This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2026.3753092

Manuscript ID TMRB-06-26-OA-0678

LLM-enhanced Multi-level Skill Assessment for Robot-Assisted Vascular Interventions

Jintao Luo[1], Shuxiang Guo[1,2], Fellow, *IEEE*, Yonggan Yan[3], Member, *IEEE*, Bingzhi Shen[1], Sheng Cao[1], Student Member, *IEEE*, Pengcheng Li[2], Senior Member, *IEEE*, Bin Wang[4], Mingchao Ding[4]

Abstract: This study presents a multi-level framework for identifying and assessing procedural manipulations in robot-assisted vascular interventions. Nine motion-level manipulations are integrated into four decision-level manipulations, and a Transformer-based algorithm is developed for automatic recognition of these categories. On the reserved, label-consistency-filtered C2 test set containing 2,146 windows, the proposed algorithm achieved an accuracy of 93.76% and a Macro-F1 score of 93.01%. The experimental data were collected from six volunteers performing simulated vascular interventions using a vascular phantom and a robotic system. In addition, a large language model was used to generate comprehensive and understandable task-assessment reports based on seven quantitative indicators. An exploratory six-participant phantom comparison showed descriptive changes in process indicators after report-guided practice; it did not establish a causal effect on skill. These results demonstrate the feasibility of the proposed framework in the present volunteer-participated phantom study.

***Index Terms*—Robot-assisted surgery, operation assessment, vascular interventional surgery (VIS) robot, large language model (LLM)**

I. Introduction

In recent years, cardiovascular diseases have remained a major health burden [1]. Robot-assisted vascular intervention has motivated research on steerable catheter systems, force feedback, compliant guidewire control, force sensing, and surgical-state perception [3], [5], [7]-[9]. These developments aim to support remote instrument manipulation and operator feedback. Manipulation recognition and assessment provide complementary information about how operators perform an interventional task [13].

Recent developments also extend beyond conventional catheter robots. Chen et al. introduced a magnetic balloon fiberbot combining a magnetically steerable tip with a wirelessly heated phase-change balloon for catheter-free angioplasty in small arteries, with phantom and ex vivo validation [6]. Organ-specific magnetic soft robots combine compliant materials and remote actuation for functions including navigation, drug delivery, thrombus extraction, therapy, and sensing [10]. Gastrointestinal soft robotic systems similarly explore compliant structures, actuation, navigation, and minimally invasive functions [11]. These advances broaden the hardware context of robotic intervention. For operator-controlled systems, understanding procedural decisions and manipulation patterns remains a complementary research need.

In surgical manipulation recognition, multiple research teams have explored interventionalists' natural operating behavior to describe operating skills qualitatively or quantitatively. Sensor technology has enabled such exploration. Initially, individual sensors like surface electromagnetic (sEMG) sensors and data gloves were used to capture interventionalists' hand movements. Yet, unimodal information proved insufficient for accurate movement modeling. Multimodal data fusion from multiple sensors has shown improved detection and recognition compared to single-sensor data. Zhou et al. developed a data-fusion framework on a vascular simulation platform, incorporating muscle activity, hand, proximal force, and finger movements [14]. Subsequently, interventionalists' skills were analyzed qualitatively and quantitatively in animal in-vivo experiments to examine skill differences across vascular branches [15]. Additionally, an integrated fusion method was proposed based on Gaussian mixture regression and an HMM-based recognition model for intravascular hand-movement recognition [13], and the effect of ensemble learning algorithms was investigated in multi-sensor information fusion [16]. Wang et al. used four sensors to collect multimodal data from interventionalists and proposed a deep learning based multimodal information fusion architecture for identifying eight common operational behaviors. A hierarchical fusion strategy based on the Shannon entropy weight method was employed for feature fusion and behavior recognition, achieving average accuracies of 96.84% (model) and 93.26% (in-vivo) [18].

Despite high recognition accuracies, these methods have drawbacks. Firstly, substantial reliance on external sensors remains. During data collection, numerous external devices attached to doctors' hands constrain instrument manipulation and may render the collected data less reliable. The complex and time-consuming data-collection process also makes it difficult to amass large amounts of data for subsequent learning-based recognition algorithms. Secondly, these studies only identify some motion-level surgical manipulations, such as advancement, rotation, etc. However, in real surgery, these are fragmented within actual surgical procedures, do not align

This work was supported by the High–level of Special Funds under Grant G03034K003 from Southern University of Science and Technology, Shenzhen, China, in part by the Shenzhen "Medical Engineering Integration"Special Support Program Project under Grant F-2024-Z99-502365, Shenzhen, China (Corresponding author: Shuxiang Guo.)

J. Luo, S. Guo, B. Shen, and S. Cao are with The School of Life Science, Beijing Institute of Technology, Beijing 100081, China

S. Guo and P. Li are with The Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China.

Y. Yan is with The School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China

B. Wang and M. Ding are with The Department of Peripheral Vascular Intervention, Aerospace Center Hospital, Beijing 100081, China

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 14,2026 at 04:09:27 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2026.3753092

with the level of doctors' decision-making, and thus are inadequate to support subsequent research on surgical evaluation.

