S.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.: Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in medicine unlocked **47**, 101504 (2024)
51. Serruys, P.W., Reiber, J.H., Wijns, W., v.d. Brand, M., Kooijman, C.J., ten Katen, H.J., Hugenholtz, P.G.: Assessment of percutaneous transluminal coronary angioplasty by quantitative coronary angiography: Diameter versus densitometric area measurements. The American Journal of Cardiology **54**(6), 482–488 (1984). https://doi.org/https://doi.org/10.1016/0002-9149(84)90235-2, https://www.sciencedirect.com/science/article/pii/0002914984902352
52. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
53. Shen, X., Han, D., Chen, C., Luo, G., Wu, Z.: An effective spatial relational reasoning networks for visual question answering. Plos one **17**(11), e0277693 (2022)
54. Stogiannidis, I., McDonagh, S., Tsaftaris, S.A.: Mind the gap: Benchmarking spatial reasoning in vision-language models. arXiv preprint arXiv:2503.19707 (2025)
55. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: International Workshop on Deep Learning in Medical Image Analysis. pp. 240–248. Springer (2017)
56. Tanno, R., Barrett, D.G.T., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Lau, C., Tu, T., Azizi, S., Singhal, K., et al.: Collaboration between clinicians and vision–language models in radiology report generation. Nature Medicine **31**, 599–608 (2025). https://doi.org/10.1038/s41591-024-03302-1
57. Tanno, R., Barrett, D.G., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Singhal, K., Azizi, S., Tu, T., et al.: Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation. arXiv preprint arXiv:2311.18260 (2023)
58. Varma, M., Lu, M., Gardner, R., Dunnmon, J., Khandwala, N., Rajpurkar, P., Long, J., Beaulieu, C., Shpanskaya, K., Fei-Fei, L., et al.: Automated abnormality detection in lower extremity radiographs using deep learning. Nature Machine Intelligence **1**(12), 578–583 (2019)
59. Varol Arısoy, M., Arısoy, A., Uysal, İ.: A vision attention driven language framework for medical report generation. Scientific Reports **15**(1), 10704 (2025)
60. van Veldhuizen, V., Botha, V., Lu, C., Cesur, M.E., Lipman, K.G., de Jong, E.D., Horlings, H., Sanchez, C.I., Snoek, C.G., Wessels, L., et al.: Foundation models in medical imaging–a review and outlook. arXiv preprint arXiv:2506.09095 (2025)

61. Wang, H., Jin, Q., Li, S., Liu, S., Wang, M., Song, Z.: A comprehensive survey on deep active learning in medical image analysis. Medical Image Analysis **95**, 103201 (2024)
62. Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A.: Multi-atlas segmentation with joint label fusion. IEEE transactions on pattern analysis and machine intelligence **35**(3), 611–623 (2012)
63. Wang, L., Wang, H., Yang, H., Mao, J., Yang, Z., Shen, J., Li, X.: Interpretable bilingual multimodal large language model for diverse biomedical tasks. arXiv preprint arXiv:2410.18387 (2024)
64. Wang, X., Figueredo, G., Li, R., Zhang, W.E., Chen, W., Chen, X.: A survey of deep-learning-based radiology report generation using multimodal inputs. Medical Image Analysis p. 103627 (2025)
65. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
66. Wu, Z., Zhang, L., Cao, C., Yu, X., Liu, Z., Zhao, L., Li, Y., Dai, H., Ma, C., Li, G., et al.: Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology nli task. IEEE Transactions on Big Data (2025)
67. Xia, Y., Pan, Z., Hu, L., Zhu, Z., et al.: Mmed-rag: Benchmarking RAG for medical multimodal LLMs. In: The Thirteenth International Conference on Learning Representations (ICLR) (2025)
68. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
69. Yan, Z., Song, S., Song, D., Li, Y., Zhou, R., Sun, W., Chen, Z., Kim, S., Ren, H., Liu, T., et al.: Samed-2: Selective memory enhanced medical segment anything model. arXiv preprint arXiv:2507.03698 (2025)
70. Yang, R., Ning, Y., Liu, N.: Retrieval-augmented generation for generative artificial intelligence in health care. npj Health Systems (2025), perspective
71. Yang, T., Jordan, T., Liu, N., Sun, J.: Common inpainted objects in-n-out of context. arXiv preprint arXiv:2506.00721 (2025)
72. Yang, T., Shi, Y., Du, M., Wu, X., Tan, Q., Sun, J., Liu, N.: Concept-centric token interpretation for vector-quantized generative models. arXiv preprint arXiv:2506.00698 (2025)
73. Zanchin, C., Yamaji, K., Rogge, C., Lesche, D., Zanchin, T., Ueki, Y., Windecker, S., Mohacsi, P., Räber, L., Sigurdardottir, V.: Progression of cardiac allograft vasculopathy assessed by serial three-vessel quantitative coronary angiography. Plos one **13**(8), e0202950 (2018)
74. Zhang, G., Navasardyan, S., Chen, L., Zhao, Y., Wei, Y., Shi, H., et al.: Mask matching transformer for few-shot segmentation. Advances in Neural Information Processing Systems **35**, 823–836 (2022)
75. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
76. Zhang, Y., Cheng, T., Zhu, L., Hu, R., Liu, L., Liu, H., Ran, L., Chen, X., Liu, W., Wang, X.: Evf-sam: Early vision-language fusion for text-prompted segment anything model. arXiv preprint arXiv:2406.20076 (2024)

77. Zhao, C., Xu, Z., Baral, P., Esposito, M., Zhou, W.: Multi-graph graph matching for coronary artery semantic labeling in invasive coronary angiograms. Pattern Recognition (2025)
78. Zhao, L., Chen, X., Chen, E.Z., Liu, Y., Chen, T., Sun, S.: Retrieval-augmented few-shot medical image segmentation with foundation models. IEEE Transactions on Neural Networks and Learning Systems (2025)
79. Zhu, J., Fan, J., Guo, S., Ai, D., Song, H., Wang, C., Zhou, S., Yang, J.: Heuristic tree searching for pose-independent 3d/2d rigid registration of vessel structures. Physics in Medicine & Biology **65**(5), 055010 (2020)
80. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in neural information processing systems **36**, 19769–19782 (2023)