Surgical AI Copilot: Energy-Based Fourier Gradient Low-Rank Adaptation for Surgical LLM Agent Reasoning and Planning

Jiayuan Huang[1, 2, 6], **Runlong He**[1, 2], **Danyal Zaman Khan**[1, 3, 7], **Evangelos B. Mazomenos**[1, 2], **Danail Stoyanov**[1, 4], **Hani Marcus**[1, 3], **Linzhe Jiang**[1, 2], **Matthew J Clarkson**[1, 2], **Mobarak I. Hoque**[1, 2, 5]

[1]UCL Hawkes Institute, University College London, UK
[2]Dept of Medical Physics & Biomedical Engineering, University College London, UK
[3]Dept of Neurosurgery, National Hospital for Neurology and Neurosurgery, UK
[4]Dept of Computer Science, University College London, UK
[5]Division of Informatics, Imaging and Data Science, The University of Manchester, UK
[6]Visual Understanding Research Group, Dept of Informatics, King's College London, UK
[7]Institute of Neurology, University College London, UK
jiayuan.1.huang@kcl.ac.uk, runlong.he.23@ucl.ac.uk, mobarak.hoque@manchester.ac.uk

Abstract

Image-guided surgery demands adaptive, real-time decision support, yet static AI models struggle with structured task planning and providing interactive guidance. Large language models (LLMs)-powered agents offer a promising solution by enabling dynamic task planning and predictive decision support. Despite recent advances, the absence of surgical agent datasets and robust parameter-efficient fine-tuning techniques limits the development of LLM agents capable of complex intraoperative reasoning. In this paper, we introduce Surgical AI Copilot, an LLM agent for image-guided pituitary surgery, capable of conversation, planning, and task execution in response to queries involving tasks such as MRI tumor segmentation, endoscope anatomy segmentation, overlaying preoperative imaging with intraoperative views, instrument tracking, and surgical visual question answering (VQA). To enable structured agent planning, we develop the PitAgent dataset, a surgical context-aware planning dataset covering surgical tasks like workflow analysis, instrument localization, anatomical segmentation, and query-based reasoning. Additionally, we propose DEFT-GaLore, a Deterministic Energy-based Fourier Transform (DEFT) gradient projection technique for efficient low-rank adaptation of recent LLMs (e.g., LLaMA 3.2, Qwen 2.5), enabling their use as surgical agent planners. We extensively validate our agent's performance and the proposed adaptation technique against other state-of-the-art low-rank adaptation methods on agent planning and prompt generation tasks, including a zero-shot surgical VQA benchmark, demonstrating the significant potential for truly efficient and scalable surgical LLM agents in real-time operative settings.

Code — https://github.com/mobarakol/SurgicalAICopilot

Introduction

The integration of AI into image-guided robotic and laparoscopic surgery has the potential to revolutionize minimally invasive procedures by providing adaptive, interactive, and real-time decision support (Maier-Hein et al. 2022). However, traditional AI models, often static and lacking the ability to perform structured task planning and surgical workflow reasoning, fall short (Gao et al. 2024), making it challenging to coordinate tasks involving preoperative imaging, intraoperative video, tracking, and navigation (Chen et al. 2024). The challenge is particularly critical in endonasal pituitary surgery, where a narrow surgical corridor and proximity to critical neurovascular structures require highly context-aware support (Khan et al. 2023). Modern large language models (LLMs) and vision-language models (VLMs), however, offer a promising avenue for addressing this challenge by enabling AI-driven systems that can interpret surgical scenes, respond to natural language queries, and dynamically plan actions based on the evolving intraoperative context (Moghani et al. 2024). Building upon these advancements, there is a critical need for a specialized AI Agent that seamlessly integrates surgical workflow understanding, dynamic task planning, and interactive decision support.

Currently, several AI Agents have been developed for various application, such as MedAgents (Tang et al. 2023) designed to mimic clinical roles, SurgRAW designed to coordinate multiple vision-language agents using chain-of-thought reasoning in robotic-assisted surgery, and LlaVa-Surge designed to understand and answer open-ended questions about surgical videos. However, these existing agents either rely solely on zero-shot inference or exhaustive fine-tuning, and critically lack specialized datasets designed explicitly for surgical agent planning and workflow reasoning. Without agent-planning datasets,surgical AI agents lack comprehensive domain knowledge, which leads to hallucinations. It makes them struggle to coordinate the complex, dynamic tasks of surgery, fail to systematically capture the structured sequence of surgical workflows, and have difficulty maintaining long-term context across surgical stages, ultimately limiting their real-time applicability and clinical reliability.