In terms of surgical manipulation assessment, early methods like the Universal Endovascular Rating Scale and Global Rating Scale (GRS) have drawbacks such as high subjectivity, large task volume, and cumbersome processes. In recent years, with the rapid development of computer technology, researchers have started to explore computer-assisted technology based on machine learning and deep learning. The objective is to establish a more objective and accurate evaluation method for vascular interventional surgery skills. Zhou et al. proposed an evaluation framework based on natural behavior to qualitatively and quantitatively evaluate the surgical skills of percutaneous coronary intervention. Myoelectric-signal sensors were attached to the fingers and arms of novice and experienced doctors to obtain signals associated with four natural behaviors: hand movement, finger movement, proximal force, and muscle movement. The Mann-Whitney U test was used to compare the behavioral characteristics of groups with different skill levels for effective behavioral selection. Finally, the Gaussian mixture model and Mahalanobis distance were used to evaluate the selected natural behavior qualitatively and quantitatively respectively, thus achieving the evaluation of the surgeon's operating skills [13], [14]. Akinyemi, T et al. proposed a novel framework that integrates high-fidelity generative models for minority-class augmentation and a specialized convolutional architecture, IAR-Net, for fine-grained recognition of seven catheterization actions. Leveraging augmented data, IAR-Net achieves 98.9% accuracy. The results demonstrate significant potential for improving procedural training and facilitating clinical adoption of robot-assisted endovascular systems [19].

However, all these evaluation methods share the following two common problems. Firstly, during the data collection process, these methods attach sensors to either the doctor's hands or medical instruments. This restricts the doctor's operation and thus affects the accuracy of the evaluation results. Secondly, the evaluation results of these methods are all abstract indicators. The resulting indicators can be difficult to interpret and translate into specific practice recommendations.

Large language models (LLMs), commonly based on Transformer architectures, can generate context-sensitive natural-language text [20]. In clinical practice, LLMs have been investigated for drafting discharge summaries, although their outputs may contain inaccuracies or omissions and require human review [21].

To address the above-mentioned research issues, this paper presents a novel robot-assisted vascular interventional surgery manipulation identification and assessment method. The main contributions are summarized as follows.

1) This study proposes a multi-level surgical manipulation classification method that integrates nine motion-level manipulations into four decision-level manipulations. A Transformer-based recognition algorithm is developed to identify these four manipulations. On the reserved, label-consistency-filtered C2 test set, the algorithm achieved an accuracy of 93.76% and a Macro-F1 score of 93.01%.
2) Data in this study were collected during volunteer-operated phantom tasks using the robotic system. Motion information was obtained from the haptic interface, and the system included force sensing (Fig. 3). The approach does not require wearable motion sensors on participants' hands. Routine robotic-session logging may facilitate future data collection, but collection during clinical procedures was not evaluated.
3) An LLM-assisted assessment method converts seven quantitative process indicators into a structured task-assessment report. An exploratory phantom comparison describes indicator changes after participants practiced with or without access to the complete report. This comparison does not establish a causal effect on skill or isolate the contribution of the LLM-generated narrative.

The remainder of this paper is structured as follows. Section II details the robot platform and method of surgical manipulation recognition; Section III introduces the method of surgical manipulation assessment. The experimental validation is performed in Section IV. Finally, Section V concludes this article.

II. The Method of Identification

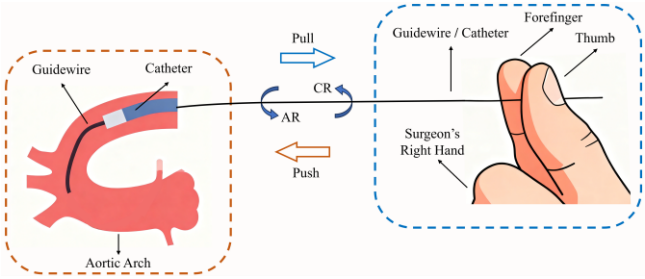


Fig. 1. Analysis of the actions performed by interventionalists when manipulating guidewire with the fingers.

During vascular intervention, guidewires are typically manipulated between the index finger and thumb while transmitted resistance is perceived through contact with the wire (Fig. 1). Guidewire maneuvers can be classified into three fundamental categories: translational, rotational, and combined motions. Translational motion involves the interventionalist grasping the guidewire with two fingers to perform side-to-side displacements, executing pushing and pulling actions to control the wire's forward and backward progression through the vasculature. When the guidewire reaches a specific position, it encounters a vessel bifurcation, such as in the carotid and coronary arteries. At this juncture, the interventionalist can rotate the guidewire either clockwise (CR) or anticlockwise (AR). Given that the tip of the guidewire features a curved section that can be oriented in various directions during rotation, interventionalists can select different vessel branches for access. Furthermore, during the procedure, interventionalists often employ a combination of translational and rotational motions to minimize vessel damage and enhance procedural efficiency. A multilevel classification approach is proposed to characterize procedural manipulation. The framework encompasses motion-level classification, decision-level recognition, and process-based

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 14,2026 at 04:09:27 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2026.3753092

assessment (Fig. 3).

A. Surgical Robot Platform

Vascular interventional surgical robots typically comprise two components: the leader end and the follower end. Surgeons operate the leader end to remotely control the follower end, which performs specific surgical procedures on patients. Fig. 2 shows the VISR designed in this study. The leader end is a commercial haptic force-feedback handle (Geomagic TouchX, 3D Systems inc., USA). Surgeons can operate this tactile handle, which collects the hand movement information of the surgeon and provides tactile force feedback. The follower end mainly consists of a drive module, a fixing module, a lead screw (EBX1605-600, Times Brilliant inc., China), and a closed-loop stepper motor (STM5657B, NiMotion inc., China). In actual surgical procedures, guidewire movement requires two degrees of freedom: linear and rotational. The drive module clamps and rotates the guide wire. The lead screw and stepper motor work together to drive the drive module in a straight line, thereby moving the guide wire forward or backward. Finally, the PID algorithm is used to control the interaction between the two ends.

