2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

RAG-RUSS: A Retrieval-Augmented Robotic Ultrasound for Autonomous Carotid Examination

Dianye Huang[1,2,*], Ziping Cong[1,*], Nassir Navab[1,2], *Fellow, IEEE*, and Zhongliang Jiang[1,2,3,†]

***Abstract*— Robotic ultrasound (US) has recently attracted increasing attention as a means to overcome the limitations of conventional US examinations, such as the strong operator dependence. However, the decision-making process of existing methods is often either rule-based or relies on end-to-end learning models that operate as black boxes. This has been seen as a main limit for clinical acceptance and raises safety concerns for widespread adoption in routine practice. To tackle this challenge, we introduce the RAG-RUSS, an interpretable framework capable of performing a full carotid examination in accordance with the clinical workflow while explicitly explaining both the current stage and the next planned action. Furthermore, given the scarcity of medical data, we incorporate retrieval-augmented generation to enhance generalization and reduce dependence on large-scale training datasets. The method was trained on data acquired from 28 volunteers, while an additional four volumetric scans recorded from previously unseen volunteers were reserved for testing. The results demonstrate that the method can explain the current scanning stage and autonomously plan probe motions to complete the carotid examination, encompassing both transverse and longitudinal planes. Code: `https://github.com/congzp/USrobot`**

I. Introduction

Ultrasound (US) is a non-invasive, real-time imaging modality with wide accessibility, establishing it as the primary diagnostic tool in contemporary clinical practice [1]. However, the quality of US imaging and the efficiency of acquiring standard imaging planes are highly dependent on the operator's proficiency in probe manipulation. This reliance on operator expertise introduces variations in examination outcomes and requires years of training to attain proficiency, which exacerbates workforce shortages, especially in under-resourced regions [2].

In order to reduce physician workloads and examination variability, over the past two decades, the autonomous robotic ultrasound scanning system (RUSS) has been extensively researched and is receiving ongoing attention [3]. Various RUSS have been developed for a broad range of clinical applications, including autonomous screening of the breast [4], lung [5], and blood vessels [6], [7]. These systems mainly focus on automating the screening process by leveraging US imaging feedback and/or external RGB-D cameras. Wang *et al.* [8] proposed a vision-servo-based autonomous carotid US scanning system using an improved Siamese network to track the target vessel. Huang *et al.* [9] proposed a two-stage process for automatic switching between transverse and longitudinal sections, combining impedance/force control with image segmentation. However, these systems mainly focus on motion control and rely heavily on rule-based policies, which struggle to exhibit higher-level understanding and dynamic decision-making capabilities.

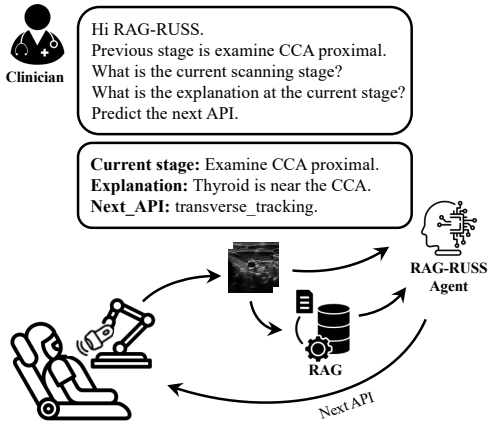


Fig. 1. An illustration showing a representative use case of the RAG-RUSS system. Given a clinician's query, RAG-RUSS retrieves similar annotated scans via RAG, then identifies the `stage`, generates an `explanation`, and predicts the `next_API`. This API is performed by the robotic arm, and the acquired US images are fed back into the system, forming a closed loop that enables autonomous, interpretable carotid artery scanning.

To address these limitations, learning-based RUSS have emerged as a promising direction, enabling more intelligent, data-driven perception and control [10]. In order to reduce the reliance on handcrafted rules, Huang *et al.* [11], [12] introduce clinical imitation learning (IL) frameworks to replicate the operational habits of experienced physicians. Given the current observation, reinforcement learning (RL) is also employed to infer the next control commands to maneuver the US probe toward the standard imaging plane [13]–[15]. To address inter-operator variability in US imaging, Jiang *et al.* propose an intelligent robotic sonographer that autonomously navigates to standard imaging planes by learning expert-level scanning strategies. In their work, a neural reward function, trained via ranked pairwise image comparisons in a self-supervised manner, captures high-level physiological knowledge. Despite the promising results achieved by the aforementioned approaches, the practical deployment of RUSS remains hindered by the inherent opacity of IL and RL paradigms. These data-driven methods typically encode expertise through only image–trajectory demonstration pairs, yielding black-box policies that mirror the opacity of human decision-making. Transparent system designs that allow physicians to monitor the internal states

* Contribute Equally. † Corresponding author(e-mail: zl.jiang@tum.de).
[1] Computer-Aided Medical Procedures and Augmented Reality (CAMP), Technical University of Munich (TUM), Germany,
[2] Munich Center for Machine Learning (MCML), Germany
[3] The University of Hong Kong, Hong Kong, China

of trained policies have therefore become essential [16]. Building upon this need for policy transparency, language emerges as a natural bridge connecting the medical robots, physicians, and patients in real-world clinical settings. It is worth noting that achieving such embodied communicative capability necessitates the integration of low-level perceptual inputs with high-level semantic reasoning, an area in which traditional rule-based and IL/RL approaches remain limited [17], [18].

