IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. XX, NO. XX, XXXX 2020 1

LMT++: Adaptively Collaborating LLMs with Multi-specialized Teachers for Continual VQA in Robotic Surgical Videos

Yuyang Du, Kexin Chen, Yue Zhan, Chang Han Low, Mobarakol Islam, Ziyu Guo, Yueming Jin, Guangyong Chen, Pheng Ann Heng *Senior Member, IEEE*,

***Abstract*— Visual question answering (VQA) plays a vital role in advancing surgical education. However, due to the privacy concern of patient data, training VQA model with previously used data becomes restricted, making it necessary to use the exemplar-free continual learning (CL) approach. Previous CL studies in the surgical field neglected two critical issues: i) significant domain shifts caused by the wide range of surgical procedures collected from various sources, and ii) the data imbalance problem caused by the unequal occurrence of medical instruments or surgical procedures. This paper addresses these challenges with a multimodal large language model (LLM) and an adaptive weight assignment strategy. First, we developed a novel LLM-assisted multi-teacher CL framework (named LMT++), which could harness the strength of a multimodal LLM as a supplementary teacher. The LLM's strong generalization ability, as well as its good understanding of the surgical domain, help to address the knowledge gap arising from domain shifts and data imbalances. To incorporate the LLM in our CL framework, we further proposed an innovative approach to process the training data, which involves the conversion of complex LLM embeddings into logits value used within our CL training framework. Moreover, we design an adaptive weight assignment approach that balances the generalization ability of the LLM and the domain expertise of conventional VQA models obtained in previous model training processes within the CL framework. Finally, we created a new surgical VQA dataset for model evaluation. Comprehensive experimental findings on these datasets show that our approach surpasses state-of-the-art CL methods.**

***Index Terms*— Surgical education, visual question answering, continual learning, multi-modal large language model**

The work was partially supported by the Research Grants Council of the Hong Kong SAR, China (Project Number: T45-401/22-N), by the Hong Kong Innovation and Technology Fund (Project Number: GHP/080/20SZ), by the Regional Joint Fund of Guangdong (Guangdong-Hong Kong-Macao Research Team Project) under Grant 2021B1515130003, by the Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, and by the Ministry of Education Tier 1 Start-up grant, National University of Singapore, Singapore (A-8001267-01-00).

Y. Du, K. Chen, Y. Zhan, Z. Guo, and P. A. Heng are with the Chinese University of Hong Kong, Hong Kong SAR, China. P. A. Heng is also with Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China (email: {dy020, yuezhan}@ie.cuhk.edu.hk, {kxchen, zyguo, pheng}@cse.cuhk.edu.hk)

C. Low and Y. Jin are with National University of Singapore, Singapore (email: e1127374@u.nus.edu, ymjin@nus.edu.sg)

M. Islam is with the Department of Medical Physics and Biomedical Engineering, University College London, London, UK (email: mobarakol.islam@ucl.ac.uk).

G. Chen is with Zhejiang Lab, Zhejiang, China (email: gychen@zhejianglab.com).

Project availability: https://github.com/yuyangdu01/LLM-CL-VQA.

An early version of this paper has been presented in IEEE ICRA 2024.

Yueming Jin is the corresponding author.

Yuyang Du, Kexin Chen, and Yue Zhan are equal contributors.

I. Introduction

HIGH-QUALITY surgical education is instrumental in the professional advancement of clinical students, as it equips them with the necessary knowledge and skills to perform complex procedures and deliver excellent patient care. However, traditional teaching methods, such as lectures and textbooks, may not always sufficiently address the diverse questions and concerns that students encounter during their learning process. Expert surgeons serve as the primary source of clinical students' surgical knowledge, but they may not always be available to provide immediate feedback due to their demanding clinical and academic responsibilities [1]–[3]. In recent years, surgical visual question answering (VQA) models have garnered significant research interest [4], [5]. These models are typically trained using expert demonstration videos or related images and can provide students with instant access to expert knowledge, enabling them to clarify their doubts and deepen their understanding of surgical procedures. Moreover, the integration of VQA models into intelligent systems strengthens their skills in grasping and interpreting surgical scenes, thus establishing the groundwork for the development of clinical assistance technologies like advanced surgical robots [6], [7].

In surgical VQA, the needs of trainees are constantly evolving, such as learning more surgical types and adapting to different clinical systems. Additionally, enhanced surgical techniques and new instruments are regularly introduced to improve patient care. This will inevitably create new surgical environments (i.e., recently developed surgical settings) and generate new question-and-answer sets, thus resulting in a myriad of fresh and creative VQA tasks. Given the rapid update of surgical knowledge in VQA tasks, it is crucial to leverage continual learning (CL) methods to overcome the catastrophic forgetting problem [8].

Catastrophic forgetting has been largely resolved in the healthcare and medical sector by early efforts that adapted CL algorithms from general domains [9], [10]. For example, [9] developed a replay-oriented CL algorithm for medical

image analysis. These studies are exemplar-based, where old patient data is accessible during model updates. However, the exemplar-based approach is infeasible in practical surgical VQA scenarios due to various constraints such as exorbitant data storage costs, data privacy concerns, and complicated and sensitive issues arising from the license of obtaining such data from different medical centers.

Recent developments in CL studies placed significant emphasis on efficient knowledge updating under data privacy constraints. Learning without Forgetting (LwF) [11] and Elastic Weight Consolidation (EWC) [12] are widely used non-exemplar-based CL approaches in medical studies like [13]. Additionally, [14] proposed CL algorithms for biological tissues and surgical tools spatial identification tasks alongside VQA, with particular attention to overlapping classes present in both the current and previous datasets.

In this paper, we observed two key characteristics within the medical and surgical domain that have been widely overlooked in previous studies. These two characteristics are *large domain shifts* and *severe data imbalance*, which give rise to poor and inaccurate performances when addressing complicated and intricate VQA tasks in the surgical field.

Domain shift is the first problem. Conventionally, a CL model is updated using the teacher-student framework. This framework involves training the latest model with the guidance of a teacher model trained on previous data. The newly developed model can be viewed as the student. During the learning phase, the student model leverages the teacher model's logits, which helps it capture the teacher's knowledge without needing access to prior data [15]. However, if the teacher encounters a completely foreign task within the new data in the training phase, it can only offer an arbitrary deduction as an estimate to guide the student [16]. This phenomenon is also known as domain shift. This issue is particularly common in surgical applications where surgical scenes from different types of surgeries, even within distinct and specialized categories, can vary greatly in appearance, due to variations in surgical tools, medical procedures, and techniques. For instance, data from a surgical procedure on the liver can be significantly different from that of a kidney surgery. This problem could be exacerbated by data collected from various sources, given the differences in surgical procedure guidelines and clinical systems applied. Significant and detrimental domain shifts are common in surgical datasets and can considerably impair the precision and effectiveness of the CL model, as the student model ends up learning only the teacher's uneducated deductions.

Data imbalance is also an important issue. In real-world surgical procedures, certain actions or instruments are encountered less frequently. For instance, tissue manipulation is commonly observed in many surgical datasets since most surgeries involve interaction with tissues. Cutting actions are also prevalent, especially in datasets centered around nephrectomy. However, within the same nephrectomy dataset, stapling is seldom observed. This is because the stapling action typically takes place after vessel severance, a specialized technique within the general nephrectomy procedure. When some classes in the training data have substantially fewer instances than others, this situation is referred to as data imbalance [17]. Previous CL algorithms have not specifically addressed the challenge of imbalanced data, thus resulting in inadequate training for underrepresented classes.