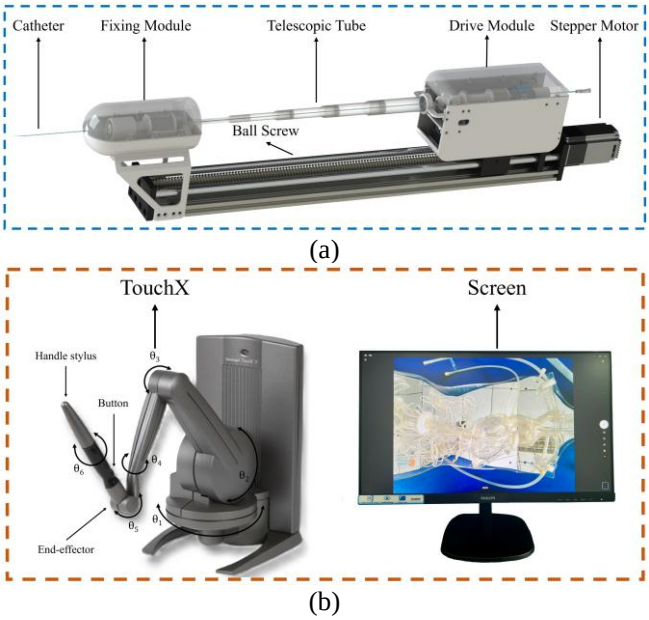


Fig. 2. The VISR designed in this study. (a) The follower end of the VISR. (b) The leader end of the VISR.

TABLE I
Category Names and Abbreviations of MLSM

Motion-level surgical manipulations	Abbreviation	(T, R)
Stop	SP	(0, 0)
Push	PH	(1, 0)
Pull	PL	(-1, 0)
Clockwise rotation	CR	(0, 1)
Anticlockwise rotation	AR	(0, -1)
Push and clockwise rotation	HC	(1, 1)
Push and anticlockwise rotation	HA	(1, -1)
Pull and clockwise rotation	LC	(-1, 1)
Pull and anticlockwise rotation	LA	(-1, -1)

B. Motion-Level Surgical Manipulations, MLSM

Based on the preceding motion analysis, manipulations are classified into nine motion-level categories. The category names and abbreviations are presented in Table I. Subsequently, relying on the control data gathered by the TouchX handle, the motion status of each sampling is logged. Translational advancement, retraction, and stopping are encoded as 1, -1, and 0, respectively. The rotational state is recorded at the same sampling time. A clockwise rotation is noted as 1, a counter - clockwise rotation as - 1, and stopping as 0. In accordance with this approach, a record of the motion state can be obtained each time a sample is taken, in the format of (TRANSLATION, ROTATION). The nine possible combinations each correspond to one motion-level category (Table I).

C. Decision-Level Surgical Manipulations, DLSM

Currently, most studies only classify surgical procedures at these motion-level surgical manipulations. However, in actual surgery, these manipulations are very brief and fragmented, and do not align with the doctor's actual surgical decision-making level. Surgical procedures that only determine the motion-level are insufficient to support subsequent assessment of surgical skills. These motion-level manipulations are therefore grouped into four decision-level categories according to procedural intent. The category names and classification criteria are presented as follows.

1) Steady Progression Phase (SPP): When the anatomical structure of blood vessels is relatively regular (with no obvious branches, stenosis, or tortuosity), surgical instruments are continuously and smoothly advanced along the axis of the target blood vessel. This stage is characterized by continuous propulsive movements. The equipment moves rapidly, and the average force applied is small. The core features are the continuity of the operation and low resistance.

2) Deviation Correction Phase (DCP): When the vascular environment is complex (such as in tortuous blood vessels and those with dense branches), after the surgical instrument collides with the blood vessel wall or strays into a non - target branch, this operation stage involves correcting the trajectory through pullback adjustments. At this stage, the pullback action is the primary operating method. The movement speed of the instrument is significantly reduced, the average force is relatively large, and the core features are the retroactivity and high resistance of the operation.

3) Vascular Super-selective Cannulation Phase (VSP): When the surgical instrument reaches the bifurcation site of the blood vessel, it enters the operation stage of selectively entering a specific branch blood vessel based on the surgical goal. This stage requires a combination of frequent instrument rotation, advancement, pullback, and compound movements to accurately locate the branch entrance. The operation speed is slow, and the movements are highly precise. The average operating force is large. The core feature is the compound and targeted nature of the movements.

4) Transient Stagnation Phase (TSP): This is a brief quiescent phase during surgery for observing vascular images, evaluating operation effects, or making real - time decisions. At this stage, the translational displacement, rotation angle, and force values of the instrument are all within a small

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 14,2026 at 04:09:27 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2026.3753092

floating range or are completely zero. The core features are the pause in the operation and the stability of the mechanical parameters.

D. DLSM Recognition Algorithm

DLSM recognition (Fig. 3) uses the MLSM representation, the guidewire-force signal, and axial translational velocity as multivariate time-series inputs.