Recently, large language models (LLMs) have shown strong associative capabilities and broad commonsense knowledge, accelerating progress in embodied artificial intelligence [19], [20]. Motivated by this progress, the research community starts exploring the use of vision-language models (VLMs) and vision-language-action (VLA) frameworks to enable intelligent, interactive agents. Wang *et al.* propose EndoChat, a specialized multimodal large language model (MLLM) designed for diverse dialogue tasks in robotic-assisted surgery, trained on the Surg-396K dataset. Ng *et al.* [21] introduce EndoVLA, a VLA model tailored for continuum robots in gastrointestinal interventions, capable of semantic polyp tracking, delineating abnormal mucosal regions, and following circular cutting markers. In the field of US, Xu *et al.* introduced USPilot [22], an embodied, LLM-powered robotic US assistant that acts as a virtual sonographer—answering patient queries, interpreting US-specific tasks through fine-tuning, and invoking APIs based on user intent. The system integrates a fine-tuned LLM with an LLM-enhanced graph neural network for API control and task planning. However, its primary focus lies in intent understanding and high-level planning, with limited capability for real-time image content analysis, constraining its effectiveness in dynamic task execution. Alternatively, Jiang *et al.* [23] proposed UltraBot, a carotid artery ultrasonography system trained on a large-scale dataset comprising 247 k image–action expert demonstrations. UltraBot employs end-to-end imitation learning to directly map real-time US images to 6-DoF probe motions, while integrating scanning, biometric measurement, plaque segmentation, and report generation into a unified workflow. However, UltraBot's decision-making process remains a black box without explicit modeling of intermediate reasoning steps or stage-wise task decomposition. While large model-based methods have shown promising results, they require extensive domain-specific training datasets, which are often prohibitively expensive to collect. To mitigate this limitation, and inspired by RAG-Driver [24] from the autonomous driving domain, we propose incorporating a retrieval-augmented generation (RAG) component to reduce dependency on large training datasets and alleviate data scarcity bottlenecks in the medical field.

In this work, we introduce RAG-RUSS, a retrieval-augmented framework for carotid US scanning that couples a VLM with an RAG component to reduce dependence on large training corpora. Unlike systems such as USPilot and UltraBot, which emphasize intent understanding or direct end-to-end action mapping, our approach leverages retrieved, similar scan contexts to enable image understanding and explanation-driven decision-making during closed-loop execution. To train RAG-RUSS, we also constructed a high-quality dataset from 32 human volumetric scans. Within a clinically defined workflow, RAG-RUSS predicts the current stage and provides stepwise explanations alongside the next_API. This design supports standardized carotid scanning by adapting evidence from prior cases while maintaining transparency throughout the examination. To the best of our knowledge, this is the first robotic ultrasound system to integrate perception, explanation, and action in a unified framework for carotid artery examination. By explaining both the current image and the subsequent action, RAG-RUSS takes an important step toward trustworthy, clinically acceptable deployment.

The rest of this paper is organized as follows: *Section II* outlines the clinical workflow for carotid artery scan and the data preparation process. *Section III* presents the architecture of the proposed RAG-RUSS. *Section IV* describes the experimental setup and reports the results. Finally, *Section V* concludes the study.

II. Examination Workflow and Data Preparation

A. Carotid Ultrasound Examination Workflow

1) Clinical Workflow for Carotid Artery Examination: The carotid arteries, located on either side of the trachea at the neck, are the primary vessels that deliver oxygen-rich blood from the heart to the brain and face. Each carotid artery comprises a common carotid artery (CCA) that bifurcates into the internal carotid artery (ICA), supplying blood to the brain and eyes, and an external carotid artery (ECA), supplying blood to the face and scalp (see Fig. 2). Atherosclerosis is one of the most prevalent carotid diseases, in which plaque accumulation leads to carotid stenosis (luminal narrowing) and elevates stroke risk, particularly for the aged population. In this context, medical US has become the standard first-line diagnostic tool due to its advantages of being radiation-free and providing real-time imaging.

For regular carotid examination, sonographers generally follow a standard workflow [9], [11]. To ensure full satisfaction of the clinical diagnostic purpose, sonographers need to fully cover the CCA and its bifurcation into the internal and external branches. The examination is typically performed first in the transverse (short-axis) plane along the lumen centerline and then in the longitudinal (long-axis) plane of the carotid artery [8]. The scan usually begins in transverse view at the proximal CCA, proceeds distally to the carotid bulb (carotid sinus), and continues across the bifurcation to visualize the ICA and ECA. The physician then switches to longitudinal views for detailed assessment and measurements, such as the vessel diameter.

2) Scanning Stage Definition: To automate the scanning procedure and facilitate the holistic understanding of the overall progress of the examination for RUSS, we adopt the above clinical workflow to an automatic system, following [9], [23]. In this study, we partition the carotid scan

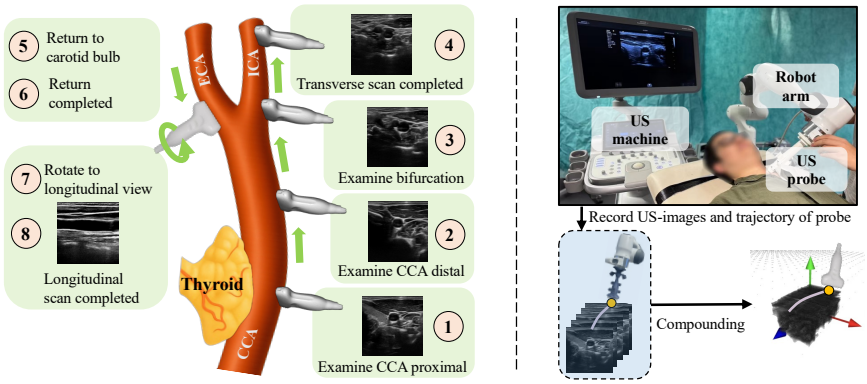


Fig. 2. Overview of the carotid ultrasound scanning workflow and data acquisition/processing pipeline. *Left:* presents eight predefined scanning stages of the carotid artery examination. *Right:* US images and the probe's pose trajectory are recorded and subsequently compounded into a 3D volumetric data for use in the simulation environment.

process into eight sequential stages (see Fig. 2), each guided by distinct anatomical cues from the US images.