In order to address the knowledge constraints arising from the two above-mentioned characteristics, we harnessed the robust generalization capabilities of multimodal large language models (LLMs). LLMs are trained with extensive datasets of images and/or text from a variety of domains [18]–[21]. The model's powerful capability to answer questions spanning a multitude of domains has garnered significant research interest in different fields. Recent investigations have observed the proficiency of state-of-the-art LLMs in answering medical queries [22]. These observations encourage the use of LLMs' responses to bridge the knowledge gaps whenever data used for model training includes new knowledge for the teacher model or is highly imbalanced. The straightforward rationale stems from situations where the information encountered exceeds the teacher's capabilities and scope; utilizing LLMs' vast medical understanding is considerably more efficient than trusting the teacher model's incomplete training (caused by the imbalanced data) or random guesses (caused by domain shifts).

In addition to the LLM-aided multi-teacher CL framework, we also introduce an adaptive weight assignment method to balance insights from multiple teachers, which include the LLM teacher and conventional CL teachers obtained in previous training phases under a CL setup. This mechanism aims to leverage the general medical insights and generalization ability of the LLM, as well as the domain expertise of previous CL models, to help the student model learn problem-solving abilities from different perspectives. The implementation of adaptive weights enables ideal model training: it draws more from a specific CL teacher model when the knowledge is from a well-understood and specifically trained domain with ample data; otherwise, it relies heavily on the LLM.

Moreover, this paper introduces a new surgical VQA dataset to validate our scheme in practical surgical environments. We devised a new QA pair generation method based on GPT-3.5 for constructing VQA datasets. We used in-context learning (ICL) [23]–[25] to enhance the analysis of text descriptions related to clinical images.

Our contributions are summarized as follows:

1) We proposed an LLM-assisted multi-teacher CL framework, termed LMT++, and established an innovative approach to extract logits from the investigated LLM, offering vital guidance for future works to incorporate LLMs under a CL setup.
2) We also developed an adaptive weighting scheme. The strategy effectively leverages the LLM's strong generalization ability and conventional CL teachers' specialized domain knowledge, thereby significantly enhancing the student model's training.
3) One new surgical VQA dataset is open-sourced, providing a valuable resource for future research. Additionally, the creation of the VQA dataset highlights a new approach for generating QA pairs using ICL.

A preliminary version of this work was presented in IEEE

ICRA 2024 [26]. In this paper, we have substantially revised and extended the original version. First, we advanced our model training process by enhancing the "single CL teacher plus an LLM" strategy with the "multiple CL teachers plus an LLM" strategy so that the catastrophic forgetting problem can be even more effectively addressed. For easier identification, we refer to the "single CL teacher" scheme in our ICRA paper as LMT, and we refer to the method presented in this paper as LMT++. Furthermore, we constructed a new and more challenging surgical VQA dataset that differs significantly from the previously developed dataset (including the one we proposed in [26]). We also visualized the diversity of the new dataset by comparing it with previous ones in terms of organ, procedure, robotic surgical system, and instrument. Our code is available at https://github.com/yuyangdu01/LLM-CL-VQA.

II. Proposed Method

A. Preliminaries

Problem Formulation and Notations: Consider a CL process with τ time periods, where $t \in \{1, ..., \tau\}$ represents a specific time period of the process. The training dataset at time t is denoted by D_t, with each element $d_{t,i} \in D_t$ representing the i^{th} training sample for time t. Each training sample $d_{t,i}$ consists of a frame in the surgical setting and a variety of corresponding clinical questions. The classes present in D_t are denoted by C_t, where element $c_{t,j}$ denotes the j^{th} class appearing in D_t. If class $c_{t,j}$ constantly appears in D_t but was absent in the earlier datasets used during model training (i.e., D_{t-1}, D_{t-2},..., D_1), we say a domain shift has occurred for class $c_{t,j}$. Additionally, if class $c_{t,j}$ appears significantly less frequently compared to other classes in D_t, then we can assume that data imbalance has occurred for class $c_{t,j}$.

Distillation Loss in Continual Learning: Knowledge distillation (KD) is an effective approach to enhance knowledge retention from prior models and mitigate the catastrophic forgetting issue without needing to revisit data used in prior training phases [27]. There are three types of KD: response-based KD, feature-based KD, and relation-based KD. This paper will focus on response-based KD. Due to the versatility and robustness of response-based KD, student and teacher models can be used alongside various network architectures. [28].

The expression below evaluates the distillation loss of the KD process:

$$L_{KD} = L_{CE}\left\langle \sigma\left(z^T/\delta\right), \sigma\left(z^S/\delta\right)\right\rangle \tag{1}$$

where $L_{CE}\langle\cdot,\cdot\rangle$ represents the cross-entropy loss; $\sigma(\cdot)$ denotes the softmax function; z^T and z^S refer to the output logits of the teacher and the student models, respectively; δ is a temperature hyperparameter that adjusts the smoothness of the probability distributions. When $\delta = 1$, the function corresponds to the standard softmax function, and as δ increases, the resulting probability distribution becomes softer, revealing additional information such as the similarity between the predicted class and other classes.

B. Multi-teacher CL Framework with LLM

In order to address the challenges arising from data imbalances and domain shifts, we incorporate a supplementary multimodal LLM teacher, known for its exceptional adaptability and generalization capability, to facilitate improved knowledge transfer. When previously obtained teacher models encounter knowledge that they are unfamiliar with, the LLM teacher would assist the student in learning from a more relevant and suitable source of knowledge.

The multi-teacher CL approach used in this paper is illustrated in Fig. 1. The general loss function L is depicted in the equation below:

$$L = \alpha L_0 + \sum_{i=1}^{t-1} \beta_i L_{KD}^i + \chi L_{KD}^{LLM} \tag{2}$$

where the hard labels governed the cross-entropy loss L_0; the KD loss between the new CL model trained at time t (i.e., the student model) and the previous CL model trained at time i is denoted as L_{KD}^i, in which $i \in \{1, 2, \ldots, t-1\}$ denotes the multiple conventional teacher models; L_{KD}^{LLM} is the KD loss between the student model and the LLM teacher; α, β_i, and χ are normalized adaptive weights of the L_0, L_{KD}^i, and L_{KD}^{LLM}, respectively. The sum of α, β_i, and χ is one. We refer readers to Section II-C for details about the weight assignment scheme.

Throughout this work, we denote the logits of the student model as z^t, the logits of the old teacher models as z^i, and the logits of the LLM teacher model as z^{LLM}. In addition, we write L_{KD}^i and L_{KD}^{LLM} as

$$L_{KD}^i = L_{CE}\left\langle \sigma\left(z^i/\delta\right), \sigma\left(z^t/\delta\right)\right\rangle \tag{3}$$

and

$$L_{KD}^{LLM} = L_{CE}\left\langle \sigma\left(z^{LLM}/\delta\right), \sigma\left(z^t/\delta\right)\right\rangle \tag{4}$$

Given that the LLM teacher is implemented within an intricate and complicated transformer network, which differs considerably from both traditional teacher and student models, several transformation steps are required to derive logits from the embeddings. Information regarding the chosen LLM and the transformation from embedding to logits is provided in the following section.