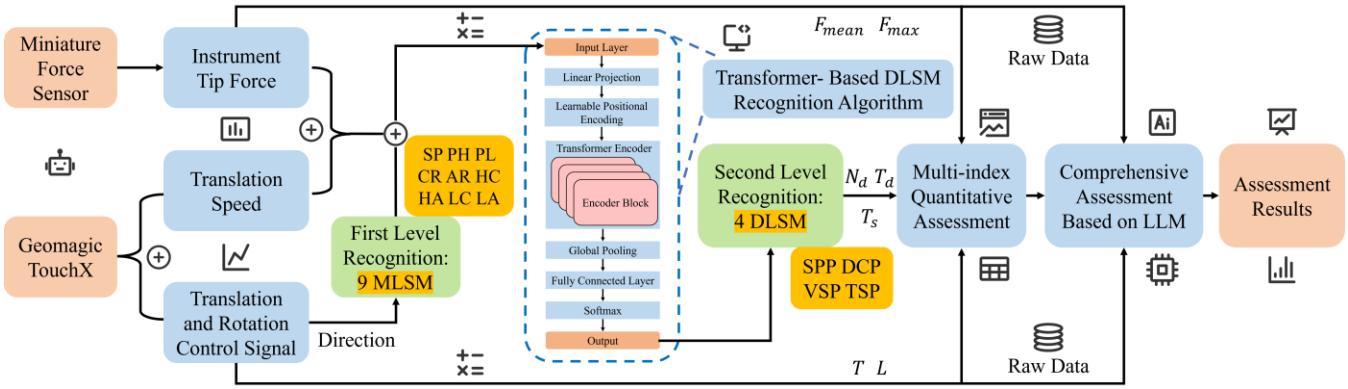


Fig. 3. The technical framework of this study illustrates the data transmission and the relationships among various components, including data collection, multi-level manipulation recognition, the DLSM recognition algorithm, multi-index quantitative assessment, and the final comprehensive assessment by the LLM.

The continuous force and translational-velocity channels are combined with the one-hot-encoded nine-class MLSM representation before linear projection. Rotational direction contributes to the MLSM label described in Section II-B.

The algorithm's core feature extraction module is built on a Transformer encoder. First, input features are projected to the model dimension *d_model* via a linear layer to match the Transformer's internal dimension. To address the Transformer's inherent lack of temporal awareness, learnable positional encoding is integrated, enabling the model to capture temporal sequence information through parameter learning and adapt flexibly to diverse temporal patterns.

Following positional encoding enhancement, features are fed into a Transformer encoder composed of *num_layers* stacked encoder blocks. Each block incorporates a multi-head self-attention mechanism and a feed-forward network, supplemented by normalization and dropout operations. The multi-head self-attention mechanism splits input features into *n_heads* subspaces for parallel attention calculation, effectively capturing inter-temporal dependencies of the same variable and cross-variable interactions at the same time step. The feed-forward network comprises two linear layers with a hidden dimension of *dim_feedforward*, using the ReLU activation function for non-linear transformation of attention outputs. Batch normalization is adopted to stabilize training, while dropout is applied to the attention and feed-forward layers to alleviate overfitting.

In the classification stage, sequence features output by the Transformer encoder are aggregated into a global feature vector via global pooling, then mapped to the category dimension through a fully connected layer. Finally, a softmax function generates category probability distributions to realize classification prediction. During training, data are loaded in batches of size *batch_size*, with the model iteratively optimized over *num_epochs* using the Adam optimizer and cross-entropy loss function. For inference, preprocessed input data undergo forward propagation to obtain classification probabilities, with the category corresponding to the maximum probability taken as the final prediction.

III. The Method Of Assessment

The assessment method (Fig. 3) combines seven quantitative process indicators with LLM-generated qualitative summaries. The intended output is an understandable report that highlights manipulation patterns and suggests areas for subsequent phantom-task practice. The present study evaluates descriptive process-indicator changes and does not establish clinical effectiveness.

A. Multi-Index Quantitative Assessment

The following seven process indicators summarize duration, motion, force, and recognized procedural phases in the phantom task. These indicators are interpreted jointly for the same task; lower values alone do not demonstrate successful cannulation, greater safety, or improved clinical skill.

1) Total Operation Time (T): T is the duration from instrument entry into the phantom to the end of the designated task. Under the same task conditions, shorter duration may indicate fewer redundant actions, but must be interpreted together with task completion and the other process indicators.

2) Average Force (Fmean) and Maximum Force (Fmax): These indicators summarize the recorded axial force signal. As used here, the signal reflects resistance transmitted to the measurement location; it is not a direct measurement of vessel-wall injury. The mean and maximum are calculated as follows:

$$F_{mean} = \frac{1}{n}\sum_{i=1}^{n} F_i \tag{1}$$

$$F_{max} = Max(F_1, F_2, \cdots, F_n) \tag{2}$$

In these formulas, Fi is the recorded axial force at the i-th sampling time, and n is the total number of samples.

3) Number of DCP Episodes (Nd) and Total DCP Duration (Td): DCP denotes pullback correction after wall contact or

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 14,2026 at 04:09:27 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2026.3753092

entry into a non-target branch. Nd counts recognized DCP episodes, and Td sums the durations of these episodes. These quantities describe correction-related behavior and depend on the recognition output; these quantities do not independently adjudicate injury or procedural success.

4) Total VSP Duration (Ts): Ts is the total duration assigned to targeted branch engagement at a bifurcation. It summarizes time spent in the recognized VSP stage under the same phantom task. Its interpretation depends on recognition quality and task context, rather than serving as a validated clinical skill score.

5) Cumulative Translational Travel (L): L is calculated from successive X-axis handle coordinates using (3). It is a handle-derived motion measure used to summarize cumulative translation during the task, rather than a reconstructed three-dimensional guidewire-tip trajectory.

$$L = \sum_{i=1}^{n-1} \sqrt{\left(x_{i+1} - x_i\right)^2} \tag{3}$$

In the formula, x_i represents the X-axis coordinate of the handle at the i-th sampling moment.