① `Examine CCA proximal`: the CCA is visualized together with the thyroid gland.
② `Examine CCA distal`: the CCA remains visible, but the thyroid gland is absent from the images.
③ `Examine bifurcation`: the carotid bulb is reached; the lumen starts diverging into ECA and ICA.
④ `Transverse scan completed`: the lumen is clearly divided into the ICA and ECA.
⑤ `Return to carotid bulb`: as the probe is moved back toward the bifurcation, the ICA and ECA gradually converge into a single lumen.
⑥ `Return completed`: the ICA and ECA converge into the common carotid artery at the carotid bulb.
⑦ `Rotate to longitudinal view`: the artery cross-section changes from an elliptical shape to an elongated, rectangular profile.
⑧ `Longitudinal scan completed`: the longitudinal cross-section of carotid artery is clearly visualized.

B. Ultrasound Simulation Using Human Volumetric Data

Due to safety concerns, it is impractical to collect all the required training scanning trajectories directly from volunteers. Following previous studies [15], we built a simulation environment based on recorded human volumetric data, in which a virtual probe is initialized to perform continuous scanning. Each paired set of simulated US images and the corresponding probe trajectory is regarded as one demonstration of a complete carotid scan. Notably, a single pre-recorded volume can generate multiple demonstrations. To mitigate the risk of overfitting, a total of 32 three-dimensional US volumes of the carotid artery were acquired and reconstructed from scans of 32 healthy volunteers.

1) Human Volumetric Data Acquisition: The data acquisition platform shown in Fig. 2 was built on a Panda robotic arm (Franka Emika, Germany) integrated with an ACUSON Juniper ultrasound (US) system (Siemens, Germany). A linear US probe (12L3, Siemens, Germany) was mounted on the robot end-effector using a custom-designed, 3D-printed probe holder. The US images were captured via a frame grabber (USB Capture HDMI, MAGEWELL, China) at a frequency of 30 Hz. Meanwhile, the robotic tracking data were recorded through a ROS interface at a higher rate of approximately 100 Hz to compensate for potential temporal misalignment between the two data streams. By stacking the 2D US images with their corresponding robot poses in 3D space, a reconstructed 3D US volume can be generated using a classic interpolation technique such as PLUS [25]. An illustrative example of the reconstructed volume is shown in Fig. 2. A total of 32 carotid volumes were acquired from 32 healthy volunteers. Each volume was manually checked to ensure that the vascular morphology was clearly visible and that the bifurcation structures were fully captured. To show the variety in the group, the volunteers' demographic and physical information is summarized in Table I.

The compounded carotid volume is visualized using OpenGL. By initializing a virtual US probe, we can mimic the standard examination workflow discussed in the previous section to generate the carotid examination demonstration with paired probe pose and B-mode image. By adjusting the virtual probe position, a set of standardized carotid examination demonstrations can be obtained. Such a human volumetric data-based simulation preserves the fidelity of medical ultrasound images while providing a flexible and practical controllable scanning trajectory generation.

TABLE I
STATISTICS OF VOLUNTEERS (N=32).

Age (years)	Height (m)	Weight (kg)	BMI
27.97 ± 3.91	1.74 ± 0.09	70.22 ± 12.24	23.06 ± 2.86

2) Expert Carotid Examination Demonstration Generation: This section introduces the implementation details for generating expert demonstrations of carotid examination using the reconstructed human volumetric data. First, a set of anatomical waypoints corresponding to different scanning stages was manually defined based on visual vessel features. To standardize the scanning trajectory generation, we further compute the centroid point of the vessel in the labeled binary

vessel mask to have a refined scanning trajectory. Then, a scanning trajectory can be generated by connecting these refined waypoints and adjusting its orientation in specific stage transitions. By automatically executing this continuous scanning trajectory in the simulation, the paired B-mode image can be obtained for individual probe locations.

For each reconstructed volume, three complete scans were conducted from the proximal segment of the CCA to the longitudinal plane visualization, as illustrated in Fig. 2. To balance the sample distribution across different scanning stages, an additional 20 scans were acquired for individual stages with relatively fewer samples, including *Transverse scan completed*, *Return completed*, and *Longitudinal scan completed*. During the transverse scanning stage, the probe was advanced in increments of 1 mm per step. In the return stage, the first step retracted 2 mm, followed by subsequent steps of 1.5 mm each. To introduce variation among scans obtained from the same volunteer, random perturbations were added to the motion commands at each step. Specifically, translational deviations of up to ± 0.3 mm in the longitudinal direction and ± 0.2 mm in the lateral direction were applied.

