arXiv:2603.01153v1 [cs.RO] 1 Mar 2026

RAG-RUSS：用于自主颈动脉检查的检索增强机器人超声

Dianye Huang[1, 2, *]，Ziping Cong[1, *]，Nassir Navab[1,2]，*IEEE* 会士，以及 Zhongliang Jiang[1, 2, †]

摘要—机器人超声（US）近年来受到越来越多的关注，被视为克服传统超声检查局限性的一种方式，例如其对操作者的高度依赖。然而，现有方法的决策过程往往基于规则，或依赖作为黑箱运行的端到端学习模型。这被视为临床接受度的一大障碍，也引发了在常规实践中广泛应用的安全问题。为应对这一挑战，我们提出 RAG-RUSS，这是一种可解释框架，能够按照临床工作流程完成全面的颈动脉检查，同时明确说明当前阶段和下一步计划采取的操作。此外，鉴于医学数据稀缺，我们引入检索增强生成，以增强泛化能力并减少对大规模训练数据集的依赖。该方法使用采集自 28 名志愿者的数据进行训练，并保留了另外四次来自此前未见志愿者的容积扫描用于测试。结果表明，该方法能够说明当前扫描阶段，并自主规划探头运动以完成颈动脉检查，涵盖横向和纵向切面。

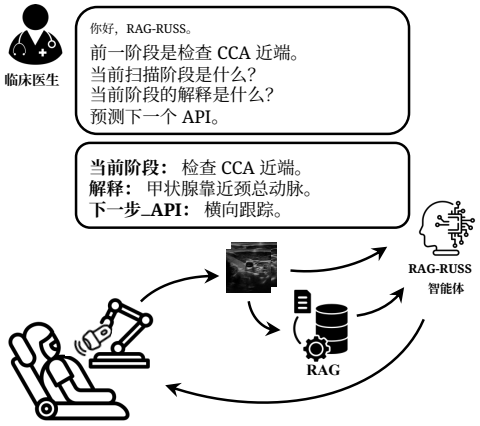


图1. RAG-RUSS 系统的典型应用场景示意图。收到临床医生的查询后，RAG-RUSS 通过检索增强生成（RAG）检索相似的已标注扫描图像，随后识别阶段、生成解释并预测下一步 API。机器人机械臂执行该 API，并将采集到的超声图像反馈给系统，从而形成闭环，实现自主且可解释的颈动脉扫描。

I. 引言

超声（US）是一种无创、实时且易于普及的成像模态，因此成为当代临床实践中的主要诊断工具 [1]。然而，超声成像的质量以及获取标准成像平面的效率高度依赖于操作者的探头操作熟练程度。对操作者专业技能的依赖会导致检查结果存在差异，且需要多年培训才能达到熟练水平，这进一步加剧了人力短缺问题，在资源匮乏地区尤为如此 [2]。

为了减轻医生的工作负担并减少检查结果的差异，过去二十多年来，自主机器人超声扫描系统（RUSS）得到了广泛研究，并持续受到关注 [3]。人们已开发出多种 RUSS，适用于广泛的临床应用，包括乳腺自主筛查 [4], 肺 [5], 和血管 [6], [7]。这些系统主要侧重于利用超声成像反馈和/或外部RGB-D相机来实现筛查过程的自动化。Wang 等人 [8] 提出了一种基于视觉伺服的自主颈动脉超声扫描系统，使用改进的孪生网络来跟踪目标血管。Huang 等人 [9] 提出了一种两阶段流程，用于在横向与纵切面之间自动切换，将阻抗/力控制与图像分割相结合。然而，这些系统主要侧重于运动控制，并严重依赖基于规则的策略，难以展现更高层次的理解能力和动态决策能力。

为解决这些局限，基于学习的 RUSS 已成为一个颇具前景的方向，能够实现更智能、数据驱动的感知与控制 [10]。为减少对人工制定规则的依赖，Huang 等人 [11] 引入临床模仿学习（IL）框架，以复现经验丰富的医生的操作习惯。基于当前观测，研究人员还采用强化学习（RL）推断下一步控制指令，引导超声探头移动至标准成像平面 [12]–[14]。为解决超声成像中的操作者间差异，Jiang 等人提出了一种智能机器人超声医师，通过学习专家级扫描策略，自主导航至标准成像平面。在他们的研究中，神经奖励函数通过自监督方式利用排序式成对图像比较进行训练，从而捕捉高层次的生理学知识。尽管上述方法取得了令人鼓舞的成果，RUSS 的实际部署仍受制于模仿学习与强化学习范式固有的不透明性。这些数据驱动方法通常仅通过图像—轨迹演示对编码专业知识，形成反映人类决策不透明性的黑箱策略。因此，能够让医生监控已训练策略内部状态的透明系统设计变得至关重要 [10]。

* 共同贡献。† 通讯作者（电子邮箱：zl.jiang@tum.de）。[1] 慕尼黑工业大学（TUM）计算机辅助医疗操作与增强现实（CAMP），德国，[2] 慕尼黑机器学习中心（MCML），德国

因此，策略透明度已成为一项关键需求 [15]。围绕这一需求，语言自然成为连接真实临床环境中的医疗机器人、医生与患者的桥梁。值得注意的是，要实现这种具身沟通能力，必须将低层次感知输入与高层次语义推理相结合，而传统的基于规则的方法以及模仿学习/强化学习方法在这方面仍存在局限 [16], [17]。