B. LLM-Based Multi-Index Assessment System

The assessment system comprises data preprocessing, DLSM recognition, calculation of the seven quantitative indicators, LLM-based qualitative analysis, report export, and a memory/database module. For each case, the LLM context is assembled from a fixed system prompt, task and instrument metadata, the seven metric values, the recognized DLSM stages and the associated time points, and benchmark records from the database for the same phantom task. The fixed prompt defines the model role, the task-specific interpretation of each metric and DLSM stage, the scoring logic, and the required report structure.

The LLM interprets precomputed quantitative indicators and recognized procedural stages rather than directly classifying surgical actions. Related research has explored LLM-generated coaching feedback for surgeons' non-technical skills [22], suggesting a potential role for natural-language feedback in surgical training.

The implemented interface supports GLM-5.1, MiniMax-M2.5, Kimi-K2.5, DeepSeek-V3.2, and Qwen-3.5. Based on task comprehension, interpretation of process indicators, identification of key procedural stages, and the specificity of improvement suggestions, GLM-5.1 was used for the reported system output. The generation temperature was set to 0.2; no manual random seed was specified by the API. The model receives the complete context in a single request and returns a structured report containing an overall score, metric-level assessment, stage-specific analysis, benchmark comparison, and actionable recommendations. Blinded expert ratings, repeated-generation consistency, and a comparison with a rule-based report were not evaluated; the present comparison therefore cannot validate an LLM-specific benefit.

IV. Experimental Validation

A. Operating Skills Data Acquisition

The data acquisition platform (Fig. 4) consists of the vascular intervention robot and a silicone endovascular evaluator (General Angiography Type C, FAIN-Biomedical Inc., Japan). An industrial camera recorded each acquisition for annotation. Three recording types were collected: single-manipulation recordings (A), mixed-manipulation recordings (B), and simulated procedural recordings (C). The procedural recordings were organized into C1 for training and optimization and C2 for final testing, as described in Section IV-C.

1) ***Single Operation Pure Sample:*** only a single operation is performed at each data acquisition, with a pause between each operation. Repeat the single operation multiple times at each sampling.
2) ***Multiple Operations Mixed Samples:*** Multiple operations are performed at each data acquisition, with a pause between each operation. The order between the four operations is not fixed at each sampling time.
3) ***Simulated Procedural Recordings (C):*** Participants performed the phantom navigation task from the femoral artery to the left subclavian artery (Fig. 5).

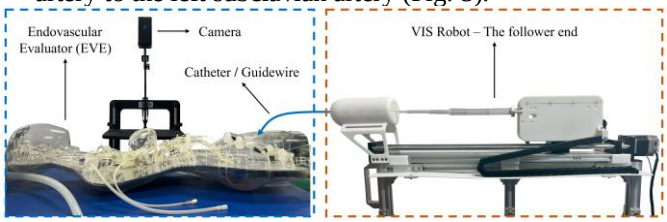


Fig. 4. Data collection and experimental platform utilizing an endovascular evaluator and the VISR.

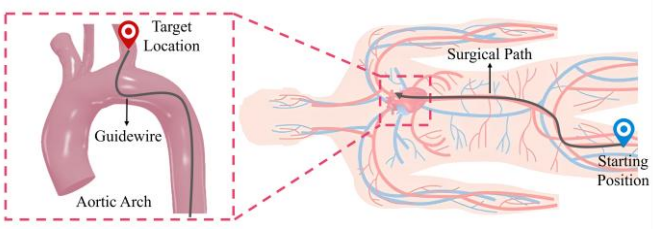


Fig. 5. Navigation path used in the phantom assessment experiment.

After establishing the data acquisition platform, six volunteers were recruited to perform the three types of data-collection tasks described above. The sampling frequency was set to 50 Hz, and a system timestamp was recorded with each data point to support subsequent annotation. An industrial camera was used to record each acquisition procedure, and written records of the manipulation process and key timestamps were maintained by a dedicated observer. All data were collected during robot-assisted procedures performed on the vascular phantom rather than during clinical interventions.

B. Data Preprocessing and Datasets Production

The guidewire-force signal and axial translational velocity were normalized, and the instantaneous nine-class MLSM state was one-hot encoded, consistently with Section II-D and Fig. 3. The four DLSM stages were manually annotated using acquisition videos, written records, and timestamps. DCP denotes pullback correction after wall contact or entry into a non-target branch, whereas VSP denotes targeted branch engagement at a bifurcation. These reference labels therefore incorporate procedural context from

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 14,2026 at 04:09:27 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2026.3753092

video; prediction uses the input signals rather than video. Recordings were segmented into 50-sample windows with a 25-sample step at 50 Hz, corresponding to 1 s and 50% overlap. Mixed-label windows were assigned the majority label within each window. Incomplete terminal windows and windows with label consistency below 80% were excluded before evaluation. The resulting recognition metrics characterize label-consistent segments, not the complete unfiltered stream.

C. DLSM Recognition Model Training and Validation

The experiments used two Intel Xeon Platinum 8124M CPUs, 64 GB of memory, and four NVIDIA GeForce RTX 2080 SUPER GPUs. Recordings A, B, and C1 yielded 11,587 valid windows for training and hyperparameter optimization. The ten C2 final-test sequences were reserved for final evaluation and were not used for training or hyperparameter selection. Segmentation of these sequences yielded 2,545 candidate windows before the exclusions described in Section IV-B.