In this paper, an open-access multimodal LLM featuring both visual and linguistic capabilities, known as InstructBLIP [29], was selected as the LLM teacher. InstructBLIP consists of three elements: an image encoder to process the input image, a text-in-text-out LLM that handles the output, and an image-text transformer that connects both modules. Due to its modular design, InstructBLIP offers high flexibility and versatility, allowing us to effectively and efficiently utilize a diverse array of text-to-text LLMs. Additionally, we selected FlanT5 [30], an instruction-tuned model derived from Transformer T5, as the text-in-text-out LLM to ensure the generalization and adaptive capabilities of our implemented model.

The embeddings derived from FlanT5's last fully connected layer are used to form a self-attention matrix, referred to as e_{LLM}. The dimension of e_{LLM} is $N \times (M+1) \times P$, where the number of classes at time t is depicted as N, i.e., the

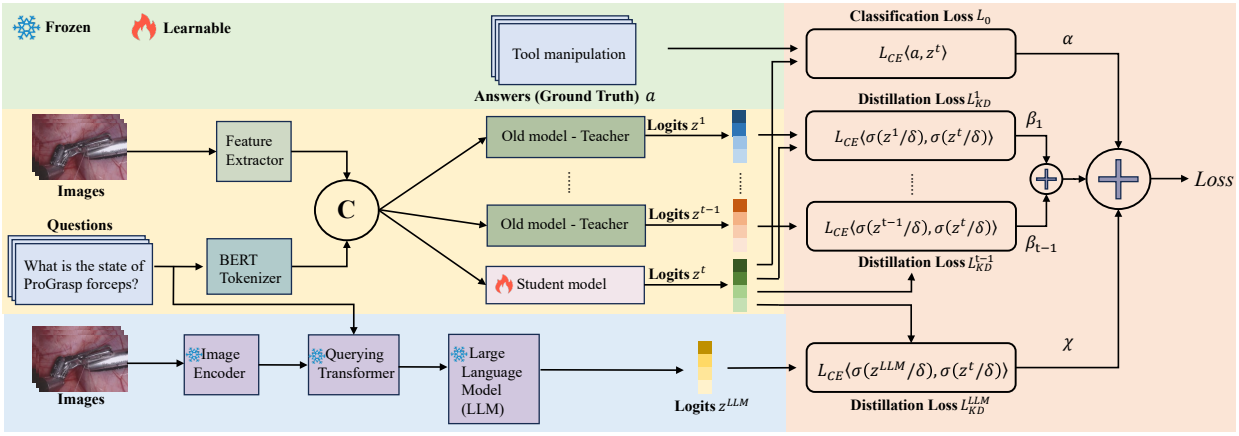


Fig. 1: The proposed multi-teacher CL framework. This system processes bimodal inputs (text and image) to generate deductions for the corresponding VQA task. Our weight adaption scheme (highlighted in the light orange zone) is formulated to balance the general information and knowledge provided by the LLM with the synthesized surgical and medical proficiency of multiple teacher models. The light-blue highlighted region illustrates the frozen LLM while the light-yellow region symbolizes the several conventional teacher models and the student model. Finally, the light-green zone represents the ground truth.

cardinality of C_t; P corresponds to the LLM's vocabulary; and M indicates the number of tokens we used to represent each class label.[1] As indicated in (4), the desired logits z^{LLM} should have a size of $N \times 1$.

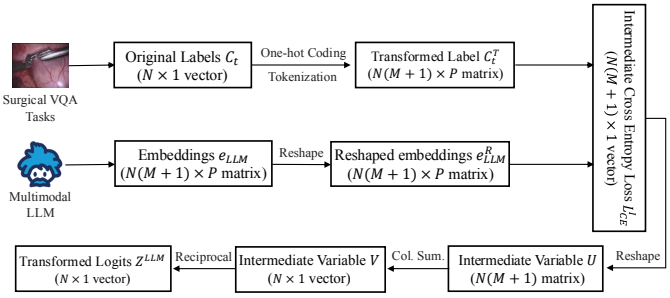


Fig. 2: The workflow logits transformation after extracting the embeddings.

Details on the transformation of $N \times (M+1) \times P$ embeddings into the $N \times 1$ logits are provided as follows. As depicted in Fig. 2, the extracted embeddings are reshaped to form a new matrix e_{LLM}^R. Subsequently, one-hot encoding and tokenization are applied to the classification label set C^t. It is important to note that both e_{LLM}^R and C^T have dimensions $N(M+1) \times P$. Next, the cross-entropy loss between e_{LLM}^R and C^T is calculated by

$$L_{CE}^I(i) = L_{CE}\left\langle e_{LLM}^R(i), C^T(i)\right\rangle \tag{5}$$

The resulting vector L_{CE}^I has dimensions of $N(M+1)\times 1$. To achieve the corresponding cross entropy for each label, we first reshape this vector to an $N \times (M+1)$ matrix and perform column summation. In the resulting $N \times 1$ vector V, the predicted loss of label $c_{t,j}$, depicted as element $V(j)$, is inversely proportionate to the likelihood of label $c_{t,j}$ selected as the eventual classification outcome. In contrast, in the desired logits vector z^{LLM}, element $z^{LLM}(j)$ ought to be directly associated with the classification probability of $c_{t,j}$ so that the output of the softmax layer, denoted as $\sigma\left(z^{LLM}/\delta\right)$, can formulate a probability vector. Hence, an additional transformation of vector V is required to ensure that its elements are directly proportional to the classification probability. A potential approach could be to invert the elements of V, i.e., $z^{LLM}(j) = 1/V(j)$.

C. Adaptive Weight Assignment

In (2), the weights of L_0, L_{KD}^i, and L_{KD}^{LLM} are represented as α, β_i, and χ, respectively. In this paper, we dynamically adjust χ and β (i.e., the total value of β_i) during the model training using the weight assignment method detailed in subsequent paragraphs while treating α as hyperparameters. And we have β_i shares β equally, i.e., we have

$$\beta_i = \beta/(t-1) \tag{6}$$

The extent of domain shift in a surgical dataset can be measured while training the model at time t by analyzing the a1verage accuracy of the previous models from 1 to D_t on the dataset utilized in the training phase. If the previous CL models perform well and attain high precision, it suggests that the models are familiar with the knowledge in D_t. Conversely, if they exhibit inferior performance and yield poor accuracy on D_t, especially relative to the LLM teacher, it suggests that the prior CL models lack adequate expertise for this training iteration. Based on the above reasoning, β and χ are adjusted according to the accuracy of the LLM and the average accuracy of the multiple CL models on D_t. A larger difference in accuracy indicates a considerable and critical domain shift. Therefore, it is ideal that a higher weight is allocated to the LLM to harness its broad knowledge base and immense generalization ability in the medical and surgical domain.

In addition to addressing domain shift, this approach also aims to mitigate the presence of data imbalance within the

[1]The second dimension of the matrix is $M+1$ rather than M because we add one additional pause token for each word to indicate the end of a word.

surgical dataset through the application of adaptive weighting. We observe that the distribution of labels within the dataset involved in the test phase was uneven, with some surgical procedures and medical tools and instruments being regularly referenced, while others are seldom exemplified. This can be illustrated using the Nephrectomy dataset example highlighted in the earlier section. In a Nephrectomy-related surgical dataset, cutting actions are commonly referenced, whereas stapling is seldom mentioned due to its specialized role in the procedure, occurring only after vessel cutting. Even if the prior CL models possessed a certain amount of knowledge and information regarding these less frequently mentioned areas, their expertise is limited due to the shortage of training data. In these scenarios, a higher weight is assigned to the LLM teacher to utilize its broad medical insights to compensate for and bridge the knowledge gaps.