C. Dataset Structure

To train the proposed RAG-RUSS with the ability to interact with human users, the dataset was designed to incorporate both visual and textual information. A sliding time window of N steps was employed. For each entry within the window, in addition to the US images, we defined three types of textual annotations: (1) the current scanning stage, such as Stage 1 *Examine CCA proximal* and Stage 2 *Examine CCA distal* (see Fig. 2); (2) an explanation of the key anatomical feature characterizing the stage; and (3) the name of the next API required to complete the given task. The stage-specific explanations were provided by our clinical collaborator according to standard criteria used to differentiate scanning stages during carotid examinations. For example, the explanations for Stages 1 and 2 are "Thyroid is near the CCA" and "Thyroid is not visible," respectively. In this study, three APIs were defined: "tracking forward," "tracking backward," and "rotation clockwise," which correspond to advancing the probe, retracting the probe, and rotating the probe from the transverse plane to the longitudinal plane, respectively. In total, we collected 15,459 items (N-length sliding window), inlcuding paired visual and textual information based on 32 reconstructed volumes. The images are in the resolution of 224×224 pixels. To avoid data leakage, four out of 32 volunteers' data is kept for testing.

For different training objectives, three sub-datasets were constructed: (1) dataset A for training the RAG module, (2) dataset B for pre-training the cross-modality projector (2-layer MLP) to align visual and textual features, and (3) dataset C for fine-tuning the cross-modality MLP together with the VLM backbone using LoRA adapters [26]. Dataset A consisted of the first and last images within each moving window, along with the textual stage description corresponding to the last frame. In total, 12,937 entries from 28 volunteers were included. For training, we iterated over each entry and generated 20 paired positive and negative samples to support contrastive learning. Dataset B was constructed from individual B-mode images paired with predefined stage-specific queries and corresponding answers that describe the anatomical features characterizing each stage. Dataset C was designed to refine the LLM backbone and includes multi-turn question–answer pairs concerning the current stage, the next API, and related instructions, thereby enabling effective interaction with human users.

III. Method

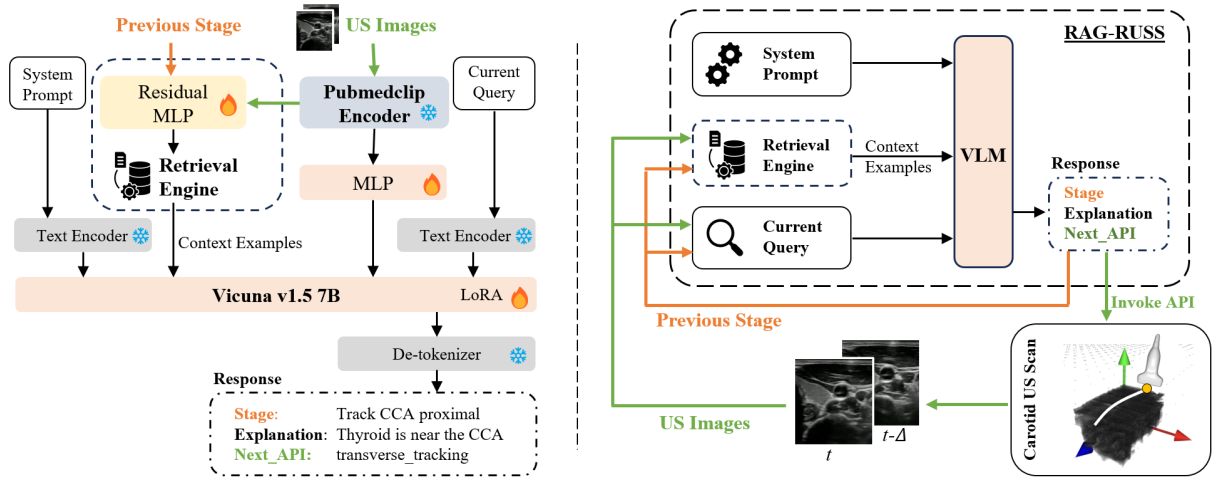
In this section, we elaborate on the structure and the training pipeline of RAG-RUSS, an intelligent sonography agent that explains its current scanning stage at each timestep, enabling the physician to monitor progress and intervene if the procedure deviates from the plan. Effective, natural communication between the agent and the physician is therefore essential. To this end, RAG-RUSS is built around an LLM to leverage its learned "common sense" and intrinsic capacity for natural-language communication, while seamlessly incorporating physician expertise.

A. Structure of RAG-RUSS

As illustrated in Fig. 3 (a), RAG-RUSS comprises an LLM, a vision foundation model trained on large-scale medical data, and a retrieval engine that extracts relevant knowledge to support decision-making. In Fig. 3 (b), during inference, RAG-RUSS follows a multi-turn question–answering workflow to address three queries: identifying the current `stage`, generating an `explanation`, and predicting the `next_API`. We also incorporate multimodal in-context learning (ICL), enabling the model to reference visual–linguistic history and thereby improve scanning accuracy and reliability. At each query step, RAG-RUSS produces interpretable text, ensuring transparency throughout the scan (see an illustrative multi-turn QA example in Fig. 4.)

1) Large Language Model: Vicuna v1.5 7B [27] is employed as the backbone LLM. Here, Vicuna takes as input the visual embeddings from a vision encoder, the system and query prompts encoded by the text encoder, and the retrieval-augmented historical context. It produces three types of outputs: the current scanning stage `stage` (see the stage definitions in *Section III-A*), an explanation of that stage `explanation`, and the prediction of the `next_API`.

2) Ultrasound Image Encoder: A frozen PubMedCLIP-ViT-B/32 [28] is adopted as the vision encoder for RAG-RUSS. This encoder is pretrained on a large-scale biomedical multimodal dataset ROCO [29], which enables it to effectively capture texture, boundary, and structural features in medical imaging modalities such as US, X-ray, and MRI. The input consists of two US frames [the most recent US image $\mathbf{I}_t \in \mathbb{R}^{224\times224}$ captured at timestep t, and the previous US image $\mathbf{I}_{t-\Delta}$ captured at timestep $t-\Delta$, as illustrated in Fig. 3 (a)]. Each image is encoded by PubMedCLIP into a sequence of patch-level tokens, where the `[CLS]` token *from the second-last layer* is fed into the RAG module, while the remaining visual tokens *from the same layer* are passed to the LLM.