Parameter-Efficient Fine-Tuning (PEFT) is a set of techniques used to adapt large pre-trained models to new tasks without updating all of the model's parameters. Although current PEFT methods such as LoRA (Hu et al. 2022) and DoRA (Liu et al. 2024) can significantly reduce trainable parameters while preserving performance, they often struggle to adapt to highly dynamic and context-sensitive surgical environments and may fail to fully capture the intricate relationships between multimodal surgical context. Although methods like GaLore (Zhao et al. 2024a) and DCT-GaLore (Modoranu et al. 2025) can enable memory-efficient optimization via gradient low-rank projections, they are not computationally efficient, and hard to capture time-varying dominant optimization direction in surgical scenarios.

In this paper, we address these limitations by introducing Surgical AI Copilot, the first surgical LLM agent planner for image-guided pituitary surgery, which is fine-tuned with a new DEFT-GaLore adaptation technique on our proposed PitAgent dataset to enable interactive planning, conversation, and action within this specific surgical context. Following the *Planner–Worker* agent framework (Xu et al. 2023; Zeng et al. 2024), our copilot explicitly separates reasoning and execution into a *Planner*, responsible for query decomposition and sub-task planning, and a *Worker* that invokes specialized surgical models to perform the corresponding actions. Our contributions are summarized as follows:

- We introduce PitAgent dataset, the first surgical context-aware dataset designed for task planning in endonasal pituitary surgery, encompassing segmentation, overlaying, instrument localization, tracking, phase identification, and transsphenoidal-specific surgical activity recognition, providing structured workflow information to enhance surgical decision-making.
- We propose DEFT-GaLore, a novel deterministic energy-based Fourier transform gradient projection method for low-rank adaptation, enabling efficient fine-tuning of LLaMA 3.2 (Grattafiori et al. 2024) and Qwen 2.5 (Zhu et al. 2025) to enhance real-time surgical decision support and task-specific adaptation.
- We evaluate our LLM-based agent's task planning, prompt generation and execution capabilities, benchmarking its adaptation performance against existing techniques, and validate its surgical VQA outputs on a public pituitary dataset, demonstrating semantically meaningful responses and superior real-world applicability.

Related Work

LLM based Agent in Medical Applications

The LLM-Agent represents a transformative advancement in surgical AI, capable of planning, memory retention, and real-time decision-making, making it an ideal interactive and collaborative copilot for surgeons (Gao et al. 2025; Kelly et al. 2024). Trained or fine-tuned on surgery-specific, large-scale datasets, the agent has the potential to plan sequences of tasks, communicate with other AI models or tools, and deliver evidence based support and visual assistance such as real time instrument-anatomy segmentation, tool tracking, diagnostic image segmentation, overlaying and visual question answering (VQA) onto intraoperative endoscopic views. Several clinical LLM-agents have been proposed and developed for various applications. For example, the Agent-Clinic benchmark (Dutta and Hsiao 2024), an automated diagnostic agent architecture for use in simulated clinical environments, where the agent dynamically corrects misdiagnoses and interacts with simulated patients to improve diagnostic accuracy. MedAide (Wei et al. 2024), a system uses multi-agent collaboration to provide integrated diagnostics and decision support, and performs well on multiple healthcare benchmarks. Additionally, efforts to develop surgical AI agents include SurgBox (Wu et al. 2024), which simulates surgical workflows, and VS-Assistant (Chen et al. 2024), which provides on-demand visual support in surgery. However, these agents either rely primarily on retrieval-augmented generation (RAG) (Gao et al. 2023) or rely solely on zero-shot inference or utilize simplistic classification-based call functions, neglecting comprehensive fine-tuning of LLMs for adaptive decision-making in surgical environments.