It is clear that both domain shifts and data imbalances influence the assignments of β and χ. As such, β and χ are expressed as

$$\beta = \theta_{DS}\beta_{DS} + \theta_{DI}\beta_{DI} \tag{7}$$

and

$$\chi = \theta_{DS}\chi_{DS} + \theta_{DI}\chi_{DI} \tag{8}$$

where hyperparameters θ_{DS} and θ_{DI} are assumed to satisfy $\theta_{DS} + \theta_{DI} = 1 - \alpha$, reflecting the importance of domain shift (DS) and data imbalance (DI) in our model training, respectively. If the focus is more on domain shift, θ_{DS} is increased while θ_{DI} is decreased. Conversely, if the focus is on data imbalance, the opposite is done. Furthermore, β_{DS} and χ_{DS} in (7) and (8) denote the weight share of the old CL teachers and the LLM teacher with respect to domain shifts, respectively; β_{DI} and χ_{DI} denote the weight share of the two types of teachers in relation to data imbalance. To ensure that $\beta + \chi = 1 - \alpha$, we need to verify that

$$\beta_{DS} + \chi_{DS} = 1 \tag{9}$$

and

$$\beta_{DI} + \chi_{DI} = 1 \tag{10}$$

We then elaborate on how β_{DS} and χ_{DS} are assigned. It is known that β_{DS} and χ_{DS} are determined by the average accuracyk

of the previous CL teachers and the accuracy of the LLM teacher on D_t. Given D_t, the average classification accuracy of the CL models from time 1 to $t-1$ and that of the LLM are represented as acc_{avg} and acc_{LLM}, respectively. To satisfy the constraints in (9) and (10), we have

$$\beta_{DS} = \frac{acc_{avg}}{acc_{avg} + acc_{LLM}} \tag{11}$$

and

$$\chi_{DS} = \frac{acc_{LLM}}{acc_{avg} + acc_{LLM}} \tag{12}$$

The issue regarding domain shift is effectively rectified using the assignment scheme in (11) and (12), as the LLM will receive a larger weight when the old CL models are generating arbitrary deductions.

Next, we discuss how β_{DI} and χ_{DI} are determined. At time t, the set of previously encountered data is denoted by $D_{t,\ldots,1}$ (i.e., $D_{t,\ldots,1} = D_t \cup D_{t-1} \ldots \cup D_1$). The k^{th} class labels in $D_{t,\ldots,1}$ are denoted by c_k, and the occurrence count of c_k is represented by d_k. To assess the extent of data imbalance in $D_{t,\ldots,1}$, we propose the concept of imbalance ratio as in [31], which is defined as:

$$IR = \max(d_k)/\min(d_k) \tag{13}$$

It is important to note that d_k in (13) cannot be zero given the definition provided.

Under the constraint in (9) and (10), we have

$$\beta_{DI} = \frac{1}{1 + \log_N IR} \tag{14}$$

and

$$\chi_{DI} = \frac{\log_N IR}{1 + \log_N IR} \tag{15}$$

where N is a hyperparameter.

From (14) and (15), it is observed that whenever the data imbalance is sever, χ_{DI} becomes large, allowing the LLM's general domain knowledge to be effectively utilized to address the knowledge gap. This approach helps to alleviate the data imbalance issue.

Finally, we give the expression of β and χ as follows:

$$\beta = \theta_{DS}\frac{acc_{avg}}{acc_{avg} + acc_{LLM}} + \theta_{DI}\frac{1}{1 + \log_N IR} \tag{16}$$

and

$$\chi = \theta_{DS}\frac{acc_{LLM}}{acc_{avg} + acc_{LLM}} + \theta_{DI}\frac{\log_N IR}{1 + \log_N IR} \tag{17}$$

Substituting (16) into (6), we have

$$\beta_i = \frac{\theta_{DS}}{t-1}\frac{acc_{avg}}{acc_{avg} + acc_{LLM}} + \frac{\theta_{DI}}{t-1}\frac{1}{1 + \log_N IR} \tag{18}$$

III. Experiments and Analysis

A. Existing Dataset

EndoVis17 comprises frames obtained from various recorded robotic abdominal surgeries originating from a public challenge dataset [32]. The curated QA pairs in this dataset contain single-word answers, which are classified into surgical maneuvers or instrument placement. A total of 5 videos were involved in the experiments, resulting in 73 frames and 376 QA pairs for the training set, and 24 frames with 96 corresponding QA pairs for the test set.

EndoVis18 also comprises frames obtained from various recorded robotic abdominal surgeries originating from a public challenge dataset [1]. In addition to the movement-related and placement-related questions in EndoVis17, EndoVis18 also includes questions about human biological organs. This introduces a domain shift at $t = 2$. Furthermore, EndoVis18 covers 5 more action classes (i.e., clipping, looping, staple, suction, and suturing), which considerably contribute to data imbalance. The dataset consists of 14 videos. The training dataset training consists of 1560 frames with 9014 QA pairs, while the test set comprises 447 frames with 2769 QA pairs.

DAISI-VQA is the VQA dataset we created in our earlier ICRA conference paper [26]. The development of DAISI-VQA

TABLE I: Data diversity. EV17, EV18, D-V, and L-V represent EndoVis17, EndoVis18, DAISI-VQA, and LRSP-VQA datasets. Number_T and Number_N denote the number of total types and new types, respectively.

	EV17	EV18	D-V	L-V
Organ	kidney	kidney	Number_T = 25, Number_N = 24	Number_T = 10, Number_N = 9
Procedure	nephrectomy	nephrectomy	uncountable	Number_T = 11, Number_N = 10
Surgical Assisting System	Da-Vinci XI	Da-Vinci X, Da-Vinci XI	No surgical assisting system involved, all operations are manually conducted	Da-Vinci X, Da-Vinci XI, Da-Vinci SI, Da-Vinci SP, Hugo RAS
Instrument	Number_T: 5	Number_T = 8, Number_N = 3	Number_T = 32, Number_N = 24	Number_T = 18, Number_N = 10

is based on the DAISI dataset reported in [33]. Image frames and instructional texts for various surgical procedures on different organs are featured in the original DAISI dataset, where each procedure is illustrated by several images accompanied by relevant texts.

For generating QA pairs, the original DAISI dataset is initially refined by removing irrelevant frames and images (such as those without any surgical content) and unnecessary descriptions (such as those detailing the hospitals, medical centers, or surgeons). Subsequently, QA pairs were generated based on the text description associated with each image. To produce appropriate questions and precise answers for each image, this paper introduces a new data creation methodology. This methodology processes the textual depictions and narrations associated with the respective images using GPT-3.5 before implementing an advanced few-shot learning method customized for LLMs called in-context learning (ICL) [24]. Specifically, we provided GPT-3.5 with a prompt that encompasses different reference QA pair examples to illustrate the formulation of surgical questions and demonstrate how answers could be derived from the related text-based descriptions. After presenting these example QA pairs, the detailed description of a new, unaddressed DAISI image was attached to the prompt, and GPT-3.5 was then instructed to generate QA pairs based on the textual description provided. Leveraging its robust analytical and emulation capabilities, GPT-3.5 analyzed the examples, understood our task requirements, and generated suitable and sensible QA pairs for the provided description.

The DAISI-VQA dataset includes 353 surgical images and 545 QA pairs. We allocated approximately 80% of this data to the training set, leaving the remainder for testing.