(a) Architecture of RAG-RUSS

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

Fig. 3. Architecture and runtime inference of the proposed RAG-RUSS for carotid artery US scanning. *Left:* System architecture comprising an LLM backbone (Vicuna v1.5, 7B [27]), the medical vision foundation model (PubMedCLIP-ViT-B/32 [28]), and a RAG module that retrieves similar scanning contexts to support decision-making. *Right:* Inputs/outputs of RAG-RUSS and the signal flow when deploying it for carotid US scanning. For more details on inputs/outputs and architecture of RAG-RUSS, refer to *Fig. 4* and *Section III-B*, respectively.

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

Fig. 4. An illustrative multi-turn QA example of RAG-RUSS. The inputs are: i). System prompt that specifies the task description along with the predefined executable APIs and scanning stages. ii). Two retrieved scanning contexts via the RAG component. iii). The current query involves the two input US images and the previously predicted stage, and then poses three questions sequentially. RAG-RUSS then outputs: i). the current `stage`, ii). a short `explanation`, and iii). the `next_API` to execute.

3) Cross-Modality Projector: Following the LLaVA [27] design, we employed a two-layer MLP as a cross-modality projector to align visual and textual features. The non-linear activation function $\sigma(\cdot)$ (GELU in this work) is applied after each linear layer. Formally, given the visual embedding $z \in \mathbb{R}^{98\times768}$, the projection is defined as:

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

where W_1 and W_2 are learnable weight matrices. The output $z_v \in \mathbb{R}^{98\times4096}$ is thus aligned with the LLM's token-embedding space. This projector is trained jointly during both the pretraining and fine-tuning.

4) RAG-based In-Context Learning: To perform the retrieval of historical scanning context, including two US frames, the previous stage, and VQA annotations (see the Context Example in Fig. 4), we first train a residual MLP (ResMLP) to project the context information into the same embedding space. As illustrated in Fig. 5, the input consists of two `[CLS]` tokens extracted from two US images (each a 768×1 vector) and the textual embedding of the previous stage. The textual previous stage is mapped into a 16-dimensional embedding using Table II. These three embeddings are concatenated into a 1552-dimensional vector, which is then passed through the ResMLP to project it into a 256-dimensional embedding space. A triplet loss in Eq. (2) is utilized to train the ResMLP.

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

where $\mathbf{a}$, $\mathbf{p}$, $\mathbf{n}$, and $\beta = 0.75$ denote the anchor, positive, negative samples, and marginal distance, respectively.

During the retrieval, the resulting embedding is compared against a database of historical scanning contexts to compute cosine similarity scores. The top-k most similar scanning contexts ($k = 2$ in this work) are retrieved and supplied to the LLM as additional context. This retrieval mechanism allows the model to reference clinically similar scenarios at query time, improving stage-recognition accuracy and overall scanning success.

B. Training Pipeline

The training of RAG-RUSS can be divided into three steps, corresponding to the three trainable modules specified in Fig. 3 (a). We first train the ResMLP of the Retrieval Engine using metric learning with the triplet loss defined

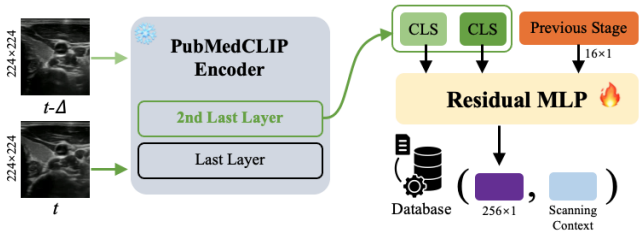


Fig. 5. Scheme for building an RAG database. The scanning context includes two US frames, the previous stage, and associated VQA annotations.

TABLE II
TEXTUAL STAGE TO EMBEDDING VECTOR MAPPING.

Stage	16-dim. Embedding Vector
Examine CCA proximal	[0.1, 0.1, 0.1, ···, 0.1]
Examine CCA distal	[0.2, 0.2, 0.2, ···, 0.2]
Examine bifurcation	[0.4, 0.4, 0.4, ···, 0.4]
Transverse completed	[0.5, 0.5, 0.5, ···, 0.5]
Return to carotid bulb	[0.6, 0.6, 0.6, ···, 0.6]
Return completed	[0.7, 0.7, 0.7, ···, 0.7]
Rotate to longitudinal view	[0.8, 0.8, 0.8, ···, 0.8]
Longitudinal scan completed	[0.9, 0.9, 0.9, ···, 0.9]

in Eq. (2). During training, the positive pairs comprised samples that shared the same scanning stage but were drawn from different volumes to encourage generalization, whereas negative pairs were drawn either from different stages within the same volume or from other stages in different volumes.

Then, following the LLaVA paradigm [30] [31], the remaining modules are trained with two sequential steps: *i). Pre-training the cross-modality projector for alignment:* Visual features extracted by the frozen PubMedCLIP encoder are aligned with textual descriptions through a two-layer MLP. *ii). Instruction fine-tuning:* Multi-turn QA pairs from the VQA dataset are used to fine-tune the cross-modality MLP and the Vicuna backbone with LoRA adapters [26]. Both stages are optimized with the same next-token prediction cross-entropy loss:

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

IV. EXPERIMENTS

To evaluate the performance of RAG-RUSS, we begin by evaluating the retrieval accuracy of the RAG component. Then, an ablation study is conducted to identify the individual contributions of the VLM and the RAG component, as well as to assess the impact of varying the number of retrieved examples.