The model was trained using the Adam optimizer with an initial learning rate of 0.001 and a weight decay of 1×10-4. The batch size was 32, the feed-forward network dimension was 512, and the maximum number of training epochs was 100. Weighted cross-entropy loss was used to mitigate the effect of class imbalance, particularly for the minority DCP class. Dropout and an early-stopping strategy with a patience of 10 epochs were employed to reduce overfitting.

A grid search combined with five-fold cross-validation within the training and optimization set was used to assist hyperparameter selection. Because this cross-validation procedure was not designed as a subject-independent evaluation, its results are not presented as an independent estimate of model generalization performance. Macro-F1 was used as the main selection metric because of the class imbalance. The selected hyperparameters were d_model=256, n_heads=4, num_layers=4, and a dropout probability of 0.1.



Fig. 6. Confusion matrices on the 2,146-window label-consistency-filtered C2 test set. (a) DTW. (b) This work. (c) MultiRocket. (d) InceptionTime.

DTW, InceptionTime, and MultiRocket were compared on the same reserved C2 test set (Table II and Fig. 6). The proposed method had the highest observed accuracy (93.76%) and Macro-F1 (93.01%). Its DCP F1 was 86.12%, 6.79 percentage points above MultiRocket (79.33%). Fig. 6(b) contains 33 VSP-to-DCP and 19 DCP-to-VSP errors, compared with 156 and 31 for DTW in Fig. 6(a). This demonstrates discrimination between the intent-annotated stages in this dataset, without establishing direct inference of operator intent. Table II reports accuracy and F1 metrics; precision and recall are reported in Fig. 6. These are single-run results; variability across repeated training runs was not quantified.

TABLE II

Comparison of Classification Performance on the Reserved C2 Phantom Test Set

Indicators	DTW	InceptionTime	MultiRocket	**Ours**
Accuracy	84.72%	90.59%	92.22%	93.76%
Macro-F1	**82.23%**	**89.51%**	**91.33%**	**93.01%**
F1-VSP	**89.37%**	**92.41%**	**93.46%**	**95.12%**
F1-DCP	**58.20%**	**77.54%**	**79.33%**	**86.12%**
F1-SPP	87.81%	90.10%	93.54%	92.32%
F1-TSP	93.53%	97.98%	98.99%	98.48%

D. Exploratory Phantom Assessment Experiment

1) Experimental Design: Six participants completed an exploratory phantom comparison using the same robot and navigation task from the femoral artery to the left subclavian artery (Fig. 5). The cohort comprised three participants experienced with the simulation and robotic system and three novices with basic training. Simple random assignment allocated three participants to each group without stratification by experience; no blinding was implemented. The report group contained one experienced participant and two novices, and the blank control group contained two experienced participants and one novice. Participants are identified consistently as S1-S3 in the report group and S4-S6 in the blank control group.

Each participant performed five trials in Exp1. For each participant and indicator, the largest and smallest of the five values were omitted from the summary, and the remaining three were averaged. All five measurements were retained in the study records. This trimmed mean is an exploratory summary intended to reduce sensitivity to occasional extreme values; it is not a formal outlier test or a replacement for repeated-measures analysis. The participant-level ROC values are reported in Table III. A larger study should report all repeated measurements and use an appropriate repeated-measures analysis.

The Exp1 indicators, recognized DLSM stages, and task-matched database benchmarks were analyzed using GLM-5.1 and the configuration in Section III-B. Participants S1-S3 received the complete assessment report and used it to guide subsequent practice. Participants S4-S6 practiced without access to the report. The intervention was the complete report package, not an isolated test of the LLM narrative.

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 14,2026 at 04:09:27 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2026.3753092

One week later, the same six participants repeated the same task five times (Exp2). The same participant-level, indicator-specific trimmed-mean procedure was applied.

TABLE III
Participant-Level ROC Between Exp1 and Exp2

Indicators	S1	S2	S3	S4	S5	S6
T (s)	-7.9%	-31.6%	-5.3%	-1.6%	-3.7%	-8.0%
L (mm)	-9.5%	-28.3%	-8.9%	-2.3%	-9.8%	-15.0%
F_{mean} (N)	-10.5%	-17.2%	5.0%	0.0%	-4.2%	8.0%
F_{max} (N)	-6.1%	-24.5%	9.7%	-3.2%	13.9%	-7.0%
T_S (s)	-19.7%	-40.6%	-20.7%	-9.9%	-5.4%	-3.7%
T_D (s)	-24.4%	-44.0%	-12.1%	-3.3%	0.4%	-27.6%
N_D	-14.3%	-33.3%	-13.8%	-7.7%	-10.3%	-7.3%

2) Results and Analyses of the Two Experiments: The relative rate of change (ROC) was utilized to quantitatively analyze the alterations in the results of the two experiments. The calculation formula was as follows:

$$ROC = \frac{Exp2 - Exp1}{Exp1} \times 100\% \quad (4)$$

Table III reports participant-level ROC values, and Table IV summarizes the mean ROC for each group. Negative ROC indicates a reduction in the measured process indicator, and positive ROC indicates an increase. These directions do not by themselves establish a change in clinical skill.

TABLE IV
Mean Participant-Level ROC by Group

Indicators	Experimental Group	Blank Control Group
T (s)	-14.9%	-4.4%
L (mm)	-15.6%	-9.0%
F_{mean} (N)	-7.6%	**1.3%**
F_{max} (N)	-7.0%	**1.2%**
T_S (s)	-27.0%	-6.3%
T_D (s)	-26.8%	-10.2%
N_D	-20.5%	-8.4%

Fig. 7 displays participant-level indicator profiles for Exp1 and Exp2. Indicators were reverse-coded and min-max normalized across the available results for each indicator, so larger radial values represent lower measured values. The radar area is a visualization of these transformed process indicators, not a validated aggregate skill score.