B. New Dataset Construction

LRSP-VQA is a new dataset constructed in this paper. It refers to the Live Recorded Surgical Procedures VQA dataset we built with 36 video demonstrations about robotic-assisted surgical operations. These videos were collected from YouTube and were selected based on their in-video narration of the surgical procedures and the quality of the recording. We segmented the videos into 150 shorter parts, with each video snippet corresponding to one surgical phase within the whole procedure. The primary objective of this operation was to guarantee that each segment contained the same set of surgical instruments throughout. Subsequently, we extracted frames from each video segment and obtained over 10,000 image frames from the 150 video segments. In the third step, we filtered out relevant portions of the video transcriptions corresponding to each specific video segment. These textual transcriptions contained the surgical instructions and explanations that were provided audibly in the video and could therefore serve as valuable descriptions for these selected surgical videos. With all the images and textual descriptions obtained above, we leveraged GPT-3.5 for QA pair generation in a way similar to the QA pair generation in DAISI-VQA.

In LRSP-VQA, there are 1,136 QA pairs, and each QA pair has an associated surgical image. We allocated an estimated 80% of the data for training and utilized the rest for testing.

After a brief introduction to each dataset, we present Table I, which provides a detailed comparison of the four datasets in terms of the organ(s) and procedure(s) involved, surgical assisting system(s) applied, and surgical instruments used. This comparison can help readers better understand the significant data diversity we introduced in the new dataset, which is crucial for validating our model's performance under severe domain shifts and data imbalances. The comparison results are as follows:

1) **Organ**: The two datasets we created show higher diversity in terms of organs involved: surgical QA pairs in DAISI-VQA and LRSP-VQA involve 25 and 10 different organs, respectively, while previous datasets (i.e., EndoVis17 and EndoVis18) focus on the kidney only.
2) **Surgical procedure**: EndoVis17 and EndoVis18 only consider nephrectomy, while LRSP-VQA investigates nine additional clinical procedures. For DAISI-VQA, detailed information about the number of clinical procedures is hard to obtain, as the construction of the DAISI dataset has filtered out information about detailed clinical procedures used. However, we note that the number of clinical procedures in DAISI-VQA at least exceeds that of LRSP-VQA, given the large number of organs investigated.
3) **Surgical assisting system**: Many surgical VQA datasets are built with videos recorded by surgical robots. For example, the surgical operation in EndoVis17 is completed by the Da Vinci X robotic surgical system [34], while the later-released EndoVis18 applies a more advanced Da Vinci Xi platform [35]. For the DAISI-VQA dataset, all surgical operations are manually conducted, which also differs significantly from EndoVis17 and EndoVis18. For our datasets, LRSP-VQA introduces two of the

TABLE II: Benchmarking experiments - Accuracy.

	Accuracy ($t = 1$ to $t = 2$)			Accuracy ($t = 2$ to $t = 3$)				Accuracy ($t = 3$ to $t = 4$)				
	EV17	EV18	Avg.	EV17	EV18	D-V	Avg.	EV17	EV18	D-V	L-V	Avg.
FT	0.2917	0.5905	0.4411	0.0938	0.3286	0.7632	0.3952	0.1250	0.1123	0.4561	0.6782	0.3429
ER	0.5417	0.5782	0.5599	0.5313	0.6071	0.7544	0.6309	0.5313	0.5544	0.7018	0.6552	0.6106
LwF	0.4479	0.5309	0.4894	0.5104	0.4745	0.6754	0.5535	0.0938	0.1322	0.3421	0.5632	0.2828
Online-EWC	0.4167	0.5002	0.4584	0.0625	0.3611	0.7368	0.3868	0.1250	0.1116	0.7018	0.6609	0.3998
EWC++	0.4792	0.4680	0.4736	0.0938	0.3734	0.7105	0.3926	0.1250	0.1098	0.5877	0.6552	0.3692
LMT++	**0.5104**	**0.5619**	**0.5362**	**0.5313**	**0.5056**	**0.7456**	**0.5942**	**0.1979**	**0.1546**	**0.6842**	**0.5747**	**0.4029**

TABLE III: Benchmarking experiments - F-score.

	Accuracy ($t = 1$ to $t = 2$)			Accuracy ($t = 2$ to $t = 3$)				Accuracy ($t = 3$ to $t = 4$)				
	EV17	EV18	Avg.	EV17	EV18	D-V	Avg.	EV17	EV18	D-V	L-V	Avg.
FT	0.1843	0.3806	0.2825	0.0327	0.0982	0.8751	0.3353	0.0436	0.0425	0.2088	0.5034	0.1996
ER	0.3344	0.3681	0.3512	0.2784	0.3792	0.8721	0.5099	0.3048	0.2740	0.8440	0.4772	0.4750
LwF	0.3034	0.2966	0.3000	0.2367	0.1708	0.3945	0.2673	0.0286	0.0274	0.1250	0.4472	0.1570
Online-EWC	0.2276	0.2012	0.2144	0.0362	0.1316	0.8648	0.3442	0.0400	0.0427	0.2400	0.4956	0.2046
EWC++	0.2229	0.2624	0.2427	0.0532	0.1922	0.7293	0.3249	0.0400	0.0339	0.2059	0.4816	0.1903
LMT++	**0.3091**	**0.3185**	**0.3138**	**0.2851**	**0.1862**	**0.8692**	**0.4468**	**0.0810**	**0.0528**	**0.3329**	**0.4512**	**0.2295**

TABLE IV: Ablation study - Accuracy.

	Accuracy ($t = 1$ to $t = 2$)			Accuracy ($t = 2$ to $t = 3$)				Accuracy ($t = 3$ to $t = 4$)				
	EV17	EV18	Avg.	EV17	EV18	D-V	Avg.	EV17	EV18	D-V	L-V	Avg.
Scenario 1	0.4271	0.5677	0.4974	0.3333	0.3398	0.7544	0.4759	0.1563	0.1322	0.5702	0.6149	0.3684
Scenario 2	0.4063	0.5702	0.4882	0.5208	0.4810	0.7456	0.5825	0.1042	0.1008	0.7193	0.6207	0.3862
Scenario 3	0.4167	0.5670	0.4918	0.5104	0.4464	0.7368	0.5645	0.1042	0.1553	0.6667	0.6379	0.3910
Scenario 4	0.4479	0.5309	0.4894	0.5104	0.4745	0.6754	0.5535	0.0938	0.1322	0.3421	0.5632	0.2828
LMT++	**0.5104**	**0.5619**	**0.5362**	**0.5313**	**0.5056**	**0.7456**	**0.5942**	**0.1979**	**0.1546**	**0.6842**	**0.5747**	**0.4029**

TABLE V: Ablation study - F-score.

	Accuracy ($t = 1$ to $t = 2$)			Accuracy ($t = 2$ to $t = 3$)				Accuracy ($t = 3$ to $t = 4$)				
	EV17	EV18	Avg.	EV17	EV18	D-V	Avg.	EV17	EV18	D-V	L-V	Avg.
Scenario 1	0.2336	0.2365	0.2351	0.1218	0.1035	0.8736	0.3663	0.0688	0.0539	0.2291	0.4983	0.2125
Scenario 2	0.2595	0.3101	0.2848	0.2195	0.1933	0.8678	0.4269	0.0515	0.0257	0.2937	0.4876	0.2147
Scenario 3	0.2658	0.2883	0.2770	0.2151	0.2090	0.8648	0.4296	0.0256	0.0231	0.2313	0.5740	0.2135
Scenario 4	0.3034	0.2966	0.3000	0.2367	0.1708	0.3945	0.2673	0.0286	0.0274	0.1250	0.4472	0.1570
LMT++	**0.3091**	**0.3185**	**0.3138**	**0.2851**	**0.1862**	**0.8692**	**0.4468**	**0.0810**	**0.0528**	**0.3329**	**0.4512**	**0.2295**

latest robotic surgical systems in the Da Vinci series (i.e., Da Vinci Si [36] and Da Vinci SP [37]) and a robotic surgical system from a new series (the Hugo RAS platform [38]).