近年来，大语言模型（LLM）展现出强大的联想能力和丰富的常识知识，推动了具身人工智能的发展 [18], [19]。受此进展启发，研究界开始探索利用视觉语言模型（VLM）和视觉-语言-动作（VLA）框架来赋能智能交互智能体。Wang 等人 提出了 EndoChat，这是一种专为机器人辅助手术中的多种对话任务设计的多模态大语言模型（MLLM），在 Surg-396K 数据集上进行训练。Ng等人 [20]推出了 EndoVLA，这是一种面向胃肠道介入治疗中连续体机器人的 VLA 模型，可实现语义息肉跟踪、异常黏膜区域描绘以及环形切割标记跟踪。在超声领域，Xu 等人推出了 USPilot [21]，这是一款由大语言模型驱动的具身机器人超声助手，可充当虚拟超声医师——回答患者的问题，通过微调理解超声特定任务，并根据用户意图调用应用程序接口。该系统将经过微调的大语言模型与大语言模型增强的图神经网络相结合，用于应用程序接口控制和任务规划。然而，其主要侧重于意图理解和高层规划，实时图像内容分析能力有限，因此难以有效执行动态任务。另一方面，Jiang 等人 [22] 提出了 UltraBot，这是一种用于颈动脉超声检查的系统，在一个大规模数据集上训练，该数据集包含 247 k 组图像-动作专家示范。UltraBot 采用端到端模仿学习，直接将实时超声图像映射为 6-DoF 探头运动，同时将扫描、生物测量、斑块分割和报告生成整合到统一工作流程中。然而，UltraBot 的决策过程仍是一个黑箱，未对中间推理步骤或分阶段任务分解进行显式建模。尽管基于大模型的方法已展现出良好前景，但它们需要大量特定领域的训练数据集，而这些数据集的采集成本往往高得令人难以承受。为缓解这一局限，并受到自动驾驶领域 RAG-Driver [23] 的启发，我们提出引入检索增强生成（RAG）组件，以减少对大型训练数据集的依赖，并缓解医学领域的数据稀缺瓶颈。

在这项工作中，我们提出了 RAG-RUSS，这是一种用于颈动脉超声扫描的检索增强框架，通过将视觉语言模型与 RAG组件相结合，降低对大型训练语料库的依赖。不同于 USPilot 和 UltraBot 等侧重于意图理解或直接端到端动作映射的系统，我们的方法利用检索到的相似扫描上下文，实现图像理解和在闭环执行过程中基于解释进行决策。为了训练 RAG-RUSS，我们还基于 32 例人体容积扫描构建了一个高质量数据集。在临床定义的工作流程中，RAG-RUSS 可预测当前阶段，并在提供逐步解释的同时给出下一步 API。该设计能够借鉴既往病例中的证据，实现标准化的颈动脉扫描，并在整个检查过程中保持透明度。据我们所知，这是首个将感知、解释和动作整合到统一框架中、用于颈动脉检查的机器人超声系统。通过解释当前图像和后续动作，RAG-RUSS 朝着可信且临床可接受的部署迈出了重要一步。

本文其余部分安排如下：第二节概述颈动脉扫描的临床工作流程及数据准备过程。第三节 介绍所提出的 RAG-RUSS 架构。第四节 描述实验设置并报告结果。最后，第五节总结本研究。

II. 检查工作流程与数据准备

A. 颈动脉超声检查工作流程

1）颈动脉检查的临床工作流程：颈动脉位于颈部气管两侧，是将富含氧气的血液从心脏输送至大脑和面部的主要血管。每条颈动脉由颈总动脉（CCA）构成，颈总动脉分叉为颈内动脉（ICA）和颈外动脉（ECA）；前者向大脑和眼睛供血，后者向面部和头皮供血（见图 2）。动脉粥样硬化是最常见的颈动脉疾病之一，斑块积聚会导致颈动脉狭窄（管腔变窄），并增加卒中风险，尤其是老年人群。在此背景下，医学超声凭借无辐射和可实时成像的优势，已成为一线标准诊断工具。

进行常规颈动脉检查时，超声医师通常遵循标准工作流程 [9], [11]。为充分满足临床诊断目的，超声医师需要全面检查颈总动脉及其分叉形成的颈内动脉和颈外动脉分支。检查通常先沿管腔中心线进行横向（短轴）平面扫描，再对颈动脉进行纵向（长轴）平面扫描 [8]。扫描通常从颈总动脉近端的横切面开始，向远端延伸至颈动脉球部（颈动脉窦），并越过分叉处以观察颈内动脉和颈外动脉。随后，医生切换至纵切面进行详细评估和测量，例如测量血管直径。

2）扫描阶段定义：为了自动化扫描流程，并帮助 RUSS 全面理解检查的整体进度，我们采用上述临床工作流程构建自动化系统，并遵循 [9], [22]。在本研究中，我们将颈动脉扫描过程划分为八个连续阶段（见图2），每个阶段均由超声图像中不同的解剖学线索引导。

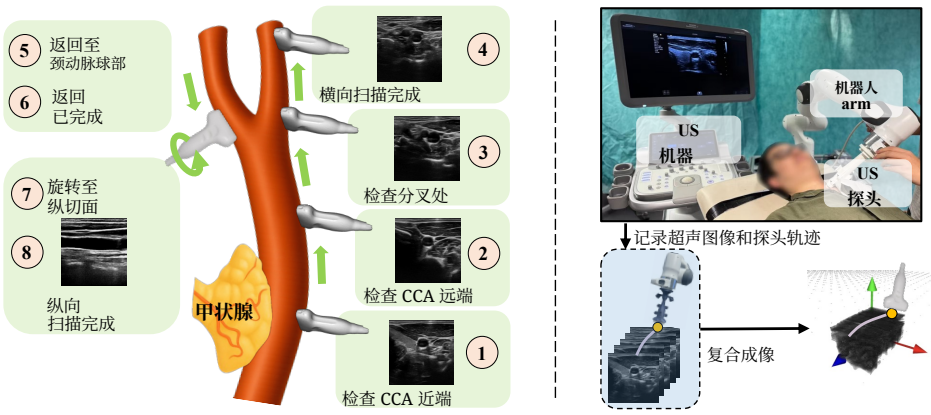


图2. 颈动脉超声扫描工作流程及数据采集/处理流程概览。左侧：展示颈动脉检查预先定义的八个扫描阶段。右侧：记录超声图像和探头的位姿轨迹，并随后将其合成为三维容积数据，用于仿真环境。