PEFT LLM Finetuning

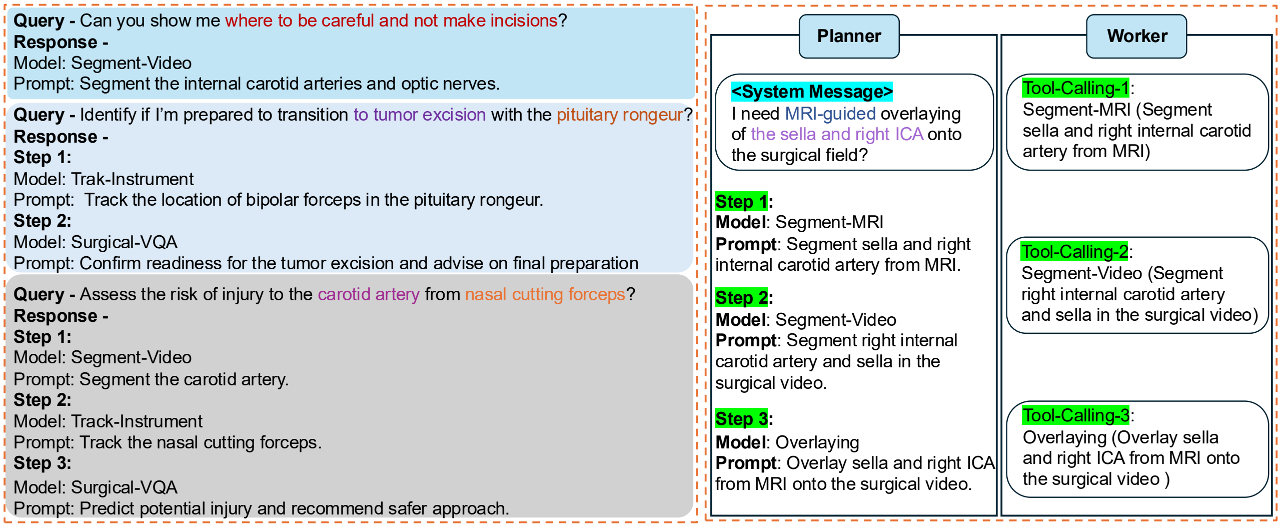
PEFT techniques can be applied to adapt LLMs efficiently to task-specific smaller datasets. There are currently two main categories of PEFT techniques designed to facilitate LLM adaptation. The first category comprises adaptation methods based on learning parameters, such as LoRA (Hu et al. 2022), MoRA (Jiang et al. 2024), and DoRA (Liu et al. 2024). The second category includes gradient low-rank projection approaches, such as GaLore (Zhao et al. 2024a), GaLore 2 (Su et al. 2025) and DCT-GaLore (Modoranu et al. 2025). While these techniques have demonstrated strong performance in reducing the number of trainable parameters, they also present certain limitations. Learning parameter-based adaptation methods may struggle to generalize to out-of-distribution tasks, particularly when the injection of additional parameters is suboptimally placed, potentially leading to degraded performance (Zhang et al. 2025; Han et al. 2024; Chen et al. 2023). On the other hand, gradient low-rank projection approaches, though efficient in terms of memory and computation, can introduce optimization instability when navigating complex loss landscapes (Zhao et al. 2024a). This is especially true in high-dimensional gradient spaces, where the choice of projection direction may constrain the model's learning capacity. In addition, the lack of surgery-specific agent datasets that encompass planning, interaction, and action within the surgical context further limits the development of truly adaptive and interactive AI copilots for surgery.

Methodology

Preliminaries: Gradient Low-Rank Projection

GaLore (Zhao et al. 2024a) is a technique for optimizing neural network training, aiming to reduce memory usage by lowering the rank of gradient matrices. For the gradient matrix $G_t \in \mathbb{R}^{m \times n}$ at iteration t , GaLore applies singular value decomposition (SVD) with rank k to obtain projection matrices $P_t \in \mathbb{R}^{m \times r}$ and $Q_t \in \mathbb{R}^{n \times r}$ as follows:

$$G_t = USV^\top \approx \textstyle\sum_{i=1}^{r} s_i u_i v_i^\top \tag{1}$$



Query - Can you show me where to be careful and not make incisions?
Response -
Model: Segment-Video
Prompt: Segment the internal carotid arteries and optic nerves.

Query - Identify if I'm prepared to transition to tumor excision with the pituitary rongeur?
Response -
Step 1:
Model: Trak-Instrument
Prompt: Track the location of bipolar forceps in the pituitary rongeur.
Step 2:
Model: Surgical-VQA
Prompt: Confirm readiness for the tumor excision and advise on final preparation

Query - Assess the risk of injury to the carotid artery from nasal cutting forceps?
Response -
Step 1:
Model: Segment-Video
Prompt: Segment the carotid artery.
Step 2:
Model: Track-Instrument
Prompt: Track the nasal cutting forceps.
Step 3:
Model: Surgical-VQA
Prompt: Predict potential injury and recommend safer approach.

<System Message>
I need MRI-guided overlaying of the sella and right ICA onto the surgical field?

Step 1:
Model: Segment-MRI
Prompt: Segment sella and right internal carotid artery from MRI.

Step 2:
Model: Segment-Video
Prompt: Segment right internal carotid artery and sella in the surgical video.