4) **Instruments**: The latest two datasets show much higher diversity in terms of instruments involved: QA pairs in DAISI-VQA and LRSP-VQA using 32 and 18 different instruments, respectively, while previous datasets involve fewer instruments, with EndoVis17 and EndoVis18 discussing 5 and 8 surgical instruments, respectively.

From the above discussion, it is evident that our datasets are significantly different from those developed previously, and an obvious domain shift can be introduced when these two datasets are applied at $t = 3$ and $t = 4$.

C. Implementation Details of Our Method and Baselines

We assessed our method LMT++ against the following algorithms in a CL setting:

1) **Fine Tune (FT)**: FT is a core approach for adaptation in CL. It involves introducing new data to update a model that has already been trained on prior tasks. While FT is capable of quickly adapting to new tasks, it is prone to catastrophic forgetting and often results in the poorest mean accuracy and performance in a CL setting [39].
2) **Experience Replay (ER)**: Through the retention of the original and prior training data in a memory buffer, ER can alleviate and prevent catastrophic forgetting. The model revisits examples from this buffer when training on new tasks, thereby retaining knowledge of the old tasks [40]. Although ER performs well in a CL process, it may not always be feasible in medical applications due to ethical concerns arising from the storage and retention of patient data. In this study, we use ER only as an upper-performance benchmark.
3) **Learning without Forgetting (LwF)**: By maintaining past knowledge in a distilled model, LwF ensures an efficient CL process, enabling the learning of new material without forgetting previously acquired information.

LwF is a renowned CL algorithm that uses knowledge distillation (KD) as part of the training procedure. [11].

4) **Elastic Weight Consolidation++ (EWC++)**: With automated weight importance adjustments and simplified Fisher information matrix calculations, EWC++ enhances the original EWC algorithm and offers greater scalability and flexibility. [41]
5) **Online Elastic Weight Consolidation (Online-EWC)**: Online-EWC is another enhanced variation of the original EWC algorithm. It improves the efficiency of updating critical model parameters, helping to safeguard prior information and knowledge when fresh insights and new data are introduced [42].

The implementation details of our method are as follows. The model is first trained on EndoVis17 at $t = 1$ with a learning rate of 5×10^{-6} and for 60 epochs, as outlined in [32]. At $t = 2$, the model is then trained on EndoVis18 with a learning rate of 5×10^{-5} and for 80 epochs, according to the settings in [1]. Subsequently, for $t = 3$, the model is trained on DAISI-VQA with a learning rate of 5×10^{-6} and for 80 epochs. Ultimately, at $t = 4$, the model is finally trained on LRSP-VQA, using a learning rate of 1×10^{-5} and for 80 epochs. The mentioned algorithms above are trained using an NVIDIA Tesla T4 GPU and implemented using PyTorch. The Adam optimizer is employed consistently across all experiments.

D. Experimental Results and Associated Discussions

Prior to the result analysis, it is important to highlight that the evaluation approach employed in this paper is not based on the model's performance on a single dataset. Rather than focusing on specific instances, we prioritize the model's *average* performance, particularly in terms of accuracy and F-score, across the different datasets evaluated at various time points. This measurement approach is essential because a CL model that struggles with catastrophic forgetting may achieve satisfactory performance on the latest dataset while failing on those it encountered in the past. A robust and adaptable CL model is capable of maintaining a strong performance across all datasets. While it may not always outperform a weaker model on a single dataset, it typically delivers better average results and performances overall. Therefore, evaluating the model's average performance across all the tested datasets is a standard practice in CL research [1], [9], [10], [32], and this paper adopts the same approach.

We first analyze the benchmarking results presented in Tables II and III. As anticipated, FT exhibits the most severe catastrophic forgetting among all tested methods, resulting in the poorest performance. ER, which represents the ideal upper bound, shows the strongest result. Our method consistently outperforms all other tested schemes throughout the process. Compared to previous approaches, our method improves model accuracy by an average of 4.54% over the second-best model from $t = 2$ to $t = 4$. Additionally, we averagely enhanced the F-score of the second-best model by an average of 15.27%.

The results highlighted our approach's exceptional capability to learn new information without losing previously acquired knowledge. We attribute the model's strong performance to two key factors: 1) the incorporation of LLM during the training phase, which effectively addresses the challenges brought about by domain shifts and data imbalances, and 2) the dynamic weight adjustment scheme, which strikes a balance between expert teacher models and the LLM teacher.

To further validate the effectiveness of each component in our proposed method, we conducted an ablation study. The following scenarios were taken into consideration:

1) When assigning weights, set aside data imbalance and prioritize domain shift. (i.e., $\theta_{DI} = 0$ and $\theta_{DS} = 1$).
2) When assigning weights, set aside domain shift and concentrate exclusively on data imbalance (i.e., $\theta_{DI} = 1$ and $\theta_{DS} = 0$).
3) Remove the adaptive weight assignment mechanism altogether and apply a fixed weight throughout the entire CL process.
4) Remove the LLM teacher and use previously obtained models as conventional teachers in the CL process.

The results of the ablation study, shown in Tables IV and V, demonstrate that our method outperforms its variations in Scenarios 1/2/3/4 at $t = 2$, $t = 3$, and $t = 4$. Fig. 3 (a) provides an example to illustrate the ablation experiments. In Scenarios 1/2/3/4, the tested model misidentifies the number of surgical instruments as 3, generating the incorrect answer "Yes". However, with all components integrated, our model successfully provides the correct answer "No". And Fig. 3 (b) illustrates the accuracy of tool state identification for the needle driver. The tested model incorrectly predicts the state as "Idle" while our method correctly identifies as "Tool Manipulation". The results emphasize that each proposed component is crucial to the final performance, proving their indispensability in our methodology.

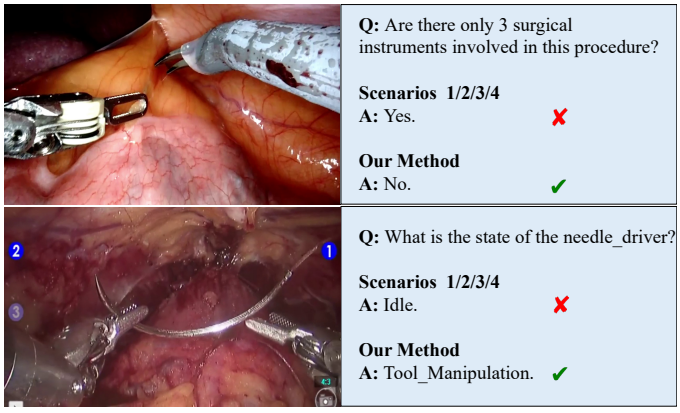


Fig. 3: Two samples in LRSP-VQA testing dataset, where only our method successfully identifies the (a) top: instrument counting, and (b) bottom: tool state during a surgical operation.