①检查 CCA 近端：颈总动脉与甲状腺同时显影。2 检查 CCA 远端：颈总动脉仍可见，但图像中已看不到甲状腺。3 检查分叉处：到达颈动脉球部；管腔开始分为颈外动脉和颈内动脉。4 横向扫描完成：管腔清晰分为颈内动脉和颈外动脉。5 返回颈动脉球部：探头向分叉处移回时，颈内动脉和颈外动脉逐渐汇合为单一管腔。6 返回完成：颈内动脉和颈外动脉在颈动脉球部汇合为颈总动脉。7 旋转至纵向视图：动脉横截面由椭圆形变为细长的矩形轮廓。8 纵向扫描完成：清晰显示颈动脉的纵向横截面。

B. 超 基于人体容积数据的超声仿真 *Data*

出于安全问题考虑，直接从志愿者处采集所有所需训练扫描轨迹并不现实。遵循以往研究 [14]，，我们基于记录的人体体积数据构建了仿真环境，并在其中初始化虚拟探头以执行连续扫描。每组配对的模拟超声图像及其对应的探头轨迹均被视为一次完整颈动脉扫描的示范。值得注意的是，单个预先记录的体积可以生成多个示范。为降低过拟合风险，我们从32名健康志愿者的扫描中采集并重建了总计32个颈动脉三维超声体积。

1）人体容积数据采集：图2所示的数据采集平台由一台Panda机械臂（Franka Emika，德国）和一套ACUSON Juniper超声（US）系统（Siemens，德国）组成。线性超声探头（12L3，Siemens，德国）通过定制三维打印探头支架安装在机器人末端执行器上。超声图像通过一台采集卡（USB Capture HDMI，MAGEWELL，中国）以30 Hz的频率采集。同时，通过ROS接口以约100 Hz的较高频率记录机器人跟踪数据，以补偿两路数据流之间可能出现的时间错位。将二维超声图像与对应的三维空间中的机器人位姿进行堆叠，即可使用PLUS [24]等经典插值技术生成重建三维超声体积。图2展示了一个重建体积示例。我们从32名健康志愿者中共采集了32个颈动脉体积。我们对每个体积进行了人工检查，以确保血管形态清晰可见且分叉结构完整捕获。为展示该群体的多样性，表I总结了志愿者的人口统计与身体信息。

使用 OpenGL 对复合颈动脉容积进行可视化。通过初始化虚拟超声探头，我们可以模拟上一节讨论的标准检查工作流程，生成配对探头位姿与 B 型图像的颈动脉检查演示。通过调整虚拟探头位置，可以获得一组标准化颈动脉检查演示。这种基于人体容积数据的仿真在保持医学超声图像保真度的同时，提供了灵活且实用的可控扫描轨迹生成。

表 I 志愿者统计（N=32）。

年龄（岁）	身高（m）	体重（kg）	BMI
27.97 ± 3.91	1.74 ± 0.09	70.22 ± 12.24	23.06 ± 2.86

2）专家颈动脉检查演示生成：本节介绍如何利用重建的人体体积数据生成颈动脉检查专家演示的实现细节。首先，根据可视化的血管特征，手动定义一组对应不同扫描阶段的解剖航点。为规范扫描轨迹生成，我们进一步计算标注后的二值血管掩膜中的血管质心点，以获得更精细的扫描轨迹。

随后，通过连接这些精细化航点并调整特定阶段转换处的方向，即可生成扫描轨迹。在仿真中自动执行这条连续扫描轨迹后，便可获取各个探头位置对应的B型图像。

对于每个重建体积，均进行了三次完整扫描，从颈总动脉近端段扫描至纵切面可视化，如图2所示。为平衡不同扫描阶段的样本分布，对样本相对较少的个别阶段额外采集了20次扫描，包括横向扫描完成、返回完成和纵向扫描完成。在横向扫描阶段，探头每步前进1 mm。在返回阶段，第一步回撤2 mm，之后每步回撤1.5 mm。为使同一志愿者获得的扫描结果有所差异，每一步的运动指令均加入了随机扰动。具体而言，沿纵向的平移偏差最大为 ± 0.3 mm，沿横向的平移偏差最大为 ± 0.2mm。

C. 数据集结构

为训练所提出的 RAG-RUSS，使其能够与人类用户交互，我们设计的数据集同时包含视觉和文本信息。我们采用了长度为 N 步的滑动时间窗口。对于窗口内的每个条目，除了超声图像外，我们还定义了三类文本注释：（1）当前扫描阶段，例如阶段1检查 *CCA* 近端 和阶段2检查 *CCA* 远端 （见图2）；（2）说明该阶段特征性解剖特征的文字；（3）完成给定任务所需的下一个 API 名称。各阶段的说明由我们的临床合作者根据颈动脉检查中用于区分扫描阶段的标准提供。例如，阶段1和阶段2的说明分别为“甲状腺靠近颈总动脉”和“未见甲状腺”。本研究定义了三个 API：“向前跟踪”“向后跟踪”和“顺时针旋转”，分别对应于推进探头、回撤探头，以及将探头从横断面旋转至纵切面。我们总共收集了15,459个条目（长度为 N 的滑动时间窗口），基于32个重建体积，包含成对的视觉和文本信息。图像分辨率为 224×224 像素。为避免数据泄漏，我们将32名志愿者中4人的数据留作测试集。

针对不同的训练目标，构建了三个子数据集：(1) 用于训练 RAG 模块的数据集 A；(2) 用于预训练跨模态投影器（2 层多层感知机），以对齐视觉和文本特征的数据集 B；(3) 使用 LoRA适配器 [25]，与视觉语言模型主干网络共同微调跨模态多层感知机的数据集 C。数据集 A 包含每个移动窗口内的首帧和末帧图像，以及与末帧对应的文本阶段描述。总计纳入了来自 28 名志愿者的 12,937 条记录。训练时，我们遍历每条记录，并生成 20 组成对的正样本和负样本，以支持对比学习。数据集 B 由单张 B 模式图像构成，每张图像均配有预定义的、针对特定阶段的查询及相应回答，用于描述各阶段的特征性解剖特征。数据集 C 旨在优化大语言模型骨干，并包含涉及当前阶段、下一个 API 以及相关指令的多轮问答对，从而实现与人类用户的有效交互。