Step 3:
Model: Overlaying
Prompt: Overlay sella and right ICA from MRI onto the surgical video.

Tool-Calling-1:
Segment-MRI (Segment sella and right internal carotid artery from MRI)

Tool-Calling-2:
Segment-Video (Segment right internal carotid artery and sella in the surgical video)

Tool-Calling-3:
Overlaying (Overlay sella and right ICA from MRI onto the surgical video)

(a) A sample annotation of PitAgent dataset

(b) How our AI Copilot employs the dataset with Planner and Worker

Figure 1: (a) denotes Query and Response samples of PitAgent dataset. (b) denotes how the *Planner-Worker* architecture employs the dataset to enable multi-step decision-making and real-time guidance for surgical scenarios.

where $\mathrm{P}_t = [u_1, u_2, \dots, u_k]$, $Q_t = [v_1, v_2, \dots, v_k]$. Then, the gradient update using the projected matrices is formulated as:

$$W_T = W_0 + \eta \sum_{t=0}^{T-1} \left[P_t\, \rho_t \left(P_t^\top G_t\, Q_t \right) Q_t^\top \right] \qquad (2)$$

Where ρ_t is an entry-wise stateful gradient regularizer (e.g., Adam), $P_t^\top G_t Q_t$ is the low-rank matrix projected into the $\mathbb{R}^{r \times r}$space and η is the learning rate. While GaLore significantly reduces memory consumption by maintaining low-rank gradient statistics, the primary limitations are the computational expense of performing SVD.

GaLore 2 (Su et al. 2025) improves the SVD to a Fast Randomized SVD, but it is still essentially an SVD, although it reduces the training time, but it still essentially uses a high computational complexity SVD, and the primary limitation is not solved.

DCT-GaLore (Modoranu et al. 2025) replaces SVD with DCT to construct an orthogonal basis and alleviates optimization information loss via a dynamic column selection strategy that selects basis vectors based on gradient similarity. However, since it samples only a few columns from a fixed, gradient-independent cosine basis containing only real components, phase information in the frequency domain is lost, reducing projection expressiveness and optimization efficiency.

Proposed Method: Surgical AI Copilot

We design the Surgical LLM Copilot by creating the PitAgent dataset and DEFT-GaLore to adapt open-source LLMs (LLaMA 3.2 and Qwen 2.5) into a surgical AI agent. This agent enables task planning and prompt generation for promptable AI models supporting surgical tasks like VQA, video segmentation, MRI segmentation, overlaying, and instrument tracking. As shown in Figure 2, our copilot integrates two core components: the *Planner* and the *Worker*. Together, they form an adaptive, interactive framework for pituitary surgery. The *Planner* acts as the cognitive center, dynamically interpreting surgeon queries, decomposing them into structured sub-tasks, and generating plans with model-specific prompts based on real-time context. Once tasks are identified and sequenced, the *Planner* delegates execution to the Worker. The *Worker* leverages specialized promptable visual and multimodal models to execute these tasks using multimodal inputs. After execution, the agent returns visual or textual results back to the user. This continuous interplay between the *Planner*'s context understanding and the *Worker*'s multimodal processing ensures precise, real-time, and contextually relevant surgical assistance.

PitAgent Dataset Currently, there is a lack of specialized datasets in the field of endonasal pituitary surgery that are explicitly designed for surgical agent planning and workflow reasoning. To address this gap, we introduce PitAgent, a surgical-domain dataset designed to facilitate fine-tuning of our surgical LLM copilot planner for performing reasoning over queries, available inputs and models to generate plans and corresponding prompts.

During the dataset construction process, and with the support of our clinical collaborators, we first identified queries that closely reflect those a surgeon would naturally pose in real surgical scenarios across 6 query-response categories. As shown in Table 1, Four categories represent the agent planning with single-model tasks, and the remaining two categories (Multitask-1 and Multitask-2) involve coordinating multiple base models. Based on the complexity of each query, we annotated the agent planning sequences with the appropriate model selection along with the prompt required to invoke the corresponding model, which serves as the response to the query. The planning annotations are aligning publicly available pituitary dataset of endoscopic videos,