IV. How LLM Works: Case Study and Discussion

Following the above experimental results, we now discuss a problem the reader may be interested in: why integrating InstructBLIP as a supplementary teacher enhances the student model's performance in the VQA task. This question becomes

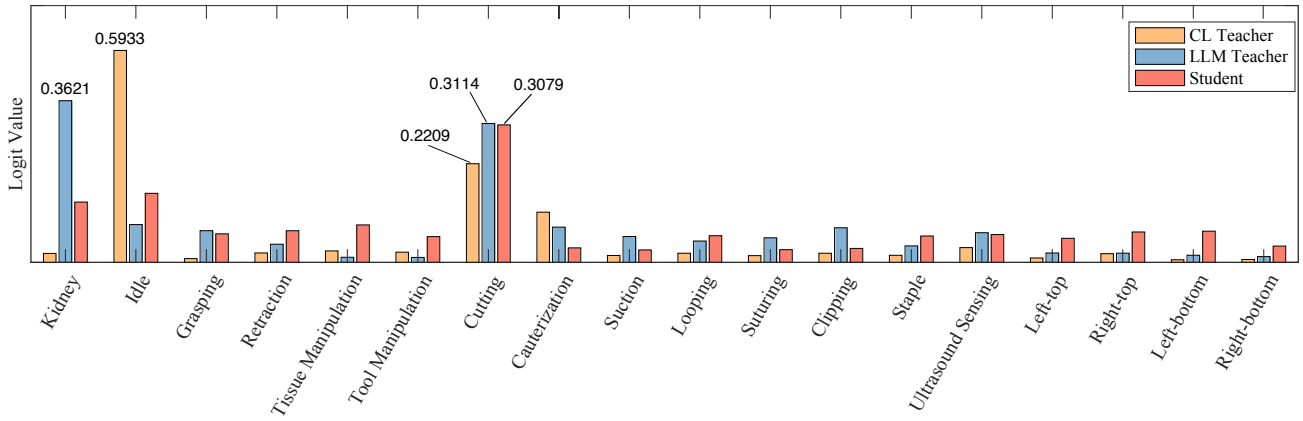
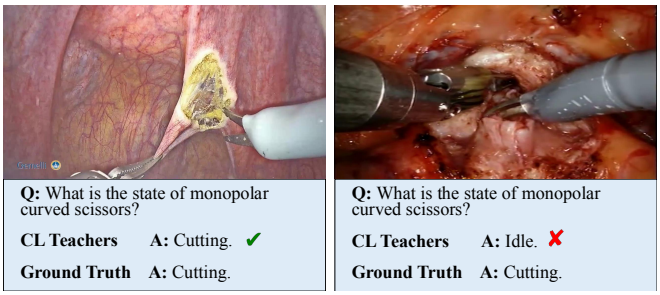


Fig. 6: Logits of CL teacher, LLM teacher, and the student for the case studied in Fig. 5.

even more interesting after we have noticed that the multi-modal LLM itself may not always have perfect performance when addressing the surgical VQA task independently. In fact, through the experiments in this paper, we observed that InstructBLIP sometimes defaulted to answering questions with "kidney" (see our later examples). This outcome is understandable, given that InstructBLIP was trained on a broad spectrum of general domain datasets, making it more accustomed to commonly used words (such as "kidney") rather than specific medical terminologies. However, it is crucial to emphasize that even if the LLM does not perform perfectly independently on the surgical task, it still makes valuable contributions within our multi-teacher CL framework thanks to the information embedded in the LLM's logits. In particular, the LLM's logits can work with traditional CL teachers under the proposed adaptive weighting scheme. The following case study illustrates how LLM logits aid the CL training process.



(a) (b)

Fig. 4: Two LRSP-VQA images that help to illustrate the conventional CL teachers' confusion between cutting and idle.

To begin with, we point out a typical confusion that conventional CL teachers may have between "cutting" and "idle" with the illustration of the two LRSP-VQA images below. A "cutting" operation usually involves the use of surgical scissors. When the scissors are fully opened (see Fig. 4a), the CL teachers can precisely identify the operation as "cutting". However, when the surgical scissor is slightly or partially opened (see Fig. 4b), the CL teachers may mistakenly classify the "cutting" operation as "idle". This confusion is reasonable, as a slightly opened surgical scissor during "cutting" may

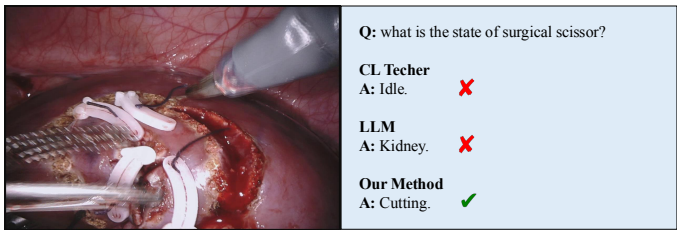


Fig. 5: An image illustrating the "cutting" operation on a kidney. The joint effort of the CL teacher and the LLM teacher results in the correct knowledge distillation to the student.

appear similar to an "idle" one.

We then take a close look at the LLM's logits in order to understand how the LLM assists conventional CL teachers in accurate categorization and effective knowledge distillation. Here we take an EndoVis18 image at $t = 2$ for example, which simplifies the analysis by avoiding introducing multiple CL teachers into the analysis (our experiment here focuses on the LLM's logits). Fig. 5 presents a "cutting" operation on a kidney, where the conventional CL teacher allocated its largest logits value to "idle" and the second biggest value to "cutting" (see Fig. 6), implying the CL teacher's wrong classification in this example. The LLM teacher, as we can also see from Fig. 6, assigns the highest logits to "kidney" and the second-highest to "cutting." Even though neither the CL teacher nor the LLM teacher independently produces precise answers, their combined capability, obtained through the weighted summation described in (2), correctly identifies the operation as "cutting". With the LLM functioning as an additional teacher in our multi-teacher framework, the student model becomes more adept at efficiently identifying the differences between similar answers. This example underscores the value of our proposed multi-teacher CL framework with adaptive weights, which effectively leverages the information in the LLM's logits to enhance overall accuracy and performance.

V. Conclusion and Future Work

We present an LLM-assisted multi-teacher framework for enhanced surgical VQA performance under a CL setup. With this innovative framework, we incorporate a multimodal LLM

into the CL training process to address the challenges of domain shift and data imbalance, effectively mitigating the catastrophic forgetting issue. The technical contributions of this paper are as follows. First, the novel data processing technique illustrated in this paper allows for the extraction of logits from complex LLM embeddings. Second, our adaptive weight assignment strategy strikes a balance between the domain-specific knowledge of previous CL teacher models and the LLM teacher's robust generalization capabilities. Third, the application of these methodologies has demonstrated high accuracy in handling VQA tasks in practical surgical scenarios. Finally, a newly released surgical VQA dataset serves as vital assets for future studies and research in this domain. Another notable contribution we made to the community is the introduction of a new research trajectory for utilizing LLMs in CL studies. In future works, we hope to explore the decomposition of representations into spatial and temporal spaces, which have higher task-invariance, to further reduce model forgetting. Additionally, integrating multimodal data, such as kinematics data from robot-assisted clinical systems, may further improve the model's performance in a CL setup.

REFERENCES

[1] L. Seenivasan, M. Islam, A. K. Krishna *et al.*, "Surgical-VQA: Visual question answering in surgical scenes using transformer," in *2022 MICCAI*. Springer, 2022, pp. 33–43.