III. 方法

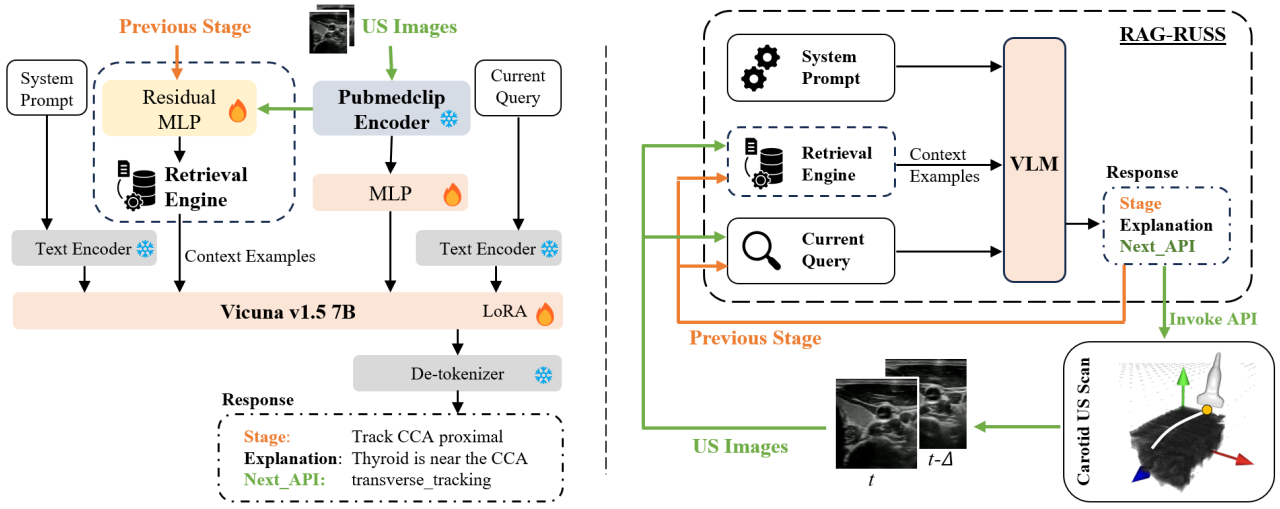
本节详细介绍 RAG-RUSS 的结构和训练流程。RAG-RUSS 是一种智能超声智能体，会在每个时间步说明当前扫描阶段，使医生能够监控进度，并在操作偏离计划时进行干预。因此，智能体与医生之间进行有效、自然的交流至关重要。为此，RAG-RUSS 以大语言模型为核心，利用其习得的“常识”和进行自然语言交流的内在能力，同时无缝融入医生的专业知识。

A. RAG-RUSS 的结构

如图 3(a)所示，RAG-RUSS 由大语言模型、基于大规模医学数据训练的视觉基础模型，以及用于提取相关知识以支持决策的检索引擎组成。如图 3(b)所示，在推理过程中，RAG-RUSS 采用多轮问答流程来处理三个查询：识别当前阶段、生成解释，以及预测下一个 API。我们还引入了多模态上下文学习（ICL），使模型能够参考视觉语言历史信息，从而提高扫描的准确性和可靠性。在每个查询步骤中，RAG-RUSS 都会生成可解释文本，确保整个扫描过程的透明度（图 4 展示了一个多轮问答示例。）

1) 大语言模型：Vicuna v1.5 7B [26] 被用作骨干大语言模型。在此，Vicuna 以视觉编码器输出的视觉嵌入、由文本编码器编码的系统提示与查询提示，以及检索增强的历史上下文作为输入。它产生三类输出：当前扫描阶段 stage（阶段定义见第三节-*A*）、对该阶段的解释 explanation，以及对下一个 API 的预测。

2）超声图像编码器：RAG-RUSS 采用冻结的 PubMedCLIP- ViT-B/32 [27]作为视觉编码器。该编码器在大规模生物医学多模态数据集 ROCO [28], 上进行了预训练，使其能够有效捕捉超声、X 射线和磁共振成像等医学成像模态中的纹理、边界和结构特征。输入由两帧超声帧 [组成：在时间步 t采集的最新超声图像 $\mathbf{I}_t \in \mathbb{R}^{224\times224}$，以及在时间步 $t-\Delta$采集的前一帧超声图像 $\mathbf{I}_{t-\Delta}$，如图 3 (a)]所示。PubMedCLIP 将每幅图像编码为一系列图像块级词元，其中，来自倒数第二层的 [CLS] 词元 被输入 RAG 模块，而来自同一层的其余视觉词元 则传递给大语言模型。



(a) Architecture of RAG-RUSS

(b) RAG-RUSS Runtime inference for carotid US scan

图 3. 所提出的 RAG-RUSS 用于颈动脉超声扫描的架构与运行时推理。左侧： 系统架构由大语言模型骨干（Vicuna v1.5，7B [26]）、医学视觉基础模型（PubMedCLIP-ViT-B/32 [27]）以及检索相似扫描上下文以支持决策的 RAG 模块组成。右侧： RAG-RUSS 的输入/输出，以及将其部署于颈动脉超声扫描时的信号流。关于 RAG-RUSS 输入/输出和架构的更多细节，分别参见图 *4* 和*III-B*节。

System Prompt

A chat between a user and an artificial intelligence assistant designed for autonomous ultrasound scanning. RAG-RUSS provides helpful, detailed, and polite answers to the user's questions, including scanning stage, explanation and prediction of next API. The executable API are only transverse_tracking, transverse_back, rotate_to_longitudinal, none. The standardized scan stages for the carotid artery ultrasound are as follows: examine CCA proximal, examine CCA distal, examine bifurcation, Transverse scan completed, return to carotid bulb, return completed, rotate to longitudinal view, longitudinal scan completed.