The report group had more negative mean ROC values than the blank control group for all seven indicators (Table IV). This descriptive pattern is compatible with changes after report-guided practice, but practice effects, baseline experience imbalance, and the small sample prevent a causal interpretation.

For example, S2 showed a larger reduction than S6 in each of the seven indicators (Table III). This is a participant-level observation and does not establish that report access caused the difference or that S2 attained a higher validated skill level.

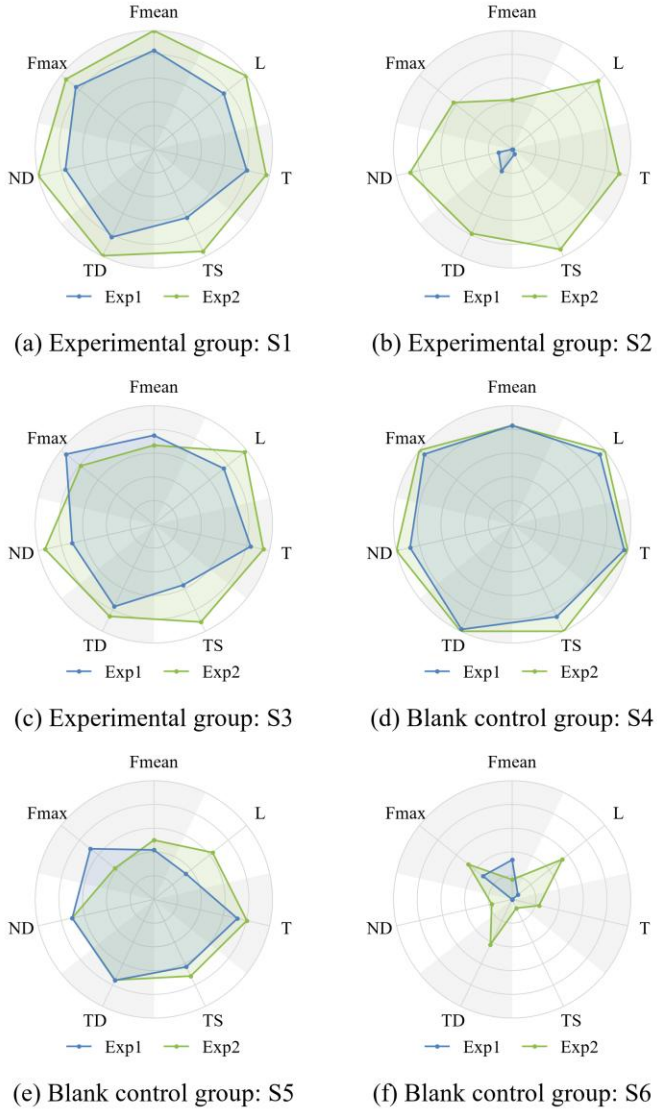


Fig. 7. The radar chart depicts the variations in seven indicators of six subjects across the two experiments.

The recognition experiment has several limitations. First, windows with a label consistency below 80%, most of which were located near transitions between procedural phases, were excluded because these windows did not provide an unambiguous single-class window-level label. Therefore, the reported accuracy does not represent performance on the complete unfiltered procedural stream. Second, the present evaluation was not designed as a subject-independent study, and leave-one-subject-out validation was not performed. The recognition results are consequently limited to the volunteers and phantom-based recordings included in this study. Third, inference latency and model-storage size were not systematically benchmarked on the intended deployment hardware. Transition-aware annotation, subject-independent validation, and computational-efficiency evaluation will be investigated in future work. Systematic ablations of input channels, window length, and architecture were not performed. MLSM encodes directional information from the control

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 14,2026 at 04:09:27 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2026.3753092

signals and may overlap with information in the continuous motion input; its incremental contribution remains unquantified.

The assessment experiment also has limitations. With only three participants per group, baseline experience imbalance, and no blinding, it supports descriptive observations rather than causal skill-improvement claims. The blank control cannot separate an LLM-specific benefit from the effect of receiving individualized feedback; indicator-only and rule-based-report controls are needed. Blinded expert ratings and repeated-generation consistency analysis are also needed. No separately adjudicated cannulation-success, vessel-contact, or perforation endpoints were included. The seven process indicators therefore do not establish procedural success or safety. Future work should prespecify these endpoints and examine how recognition errors affect stage-based metrics and reports.

V. Conclusion

This study presents a multi-level framework for procedural manipulation recognition and assessment in robot-assisted vascular intervention. Nine motion-level manipulations were integrated into four decision-level manipulations, and a Transformer-based algorithm was developed for automatic recognition of these categories. On the reserved, label-consistency-filtered C2 test set containing 2,146 windows, the proposed model achieved an accuracy of 93.76% and a Macro-F1 score of 93.01%. These results demonstrate the feasibility of the proposed recognition method for label-consistent segments in the present volunteer-participated phantom study.

The LLM-assisted system organizes seven quantitative process indicators into structured assessment reports. The six-participant phantom comparison provides descriptive observations after report-guided practice; it does not establish causal skill improvement, an LLM-specific benefit, or clinical effectiveness. Recognition on unfiltered phase transitions, subject-independent generalization, and real-time deployment also remain unvalidated. Future work will address transition-aware annotation, independent participants, input and architecture ablations, deployment-hardware benchmarking, active feedback controls, report-quality validation, and prespecified task-success and safety endpoints.