[2] L. Bai, M. Islam, and H. Ren, "Co-attention gated vision-language embedding for visual question localized-answering in robotic surgery," *arXiv preprint arXiv:2307.05182*, 2023.

[3] L. Seenivasan, M. Islam, G. Kannan *et al.*, "SurgicalGPT: End-to-end language-vision GPT for visual question answering in surgery," *arXiv preprint arXiv:2304.09974*, 2023.

[4] B. D. Nguyen, T.-T. Do, B. X. Nguyen *et al.*, "Overcoming data limitation in medical visual question answering," in *2019 MICCAI*. Springer, 2019, pp. 522–530.

[5] Y. Khare, V. Bagal, M. Mathew *et al.*, "Mmbert: Multimodal Bert pretraining for improved medical vqa," in *2021 ISBI*. IEEE, 2021, pp. 1033–1036.

[6] H. Wang, Y. Jin, and L. Zhu, "Dynamic interactive relation capturing via scene graph learning for robotic surgical report generation," in *2023 IEEE ICRA*, 2023, pp. 2702–2709.

[7] Z. Zhao, Y. Jin, and P. Heng, "TraSeTR: Track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery," in *2022 IEEE ICRA*, 2022, pp. 11 186–11 193.

[8] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[9] K. Shu, H. Li, J. Cheng *et al.*, "Replay-oriented gradient projection memory for continual learning in medical scenarios," in *2022 IEEE BIBM*, 2022, pp. 1724–1729.

[10] M. M. Derakhshani, I. Najdenkoska, T. van Sonsbeek *et al.*, "Lifelonger: A benchmark for continual disease classification," in *2022 MICCAI*. Springer, 2022, pp. 314–324.

[11] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, 2017.

[12] J. Kirkpatrick, R. Pascanu, N. Rabinowitz *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci.*, vol. 114, no. 13, pp. 3521–3526, 2017.

[13] M. Lenga, H. Schulz, and A. Saalbach, "Continual learning for domain adaptation in chest x-ray classification," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 413–423.

[14] L. Bai, M. Islam, and H. Ren, "Revisiting distillation for continual learning on visual question localized-answering in robotic surgery," *arXiv preprint arXiv:2307.12045*, 2023.

[15] M. Phuong and C. Lampert, "Towards understanding knowledge distillation," in *2019 ICML*. PMLR, 2019, pp. 5142–5151.

[16] C. Simon, M. Faraki, Y.-H. Tsai *et al.*, "On generalizing beyond domains in cross-domain continual learning," in *2022 CVPR*, 2022, pp. 9265–9274.

[17] C. D. Kim, J. Jeong, and G. Kim, "Imbalanced continual learning with partitioning reservoir sampling," in *2020 ECCV*. Springer, 2020, pp. 411–428.

[18] A. Belyaeva, J. Cosentino, F. Hormozdiari *et al.*, "Multimodal LLMs for health grounded in individual-specific data," *arXiv preprint arXiv:2307.09018*, 2023.

[19] Y. Du, S. C. Liew, K. Chen *et al.*, "The power of large language models for wireless communication system development: A case study on FPGA platforms," *arXiv preprint arXiv:2307.07319*, 2023.

[20] Z. Guo, R. Zhang, X. Zhu *et al.*, "Point-bind & point-LLM: Aligning point cloud with multi-modality for 3D understanding, generation, and instruction following," *arXiv preprint arXiv:2309.00615*, 2023.

[21] H. Cui, Y. Du, Q. Yang *et al.*, "LLMind: Orchestrating AI and IoT with LLMs for complex task execution," *arXiv preprint arXiv:2312.09007*, 2023.

[22] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan *et al.*, "Large language models in medicine," *Nature medicine*, pp. 1–11, 2023.

[23] K. Chen, J. Li, K. Wang *et al.*, "Towards an automatic ai agent for reaction condition recommendation in chemical synthesis," *arXiv preprint arXiv:2311.10776*, 2023.

[24] S. Min, X. Lyu, A. Holtzman *et al.*, "Rethinking the role of demonstrations: What makes in-context learning work?" *arXiv preprint arXiv:2202.12837*, 2022.

[25] K. Chen, H. Cao, J. Li *et al.*, "An autonomous large language model agent for chemical literature data mining," *arXiv preprint arXiv:2402.12993*, 2024.

[26] K. Chen, Y. Du, T. You *et al.*, "Llm-assisted multi-teacher continual learning for visual question answering in robotic surgery," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 10 772–10 778.

[27] J. Gou, B. Yu, S. J. Maybank *et al.*, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[28] X. Dai, Z. Jiang, Z. Wu *et al.*, "General instance distillation for object detection," in *2021, CVPR*, 2021, pp. 7842–7851.

[29] D. Wenliang, L. Junnan, L. Dongxu *et al.*, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," *arXiv preprint arXiv:2305.06500*, 2023.

[30] H. W. Chung, L. Hou, S. Longpre *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[31] F. Thabtah, S. Hammoud, F. Kamalov *et al.*, "Data imbalance in classification: Experimental evaluation," *Information Sciences*, vol. 513, pp. 429–441, 2020.

[32] L. Bai, M. Islam, L. Seenivasan *et al.*, "Surgical-VQLA: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery," *arXiv preprint arXiv:2305.11692*, 2023.

[33] E. Rojas-Muñoz, K. Couperus, and J. Wachs, "Daisi: database for AI surgical instruction," *arXiv preprint arXiv:2004.02809*, 2020.

[34] D. R. Yates, C. Vaessen, and M. Roupret, "From Leonardo to Da Vinci: the history of robot-assisted surgery in urology," *BJU international*, vol. 108, no. 11, pp. 1708–1713, 2011.

[35] J. C.-Y. Ngu, C. B.-S. Tsang, and D. C.-S. Koh, "The Da Vinci Xi: a review of its capabilities, versatility, and potential role in robotic colorectal surgery," *Robotic Surgery: Research and Reviews*, pp. 77–85, 2017.

[36] K.-Y. Lei, W.-J. Xie, S.-Q. Fu *et al.*, "A comparison of the Da Vinci Xi vs. Da Vinci Si surgical systems for radical prostatectomy," *BMC surgery*, vol. 21, pp. 1–6, 2021.

[37] R. Liu, Q. Liu, G. Zhao *et al.*, "Single-port robotic pancreatic surgery using the Da Vinci SP system: A retrospective study on prospectively collected data in a consecutive patient cohort," *International Journal of Surgery*, vol. 104, p. 106782, 2022.

[38] N. Ragavan, S. Bharathkumar, P. Chirravur *et al.*, "Evaluation of Hugo RAS system in major urologic surgery: our initial experience," *Journal of Endourology*, vol. 36, no. 8, pp. 1029–1035, 2022.

[39] S. W. Lei, D. Gao, J. Z. Wu *et al.*, "Symbolic replay: Scene graph as prompt for continual learning on vqa task," in *2023 AAAI*, vol. 37, no. 1, 2023, pp. 1250–1259.

[40] S. Zhang and R. S. Sutton, "A deeper look at experience replay," *arXiv preprint arXiv:1712.01275*, 2017.

[41] A. Chaudhry, P. K. Dokania, T. Ajanthan *et al.*, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *2018 ECCV*, 2018, pp. 532–547.

[42] F. Huszár, "Note on the quadratic penalties in elastic weight consolidation," *Proc. Nat. Acad. Sci.*, vol. 115, no. 11, pp. E2496–E2497, 2018.