Context Example 1

Human: This video records carotid artery ultrasound examination: <video> . Previous stage is examine CCA proximal. What is the current scanning stage?
RAG-RUSS: Examine CCA proximal
Human: What is the explanation at the current stage?
RAG-RUSS : Thyroid is near the CCA.
Human: Predict the next action API.
RAG-RUSS : transverse_tracking

Context Example 2 in the same format

Current Query

Human: This video records carotid artery ultrasound examination: <video> . Previous stage is examine CCA proximal. What is the current scanning stage?
RAG-RUSS : Examine CCA proximal
Human: What is the explanation at the current stage?
RAG-RUSS : Thyroid is near the CCA.
Human: Predict the next action API.
RAG-RUSS : transverse_tracking

图 4. RAG-RUSS 的多轮问答示例。输入包括：i). 系统提示，其中指定了任务描述以及预定义的可执行API和扫描阶段。ii). 通过 RAG组件检索到的两个扫描上下文。iii). 当前查询涉及两张输入超声图像和先前预测的阶段，并依次提出三个问题。随后，RAG-RUSS 输出：i). 当前阶段，ii). 简短解释，以及 iii). 下一步要执行的 API。

3）跨模态投影器：遵循 LLaVA [26]的设计，我们采用两层多层感知机作为跨模态投影器，以对齐视觉和文本特征。每个线性层之后均应用非线性激活函数 $\sigma(\cdot)$（本文中使用 GELU）。形式上，给定视觉嵌入 $z \in \mathbb{R}^{98\times768}$，投影定义如下：

$$z_v = \sigma\big(W_2 \cdot \sigma(W_1 \cdot z)\big), \tag{1}$$

其中， W_1 和 W_2 是可学习的权重矩阵。因此，输出 $z_v \in \mathbb{R}^{98\times4096}$ 与大语言模型的 token 嵌入空间对齐。该投影器在预训练和微调阶段均进行联合训练。

4）基于*RAG*的上下文学习： 为了检索历史扫描上下文（包括两帧超声帧、前一阶段和视觉问答标注；见图 4 中的上下文示例），我们首先训练一个残差多层感知机（ResMLP），将上下文信息投影到相同的嵌入空间。如图5所示，输入由从两张超声图像中提取的两个［CLS］标记（每个都是一个 768×1 向量）以及前一阶段的文本嵌入组成。前一阶段的文本使用表II映射为 16 维嵌入。将这三个嵌入拼接成一个 1552 维向量，再输入 ResMLP，将其投影到 256 维嵌入空间。采用式（2）中的三元组损失训练 ResMLP。

$$\mathcal{L}_{\text{tri}}(\mathbf{a},\mathbf{p},\mathbf{n}) = \max\left(\|\mathbf{a}-\mathbf{p}\|_2 - \|\mathbf{a}-\mathbf{n}\|_2 + \beta, 0\right) \tag{2}$$

其中，$\mathbf{a}$、$\mathbf{p}$、$\mathbf{n}$ 和 $\beta = 0.75$ 分别表示锚点、正样本、负样本和间隔距离。

检索过程中，将生成的嵌入与历史扫描上下文数据库进行比较，以计算余弦相似度分数。系统检索出最相似的前-k 个扫描上下文（本研究中为 $k = 2$ 个），并将其作为额外上下文提供给大语言模型。在查询时，这一检索机制使模型能够参考临床相似场景，从而提高阶段识别准确率和整体扫描成功率。

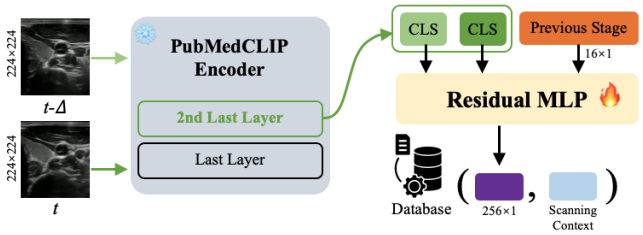


图5. 构建 RAG 数据库的方案。扫描上下文包括两个超声帧、前一阶段以及相关的视觉问答标注。

表II 文本阶段到嵌入向量的映射。

阶段	16 维嵌入向量
检查 CCA 近端	[0.1, 0.1, 0.1, ···, 0.1]
检查 CCA 远端	[0.2, 0.2, 0.2, ···, 0.2]
检查分叉处	[0.4, 0.4, 0.4, ···, 0.4]
横断面检查完成	[0.5、0.5、0.5、 ···、0.5]
返回颈动脉球部	[0.6, 0.6, 0.6, ···, 0.6]
返回完成	[0.7, 0.7, 0.7, ···, 0.7]
旋转至纵向视图	[0.8, 0.8, 0.8, ···, 0.8]
纵向扫描完成	[0.9, 0.9, 0.9, ···, 0.9]

B. 训练流程

RAG-RUSS 的训练可分为三个步骤，分别对应图 3 (a) 中指定的三个可训练模块。我们首先使用式 (2) 中定义的三元组损失，通过度量学习训练检索引擎的 ResMLP。训练期间，正样本对由扫描阶段相同但来自不同体积的样本组成，以促进泛化；负样本对则来自同一体积中的不同阶段，或来自不同体积中的其他阶段。

然后，遵循 LLaVA 范式 [29] [30]，，其余模块通过以下两个连续步骤进行训练：*i).* 预训练跨模态投影器以实现对齐：由冻结的 PubMedCLIP 编码器提取的视觉特征通过两层多层感知机与文本描述进行对齐。*ii).* 指令微调：使用来自视觉问答数据集的多轮问答对，通过 LoRA 适配器 [25] 对跨模态多层感知机和 Vicuna 主干网络进行微调。两个阶段均采用相同的下一词元预测交叉熵损失进行优化：

$$\mathcal{L}_{CE} = -\sum_{i=n+1}^{L} y_l \log P(x_l \mid z_{1:n}). \tag{3}$$

IV. 实验

为评估 RAG-RUSS 的性能，我们首先评估 RAG组件的检索准确率。随后开展消融研究，以确定视觉语言模型与 RAG组件各自的贡献，并评估检索示例数量变化所带来的影响。