A. Implementation Details

All experiments were conducted on a server with an NVIDIA A100 GPU (80 GB memory). Training was performed in 3 sequential stages using the AdamW optimizer. *i) ResMLP module:* The ResMLP constructing the RAG component was trained for 100 epochs with batch sizes of 32 for training and 64 for validation, a learning rate of 3×10^{-6}, weight decay of 1×10^{-5}, and cosine decay scheduling with 10% warm-up. *ii) VLM pretraining:* The cross-modality projector in VLM was pretrained for 5 epochs on the training dataset without validation, using a batch size of 32, a learning rate of 1×10^{-5}, cosine decay scheduling with 3% warm-up, and no weight decay. *iii) ICL instruction tuning:* The MLP and LoRA adapters of the LLM were trained on the ICL dataset using the same settings as VLM pretraining.

B. RAG Retrieval Accuracy

We evaluate the performance of RAG module using the Top@k metric, defined as follows: for each query, Top@k is set to 1 if at least one relevant context example appears among the top-k retrieved results; otherwise, it is 0. The final score is reported as the average over all queries. Table IV summarizes the retrieval accuracy on the test set of dataset A. On average, the retriever attains 70.0% Top@1 and 76.2% Top@2, with 6.2% gain from adding one extra context example. Stages with strong, distinctive visual cues perform best, such as *Examine CCA proximal* (95.3% Top@1) and *Rotate to longitudinal view* (94.2% Top@1), suggesting the retriever reliably identifies canonical appearances. In contrast, *Transverse scan completed* is the most difficult stage, yields the lowest accuracy (37.3% Top@2). While the "return" phases remain challenging (*Return to carotid bulb*: 59.7% Top@2; *Return completed*: 76.9% Top@2), the largest Top@2 gains are observed for these stages (11.7% and 9.9%, respectively), indicating ambiguity that benefits from multiple references. In addition, *Examine bifurcation* shows mid-level performance (66.7% Top@1, 74.9% Top@2), likely reflecting higher anatomical variability across cases. These results highlight two key observations: (i) using retrieval with k=2 improves robustness under uncertainty, and (ii) visually similar stages, especially during return and completion, require more discriminative training, such as harder negative mining, to enhance stage-level separation.

TABLE III
CONTEXT EXAMPLE RETRIEVAL ACCURACY

Stage	Top@1 Acc. ↑	Top@2 Acc.↑
Examine CCA proximal	95.3%	96.2%
Examine CCA distal	85.5%	89.3%
Examine bifurcation	66.7%	74.9%
Transverse scan completed	30.5%	37.3%
Return to carotid bulb	48.0%	59.7%
Return completed	67.0%	76.9%
Rotate to longitudinal view	94.2%	95.5%
Longitudinal scan completed	72.9%	80.0%
Average	70.0%	76.2%

C. Ablation Study

To investigate the role of each component in RAG-RUSS and the impact of the number of retrieval examples, we conducted the following ablation experiment comparison: i). *RAG Only:* Only RAG component was used to retrieve the most similar scanning context for stage prediction. ii). *VLM Only:* This baseline relied solely on the VLM for prediction without retrieval augmentation. iii). *RAG-RUSS@1*

TABLE IV

STAGE-LEVEL RECOGNITION ACCURACY UNDER DIFFERENT ABLATION SETTINGS.

Method	CCA proximal	CCA distal	Bifurcation	Trans. completed	Return completed	Long. completed	Avg acc.
RAG only	96.5%	82.3%	**84.3%**	50.0%	25.0%	50.0%	64.7%
VLM only	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 and RAG-RUSS@2 denote configurations in which the RAG component retrieves one or two in-context learning examples, respectively, for the RAG-RUSS input. Trans. completed: Transverse scan completed. Long. completed: Longitudinal scan complete. Avg. acc.: Average accuracy.

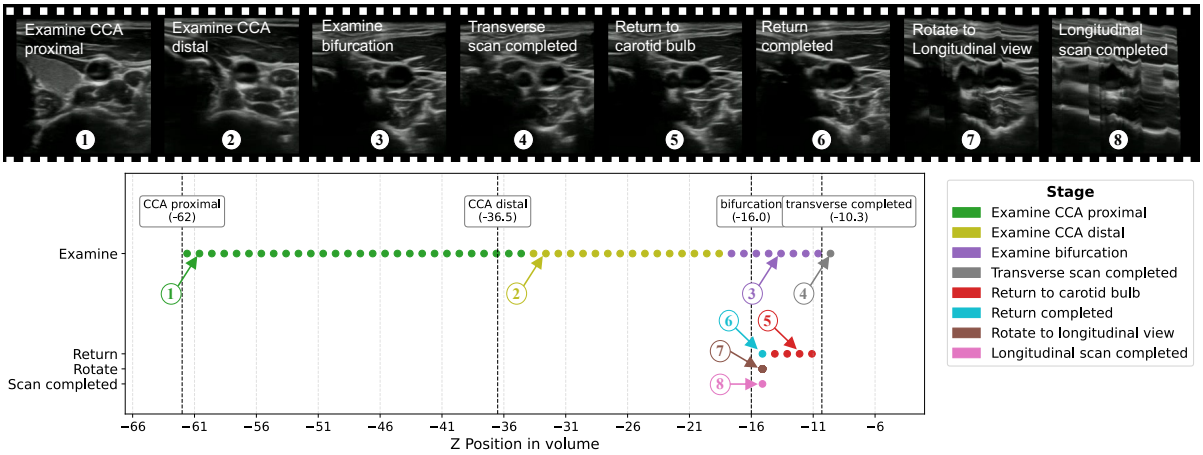


Fig. 6. Visualization example of the closed-loop evaluation results of RAG-RUSS@2. The upper panel presents US images corresponding to eight stages predicted by the model. The lower panel illustrates the complete scanning trajectory together with the annotated scanning regions, providing a visualization of the model's recognized progression through different stages.