References

[1] M. Liu, X. He, X. Yang, “Summary of the China Cardiovascular Health and Disease Report 2023 (Prevalence of Cardiovascular Diseases and the Status of Interventional Diagnosis and Treatment),” *Chinese Journal of Interventional Cardiology*, vol. 32, no. 10, pp. 541-550, 2024.

[2] S. Chen, G. Zhang, J. Lai, B. Shen, “Hierarchical part-based generative model for realistic 3d blood vessel,” *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2025, pp. 257–267.

[3] Y. Yan, S. Guo, H. Shen, B. Wang and M. Ding, “A Novel Dexterous Steerable Catheter Robot System: Design, Modeling and Evaluation,” *IEEE Robotics and Automation Letters*, vol. 10, no. 9, pp. 8770-8777, Sept. 2025.

[4] Y. Yan et al., “5-DOF Microcoil Positioning System Utilizing Single-Axis Electromagnetic Transmitter,” *IEEE/ASME Transactions on Mechatronics*, vol. 30, no. 6, pp. 5213-5224, Dec 2025.

[5] X. Bao, S. Guo, C. Yang and L. Zheng, “Haptic Interface with Force and Torque Feedback for Robot-Assisted Endovascular Catheterization,” *IEEE/ASME Transactions on Mechatronics*, vol. 29, no. 2, pp. 1111-1125, 2024.

[6] H. Chen, Y. Sun, H. Peng, Y. Lin, N. Zhang, J. Li, M. Sui, C. Cai, S. Ye, R. Li, and L. Wang, “A magnetic balloon fiberbot for catheter-free angioplasty in small arteries,” *Soft Sci.*, vol. 6, Art. no. 48, 2026, doi: 10.20517/ss.2026.28.

[7] C. Lyu, S. Guo, W. Zhou, “A deep-learning-based guidewire compliant control method for the endovascular surgery robot,” *Micromachines*, vol. 13, no. 12, p. 2237,2022.

[8] C. Lyu, S. X. Guo, Y. G. Yan et al., “Deep-Learning-Based Force Sensing Method for a Flexible Endovascular Surgery Robot,” *IEEE Transactions on Instrumentation and Measurement*, vol. 73, p. 10, 2024 2024, Art no. 7503910.

[9] Y. Yan, H. Wang, H. Yu, “Machine learning-based surgical state perception and collaborative control for a vascular interventional robot,” *IEEE Sensors Journal*, vol. 22, no. 7, pp. 7106–7118, 2022.

[10] Z. Chen, S. Yang, J. Law, X. Du, and Y. Sun, “Magnetic soft robots in medicine: Material and structural designs for organ-specific applications,” *Soft Sci.*, vol. 6, Art. no. 3, 2026, doi: 10.20517/ss.2025.97.

[11] M. Du, Z. Xie, and X. Wang, “Advances of soft robotics for gastrointestinal tract applications,” *Soft Sci.*, vol. 5, Art. no. 51, 2025, doi: 10.20517/ss.2025.61.

[12] Y. Yan, S. X. Guo, Z. J. Lin et al., “Grasping-Force-Based Passive Safety Method for a Vascular Interventional Surgery Robot System,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, 2023.

[13] X. H. Zhou, X. L. Xie, Z. Q. Feng et al., “A Multilayer and Multimodal-Fusion Architecture for Simultaneous Recognition of Endovascular Manipulations and Assessment of Technical Skills,” *IEEE Transactions on Cybernetics*, vol. 52, no. 4, pp. 2565-2577, Apr 2022.

[14] X. Zhou, G. Bian, X. Xie, “Qualitative and Quantitative Assessment of Technical Skills in Percutaneous Coronary Intervention: In Vivo Porcine Studies,” *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 2, pp. 353-364, Feb. 2020.

[15] X. Zhou, X. Xie, S. Liu, “Surgical Skill Assessment Based on Dynamic Warping Manipulations,” *IEEE Transactions on Medical Robotics and Bionics*, vol. 4, no. 1, pp. 50-61, Feb. 2022.

[16] X. Zhou, X. Xie, S. Liu, “Learning Skill Characteristics From Manipulations,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 9727-9741, Dec. 2023.

[17] S. Wang, Z. Liu, W. Yang, “Learning-Based Multimodal Information Fusion and Behavior Recognition of Vascular Interventionists' Operating Skills,” *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 9, pp. 4536-4547, Sept. 2023.

[18] S. Wang, W. L. Zhang, L. Zhao, G. Z. Yang, and L. Xie, “Task-Oriented Interventionists' Operating Skills Recognition Based on Multimodal Data Hierarchical Fusion,” *IEEE Sensors Journal*, vol. 25, no. 11, pp. 20732-20742, Jun 1 2025.

[19] T. Akinyemi, O. Omisore, W. J. Du et al., “IAR-Net: Tabular Deep Learning Model for Interventionalist's Action Recognition,” *Advanced Intelligent Systems*, 2025.

[20] F. Liu et al., “Application of large language models in medicine,” Nat. Rev. Bioeng., vol. 3, no. 6, pp. 445–464, 2025, doi: 10.1038/s44222-025-00279-5.

[21] C. Y. K. Williams et al., “Physician- and Large Language Model–Generated Hospital Discharge Summaries,” JAMA Intern. Med., vol. 185, no. 7, pp. 818–825, 2025, doi: 10.1001/jamainternmed.2025.0821.

[22] M. Obuseh et al., “Feasibility of large language models for assessing and coaching surgeons’ non-technical skills,” npj Health Syst., vol. 2, Art. no. 25, 2025, doi: 10.1038/s44401-025-00027-2.

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 14,2026 at 04:09:27 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.