A. 实现细节

所有实验均在配备 NVIDIA A100 GPU（80 GB 显存）的服务器上进行。训练使用 AdamW 优化器分 3 个连续阶段完成。*i) ResMLP* 模块：构成 RAG组件的 ResMLP 训练了 100 个 epoch，训练批大小为 32，验证批大小为 64，学习率为 3×10^{-6}，权重衰减为 1×10^{-5}，并采用带 10 % 预热的余弦衰减调度。*ii) VLM* 预训练： 视觉语言模型中的跨模态投影器在训练数据集上预训练了 5 个 epoch，不进行验证，批大小为 32，学习率为 1×10^{-5}，采用带 3 % 预热的余弦衰减调度，且不使用权重衰减。*iii) ICL* 指令微调： 大语言模型的多层感知机和 LoRA适配器在 ICL 数据集上训练，设置与 VLM 预训练相同。

B. RAG 检索准确率

我们使用 Top@k 指标评估 RAG 模块的性能，其定义如下：对于每个查询，如果前 k 个检索结果中至少有一个相关上下文示例，则 Top@k 记为 1；否则记为 0。最终得分为所有查询的平均值。表 IV 总结了在数据集 A 测试集上的检索准确率。平均而言，检索器的 Top@1 为 70.0%，Top@2 为 76.2%；增加一个上下文示例后，得分提升了 6.2%。具有明显且独特视觉线索的阶段表现最佳，例如检查 *CCA* 近端 （Top@1 为 95.3%）和旋转至纵向视图 （Top@1 为 94.2%），这表明检索器能够可靠地识别典型外观。相比之下，横向扫描完成 是最难的阶段，准确率最低（Top@2 为 37.3%）。虽然“返回”阶段仍然具有挑战性（返回颈动脉球部：Top@2 为 59.7%；返回完成：Top@2 为 76.9%），但这些阶段的 Top@2 提升幅度最大（分别为 11.7% 和 9.9%），表明其存在歧义，而多条参考示例有助于应对这种歧义。此外，检查分叉处的表现处于中等水平（Top@1 为 66.7%，Top@2 为 74.9%），这可能反映了不同病例之间较高的解剖变异性。这些结果揭示了两个关键发现：（i）在检索中使用 k=2 能够提升不确定情况下的鲁棒性；（ii）视觉上相似的阶段，尤其是返回和完成阶段，需要更具区分性的训练，例如难负样本挖掘，以增强阶段级区分能力。

表III 上下文示例检索准确率

阶段	Top@1 准确率 ↑	Top@2 准确率↑
检查 CCA 近端	95.3%	96.2%
检查 CCA 远端	85.5%	89.3%
检查分叉处	66.7%	74.9%
横向扫描完成	30.5%	37.3%
返回颈动脉球部	48.0%	59.7%
返回完成	67.0%	76.9%
旋转至纵向视图	94.2%	95.5%
纵向扫描完成	72.9%	80.0%
平均值	70.0%	76.2%

表 IV 不同消融设置下的阶段级识别准确率。

方法	颈总动脉近端	颈总动脉远端	分叉	横向完成	返回完成	纵向完成	平均准确率
仅RAG	96.5%	82.3%	**84.3%**	50.0%	25.0%	50.0%	64.7%
仅VLM	92.6%	93.3%	69.2%	75.0%	50.0%	50.0%	71.7%
RAG-RUSS@1	69.9%	92.1%	53.6%	50.0%	**75.0%**	**75.0%**	69.3%
RAG-RUSS@2	**97.8%**	**95.0%**	81.7%	**100.0%**	25.0%	**75.0%**	**79.1%**

RAG-RUSS@1 和 RAG-RUSS@2分别表示 RAG组件为 RAG-RUSS输入检索一个或两个上下文学习示例的配置。Trans.completed：横向扫描已完成。Long.completed：纵向扫描已完成。Avg.acc.：平均准确率。

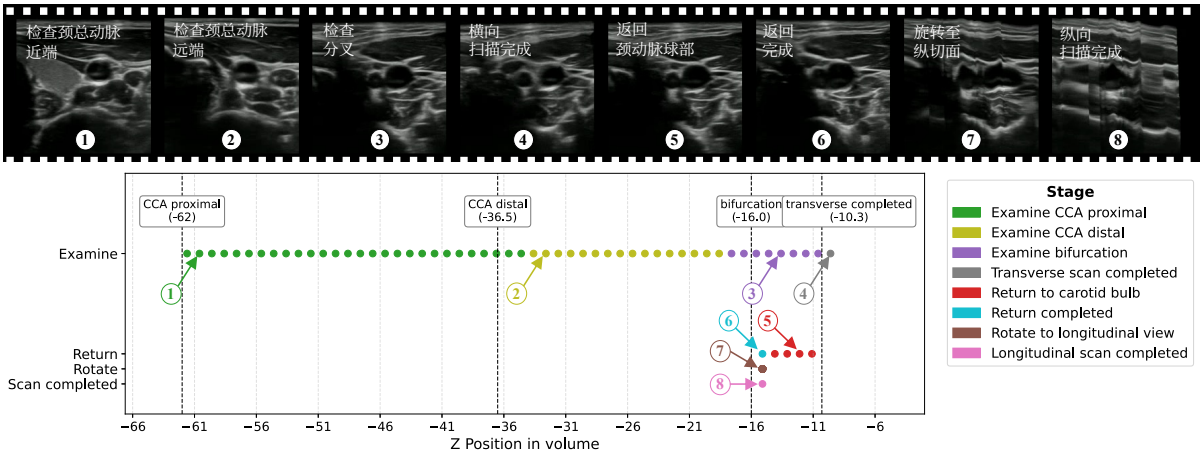


图6. RAG-RUSS@2 闭环评估结果的可视化示例。上方面板展示了模型预测的八个阶段对应的超声图像。下方面板展示了完整的扫描轨迹及标注的扫描区域，可视化呈现模型对不同阶段进展的识别结果。