and iv). *RAG-RUSS@2*: These configurations used the full framework, with the RAG component retrieving one or two in-context learning examples, respectively, as input to RAG-RUSS.

Accordingly, we conducted a closed-loop evaluation to assess stage prediction accuracy during continuous scanning across four experimental settings on four unseen carotid artery volumes in the test set. The stage-wise accuracy results are presented in Table IV. For the stages *Examine CCA proximal*, *Examine CCA distal*, and *Examine bifurcation*, accuracy is defined as the ratio between the duration of correctly recognized inferences and the annotated ground-truth duration. For *Transverse scan completed*, accuracy is determined by whether the model successfully reaches or surpasses the labeled waypoint. For *Return completed*, accuracy is defined as the predicted stage occurring within the annotated region between the bifurcations. Finally, for *Longitudinal scan completed*, accuracy is assessed based on whether the longitudinal cross-section of the carotid artery is clearly visualized. An intuitive visualization for automatically completing a full-loop carotid examination on unseen human volumetric data using RAG-RUSS@2 is shown in Fig. 6.

The results are summarized in Table IV. Among all configurations, RAG Only performs the worst with an average accuracy of 64.7%. In contrast, RAG-RUSS@2 (with two retrieved examples) achieves the highest average accuracy (79.1%), demonstrating strong performance on anatomically well-defined stages such as *CCA proximal* (97.8%), *CCA distal* (95.0%), and *Transverse scan completed* (100.0%). These results suggest that combining VLM with multiple, diverse retrievals is particularly effective for recognizing clear anatomical landmarks and identifying stage completion events. The VLM Only configuration yields the most stable overall performance among the baselines, achieving an average accuracy of 71.7%. It performs reliably on stages like *CCA distal* (93.3%) and *Transverse scan completed* (75.0%), but lacks the case-specific contextualization that retrieval provides.

D. Discussion

The experimental results demonstrate that RAG-RUSS@2 achieves the best overall performance, highlighting the effectiveness and benefits of augmenting VLMs with retrieval-based context for US scanning. However, a closer examination reveals that *Return completed* emerges as the most challenging stage across all configurations. Interestingly, while RAG-RUSS@1 (with one retrieved example) achieves the highest accuracy on this stage (75.0%), performance drops sharply to 25.0% in RAG-RUSS@2, suggesting that retrieval-induced noise or conflicting contextual signals can hinder model performance. This highlights that more context is not always better, especially when temporal ambiguity is involved. Several limitations remain. RAG-RUSS requires significant computational resources, as longer inputs from multiple retrieved examples increase inference time. Additionally, the current system operates with discrete API-based control, rather than continuous action execution, which may constrain its adaptability in fine-grained scanning tasks. Furthermore, the approach has only been validated based

on the prerecorded human data. Practical challenges in real scenarios, such as deformation, will pose additional challenges in future deployment. Nevertheless, RAG-RUSS demonstrates that integrating retrieval-augmented generation with vision–language reasoning can enable accurate, interpretable, and autonomous US scanning. This is an essential step toward enhancing trust and acceptance among both sonographers and patients.

V. Conclusion

This work presents RAG-RUSS, an interpretable and autonomous framework for carotid US scanning that follows established clinical workflows and leverages VLM. To facilitate learning within a structured clinical workflow, we constructed a high-quality dataset from 32 human volumetric scans. Building on this dataset, we propose RAG-RUSS, which integrates an LLM, a vision foundation model, and a retrieval engine. In particular, the retrieval-augmented in-context learning strategy enhances adaptability and generalization while reducing the amount of training data required. Experimental results demonstrate that RAG-RUSS can autonomously perform full-stage carotid scanning in accordance with clinical protocols. Moreover, by providing explanations of the current image and the subsequent action, the system takes an important step toward developing trustworthy and acceptable RUSS for routine deployment.

References

[1] Z. Izadifar, P. Babyn, and D. Chapman, "Mechanical and biological effects of ultrasound: a review of present knowledge," *Ultrasound in medicine & biology*, vol. 43, no. 6, pp. 1085–1104, 2017.

[2] S. Sippel, K. Muruganandan, A. Levine, and S. Shah, "Use of ultrasound in the developing world," *International journal of emergency medicine*, vol. 4, no. 1, p. 72, 2011.

[3] Z. Jiang, S. E. Salcudean, and N. Navab, "Robotic ultrasound imaging: State-of-the-art and future perspectives," *Medical image analysis*, vol. 89, p. 102878, 2023.

[4] M. K. Welleweerd, A. G. de Groot, V. Groenhuis, F. J. Siepel, and S. Stramigioli, "Out-of-plane corrections for autonomous robotic breast ultrasound acquisitions," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 515–12 521.

[5] X. Ma, Z. Zhang, and H. K. Zhang, "Autonomous scanning target localization for robotic lung ultrasound imaging," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 9467–9474.