C. 消融研究

为探究 RAG-RUSS 中各组件的作用以及检索示例数量的影响，我们进行了以下消融对比实验：i). 仅*RAG*：仅使用 RAG组件检索最相似的扫描上下文，用于阶段预测。ii).仅*VLM*：该基线仅依靠 VLM 进行预测，不进行检索增强。iii).*RAG-RUSS@1*和 iv). *RAG-RUSS@2*：这两种配置均采用完整框架，由 RAG组件分别检索一条或两条上下文学习示例，作为 RAG-RUSS 的输入。

因此，我们进行了闭环评估，以评估测试集中四个未见颈动脉体积数据在四种实验设置下连续扫描过程中的阶段预测准确率。各阶段准确率结果见表 IV。对于阶段检查 *CCA* 近端、检查 *CCA* 远端和检查分叉处，准确率定义为正确识别的推理时长与标注真实时长之比。对于横向扫描完成，准确率根据模型是否成功到达或超过所标注的航点来确定。对于 返回完成，准确率定义为预测阶段出现在分叉处之间的标注区域内。最后，对于纵向扫描完成，准确率根据颈动脉的纵向横截面是否被清晰可视化来评估。关于在未见过的使用 RAG-RUSS@2 的人体容积数据如图6所示。

结果汇总于表 IV。在所有配置中，仅RAG表现最差，平均准确率为64.7%。相比之下，RAG-RUSS@2（使用两个检索示例）的平均准确率最高（79.1%），在解剖结构定义明确的阶段表现出色，例如颈总动脉近端 （97.8%）、颈总动脉远端 （95.0%）和横向扫描完成 （100.0%）。这些结果表明，将视觉语言模型与多个多样化的检索结果结合起来，对于识别清晰的解剖标志和判定阶段完成事件尤其有效。在基线配置中，仅VLM的整体表现最稳定，平均准确率为71.7%。它在颈总动脉远端 （93.3%）和横向扫描完成 （75.0%）等阶段表现可靠，但缺少检索所提供的针对具体病例的上下文信息。

D. 讨论

实验结果表明，RAG-RUSS@2 的整体表现最佳，凸显了通过检索上下文增强视觉语言模型以进行超声扫描的有效性及优势。然而，仔细考察后发现，在所有配置中，返回完成 都是最具挑战性的阶段。有趣的是，RAG-RUSS@1（检索一个示例）在该阶段的准确率最高（75.0%），而 RAG-RUSS@2 的表现则骤降至 25.0%，这表明检索引入的噪声或相互冲突的上下文信号可能会

妨碍模型性能。这表明上下文并非越多越好，尤其是在存在时间歧义的情况下。目前仍存在若干局限。RAG-RUSS 需要大量计算资源，因为从多个检索示例中获取的较长输入会增加推理时间。此外，当前系统采用基于 API 的离散控制，而非连续动作执行，这可能会限制其在精细扫描任务中的适应性。此外，该方法仅基于预先录制的人体数据进行了验证。在真实场景中，形变等实际挑战将给未来的部署带来更多困难。尽管如此，RAG-RUSS 表明，将检索增强生成与视觉语言推理相结合，能够实现准确、可解释且自主的超声扫描。这是提升超声医师和患者信任与接受度的重要一步。

V. 结论

本文提出了 RAG-RUSS，这是一种可解释的自主颈动脉超声扫描框架，遵循既定的临床工作流程并利用视觉语言模型。为促进在结构化临床工作流程中的学习，我们基于 32 组人体容积扫描构建了一个高质量数据集。在此数据集的基础上，我们提出了 RAG-RUSS，该框架整合了大语言模型、视觉基础模型和检索引擎。尤其是，检索增强的上下文学习策略提高了适应性和泛化能力，同时减少了所需的训练数据量。实验结果表明，RAG-RUSS 能够按照临床方案自主完成全阶段颈动脉扫描。此外，通过解释当前图像和后续动作，该系统朝着开发值得信赖且可被接受、能够常规部署的 RUSS 迈出了重要一步。

参考文献

[1] Z. Izadifar、P. Babyn 和 D. Chapman，“超声的机械与生物学效应：现有知识综述”，《医学与生物学中的超声》，第 43 卷，第 6 期，第 1085–1104 页，2017年。[2] S. Sippel、K. Muruganandan、A. Levine 和 S. Shah，“超声在发展中世界的应用”，《国际急诊医学杂志》，第 4 卷，第 1 期，第 72 页，2011年。[3] Z. Jiang、S. E. Salcudean 和 N. Navab，“机器人超声成像：最新进展与未来展望”，《医学图像分析》，第 89 卷，第 102878 页，2023年。[4] M. K. Welleweerd、A. G. de Groot、V. Groenhuis、F. J. Siepel 和 S. Stramigioli，“自主机器人乳腺超声采集的平面外校正”，载于*2021 IEEE* 国际机器人与自动化会议 *(ICRA)*。IEEE，2021年，第 12 515–12 521 页。[5] X. Ma、Z. Zhang 和 H. K. Zhang，“机器人肺部超声成像的自主扫描目标定位”，载于*2021 IEEE/RSJ* 国际智能机器人与系统会议 *(IROS)*。IEEE，2021年，第 9467–9474 页。[6] D. Huang、Y. Bi、N. Navab 和 Z. Jiang，“机器人超声检查中的运动放大：实现脉动感知动脉分割”，载于*2023 IEEE/RSJ* 国际智能机器人与系统会议 *(IROS)*。IEEE，2023 年，第 6565–6570 页。[7] D. Huang 、 C. Yang、M. Zhou、A. Karlas、N. Navab 和 Z. Jiang，“使用虚拟夹具的机器人辅助深静脉血栓超声检查”，《*IEEE* 自动化科学与工程汇刊》，第 22 卷，第 381–392 页，2024。

[8] Z. Wang, Y. Han, B. Zhao, H. Xie, L. Yao, B. Li, M. Q.-H. Meng 和 Y. Hu，《采用视觉伺服导航进行颈动脉超声扫描的自主机器人系统》，IEEE医学机器人与仿生学汇刊，2024年。[9] Q. Huang、B. Gao 和 M. Wang，《用于颈动脉的机器人辅助自主超声成像》，IEEE仪器与测量汇刊，第73卷，第1–9页，2024年。[10] Y. Bi、Z. Jiang、F. Duelmer、D. Huang 和 N. Navab，《机器人超声成像中的机器学习：挑战与展望》，《年度控制、机器人与自主系统评论》，第7卷，2024年。[11] Y. Huang、W. Xiao、C. Wang、H. Liu、R. Huang 和 Z. Sun，《迈向基于临床方案模仿学习的全自主超声扫描机器人》，IEEE机器人与自动化快报，第6卷，第2期，第3671–3678页，2021年。[12] Z. Jiang、Y. Bi、M. Zhou、Y. Hu、M. Burke 和 N. Navab，《智能机器人超声医师：基于互信息的解耦奖励学习，仅需少量演示》，国际机器人研究期刊，第43卷，第7期，第981–1002页，2024。[13] K. Su、J. Liu、X. Ren、Y. Huo、G. Du、W. Zhao、X. Wang、B. Liang、D. Li 和 P. X. Liu，《用于甲状腺扫描的全自主机器人超声系统》，《自然·通讯》，第15卷，第1期，第4004页，2024。[14] K. Li、J. Wang、Y. Xu、H. Qin、D. Liu、L. Liu 和 M. Q.-H. Meng，《利用深度强化学习使超声探头自主导航至标准扫描平面》，载于 ICRA。IEEE，2021年，第8302–8308页。[15] T. Song、F. Li、Y. Bi、A. Karlas、A. Yousefi、D. Branzan、Z. Jiang、U. Eck 和 N. Navab，《智能虚拟超声医师（IVS）：促进医生-机器人-患者通信》，arXiv 预印本 arXiv:2507.13052，2025。[16] J. W. Kim、J.-T. Chen、P. Hansen、L. X. Shi、A. Goldenberg、S. Schmidgall、P. M. Scheikl、A. Deguet、B. M. White、D. R. Tsai 等，《SRT-H：通过语言条件模仿学习实现自主手术的分层框架》，《科学·机器人学》，第10卷，第104期，文章编号 eadt5254，2025年。[17] N. Roy、I. Posner、T. Barfoot、P. Beaudoin、Y. Bengio、J. Bohg、O. Brock、I. Depatie、D. Fox、D. Koditschek 等，《从机器学习到机器人学：具身智能的挑战与机遇》，arXiv 预印本 arXiv:2110.15245，2021年。[18] Y. Liu、W. Chen、Y. Bai、X. Liang、G. Li、W. Gao 和 L. Lin，《对齐网络空间与物理世界：具身人工智能综合综述》，IEEE/ASME机电一体化汇刊，2025。[19] D. Driess、F. Xia、M. S. Sajjadi、C. Lynch、A. Chowdhery、A. Wahid、J. Tompson、Q. Vuong、T. Yu、W. Huang 等，《PaLM-E：具身多模态语言模型》，2023年。[20] C. K. Ng、L. Bai、G. Wang、Y. Wang、H. Gao、K. Yuan、C. Jin、T. Zeng 和 H. Ren，《EndoVLA：用于内镜检查自主跟踪的双阶段视觉-语言-动作模型》，arXiv 预印本 arXiv:2505 15206，2025。。[21] M. Chen、S. Fan、G. Cao、Y.-h. Liu 和 H. Liu，《USPilot：由大语言模型增强图规划器的具身机器人超声助理系统》，arXiv 预印本 arXiv:2502 12498，2025。。[22] H. Jiang、A. Zhao、Q. Yang、X. Yan、T. Wang、Y. Wang、N. Jia、J. Wang、G. Wu、Y. Yue 等，《借助大规模学习型机器人系统迈向专家级自主颈动脉超声检查》，《自然·通讯》，第16卷，第1期，第7893页，2025。[23] J. Yuan、S. Sun、D. Omeiza、B. Zhao、P. Newman、L. Kunze 和 M. Gadd，《RAG-Driver：利用检索增强的上下文多模态大语言模型学习生成可泛化的驾驶解释》，载于《机器人学：科学与系统》，2024年。[24] A. Lasso、T. Heffter、A. Rankin、C. Pinter、T. Ungi 和 G. Fichtinger，《PLUS：用于超声引导介入系统的开源工具包》，IEEE生物医学工程汇刊，第61卷，第10期，第2527–2537页，2014年。[25] E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang、W. Chen 等，《LoRA：大语言模型的低秩自适应》。ICLR，第1卷，第2期，第3页，2022年。[26] L. Zheng、W.-L. Chiang、Y. Sheng、S. Zhuang、Z. Wu、Y. Zhuang、Z. Lin、Z. Li、D. Li、E. Xing 等，《使用 MT-Bench 和 Chatbot Arena 评估以大语言模型为评判者的评判能力》，神经信息处理系统进展，第36卷，第46 595–46 623页，2023年。[27] S. Eslami、G. De Melo 和 C. Meinel，《CLIP 在医学领域的视觉问答任务中是否能像在通用领域一样发挥作用？》，arXiv 预印本 arXiv:2112.13906，2021年。

[28] J. R¨uckert、L. Bloch、R. Br¨ungel、A. Idrissi-Yaghir、H. Sch¨afer、C. S. Schmidt、S. Koitka、O. Pelka、A. B. Abacha、A. G. Seco de Herrera 等，《Rocov2：放射学情境对象第2版，一种更新的多模态图像数据集》，*Scientific Data*，第11卷，688，2024年1，第。[29] H. Liu、C. Li、Q. Wu 和 Y. J. Lee，《视觉指令微调》，神经信息处理系统进展，第36卷，第34 892–34 916页，2023年。[30] F. Li、R. Zhang、H. Zhang、Y. Zhang、B. Li、W. Li、Z. Ma 和 C. Li，《Llava-next-interleave：应对大型多模态模型中的多图像、视频和三维任务》，*arXiv* 预印本 *arXiv:2407.07895*，2024。