[6] D. Huang, Y. Bi, N. Navab, and Z. Jiang, "Motion magnification in robotic sonography: enabling pulsation-aware artery segmentation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 6565–6570.

[7] D. Huang, C. Yang, M. Zhou, A. Karlas, N. Navab, and Z. Jiang, "Robot-assisted deep venous thrombosis ultrasound examination using virtual fixture," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 381–392, 2024.

[8] Z. Wang, Y. Han, B. Zhao, H. Xie, L. Yao, B. Li, M. Q.-H. Meng, and Y. Hu, "Autonomous robotic system for carotid artery ultrasound scanning with visual servo navigation," *IEEE Transactions on Medical Robotics and Bionics*, 2024.

[9] Q. Huang, B. Gao, and M. Wang, "Robot-assisted autonomous ultrasound imaging for carotid artery," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–9, 2024.

[10] Y. Bi, Z. Jiang, F. Duelmer, D. Huang, and N. Navab, "Machine learning in robotic ultrasound imaging: Challenges and perspectives," *Annu. Rev. Control. Robotics Auton.*, vol. 7, 2024.

[11] Y. Huang, W. Xiao, C. Wang, H. Liu, R. Huang, and Z. Sun, "Towards fully autonomous ultrasound scanning robot with imitation learning based on clinical protocols," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3671–3678, 2021.

[12] H. Huang, W. Zheng, J. Duan, K. Huang, J. Guo, and C. Yang, "Learning globally stable neural imitation policies via reshaped energy functions," *IEEE/ASME Transactions on Mechatronics*, pp. 1–11, 2026.

[13] Z. Jiang, Y. Bi, M. Zhou, Y. Hu, M. Burke, and N. Navab, "Intelligent robotic sonographer: Mutual information-based disentangled reward learning from few demonstrations," *The International Journal of Robotics Research*, vol. 43, no. 7, pp. 981–1002, 2024.

[14] K. Su, J. Liu, X. Ren, Y. Huo, G. Du, W. Zhao, X. Wang, B. Liang, D. Li, and P. X. Liu, "A fully autonomous robotic ultrasound system for thyroid scanning," *Nat. Comm.*, vol. 15, no. 1, p. 4004, 2024.

[15] K. Li, J. Wang, Y. Xu, H. Qin, D. Liu, L. Liu, and M. Q.-H. Meng, "Autonomous navigation of an ultrasound probe towards standard scan planes with deep reinforcement learning," in *ICRA*. IEEE, 2021, pp. 8302–8308.

[16] T. Song, F. Li, Y. Bi, A. Karlas, A. Yousefi, D. Branzan, Z. Jiang, U. Eck, and N. Navab, "Intelligent virtual sonographer (ivs): Enhancing physician-robot-patient communication," *arXiv preprint arXiv:2507.13052*, 2025.

[17] 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 *et al.*, "Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning," *Science robotics*, vol. 10, no. 104, p. eadt5254, 2025.

[18] N. Roy, I. Posner, T. Barfoot, P. Beaudoin, Y. Bengio, J. Bohg, O. Brock, I. Depatie, D. Fox, D. Koditschek *et al.*, "From machine learning to robotics: Challenges and opportunities for embodied intelligence," *arXiv preprint arXiv:2110.15245*, 2021.

[19] Y. Liu, W. Chen, Y. Bai, X. Liang, G. Li, W. Gao, and L. Lin, "Aligning cyber space with physical world: A comprehensive survey on embodied ai," *IEEE/ASME Transactions on Mechatronics*, 2025.

[20] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang *et al.*, "Palm-e: An embodied multimodal language model," 2023.

[21] C. K. Ng, L. Bai, G. Wang, Y. Wang, H. Gao, K. Yuan, C. Jin, T. Zeng, and H. Ren, "Endovla: Dual-phase vision-language-action model for autonomous tracking in endoscopy," *arXiv preprint arXiv:2505.15206*, 2025.

[22] M. Chen, S. Fan, G. Cao, Y.-h. Liu, and H. Liu, "Uspilot: An embodied robotic assistant ultrasound system with large language model enhanced graph planner," *arXiv preprint arXiv:2502.12498*, 2025.

[23] H. Jiang, A. Zhao, Q. Yang, X. Yan, T. Wang, Y. Wang, N. Jia, J. Wang, G. Wu, Y. Yue *et al.*, "Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system," *Nature Communications*, vol. 16, no. 1, p. 7893, 2025.

[24] J. Yuan, S. Sun, D. Omeiza, B. Zhao, P. Newman, L. Kunze, and M. Gadd, "Rag-driver: Generalisable driving explanations with retrieval-augmented in-context multi-modal large language model learning," in *Robotics: Science and Systems*, 2024.

[25] A. Lasso, T. Heffter, A. Rankin, C. Pinter, T. Ungi, and G. Fichtinger, "Plus: open-source toolkit for ultrasound-guided intervention systems," *IEEE transactions on biomedical engineering*, vol. 61, no. 10, pp. 2527–2537, 2014.

[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[27] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in neural information processing systems*, vol. 36, pp. 46 595–46 623, 2023.

[28] S. Eslami, G. De Melo, and C. Meinel, "Does clip benefit visual question answering in the medical domain as much as it does in the general domain?" *arXiv preprint arXiv:2112.13906*, 2021.

[29] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. Seco de Herrera *et al.*, "Rocov2: Radiology objects in context version 2, an updated multimodal image dataset," *Scientific Data*, vol. 11, no. 1, p. 688, 2024.

[30] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[31] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models," *arXiv preprint arXiv:2407.07895*, 2024.