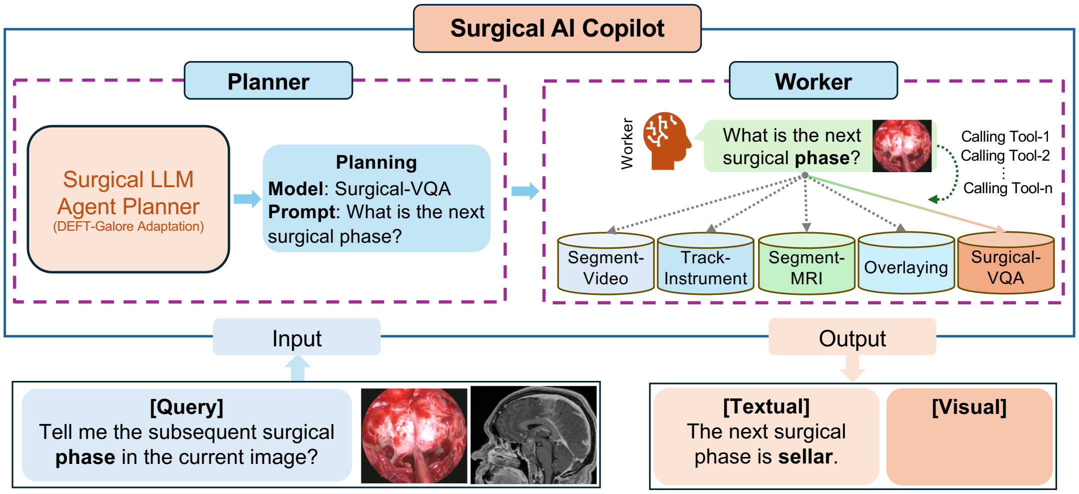


Figure 2: The *Planner-Worker* architecture of our Surgical LLM Agent. The *Planner* analyzes the surgeon's query and generates planning prompts. Based on these prompts, the *Worker* dynamically calls multiple promptable AI models such as segmentation, tracking, and MRI registration to collaboratively answer surgical queries.

Table 1: Key content and number of unique queries of each category in PitAgent dataset. Q. denotes unique quantity of queries.

Category	Q.	Key Content
Surgical-VQA	52	Surgical step, Surgical phase, Next surgical phase, Next surgical step, Number of instruments, Instrument location, Surgical activity, Instrument recognition, Remaining time
Segment-Video	27	Sella, Clival Recess, Left Carotid, Right Carotid, Left Optic Protuberance, Right Optic Protuberance, Carotid, Optic Protuberance, All Anatomical (Neurovascular) Structures
Segment-MRI	15	Pituitary Adenoma, Pituitary Tumor, Internal Carotid Artery (ICA)
Track-Instrument	27	Bipolar Forceps, Cottle Elevator, Cup Forceps, Dural Scissors, Freer Elevator, Hemostatic Foam, Irrigation Syringe, Kerrison Rongeur, Micro-Doppler Probe, Nasal Cutting Forceps, Pituitary Rongeur, Retractable Knife, Ring Curette, Spatula Dissector, Stealth Pointer, Suction Cannula, Surgical Drill, Tissue Glue
Multitask-1 (2 models)	90	Segment-Video→Segment-MRI, Segment-Video→Track-Instrument, Segment-Video→Surgical-VQA, Track-Instrument→Surgical-VQA, Overlaying→Surgical-VQA, Surgical-VQA →Surgical-VQA
Multitask-2 (3 models)	50	Segment-Video→Track-Instrument→Surgical-VQA, Surgical-VQA→Surgical-VQA→Surgical-VQA, Model1→Model2→Overlaying

tasks of video segmentation (Mao et al. 2024), MRI segmentation (Zhao et al. 2024b), tracking (Islam, VS, and Ren 2020), overlaying (Enkaoua et al. 2023), and Surgical-VQA (He et al. 2024). Some examples of the annotated samples are illustrated in Figure 1(a), it shows how our LLM copilot employs the dataset with *Planner* and *Worker*. Subsequently, all query-response pairs were reviewed and validated by clinical collaborators. The final number of unique query-response pairs per category is summarized in Table 1. Among these categories, the Multitask-1 category contains the largest number of pairs (90), while Segment-MRI includes the fewest (15). The unique question numbers of Surgical-VQA, Segment-Video, Track-Instrument, and Multitask-2 are 52, 27, 27, and 50 respectively. Each finalized unique query was further rephrased 50 times, resulting in a total of 13,050 query-response pairs. Table 1 also details the scope of each task: Surgical-VQA covers 8 surgical concepts, Segment-Video handles 6 anatomical structures, Segment-MRI focuses on 3 anatomical structures, and Track-Instrument includes 18 instrument types. multitasks includes two or three combinations of promptable models. Finally, we split the query-response pairs into training and testing sets in an 8:2 ratio.

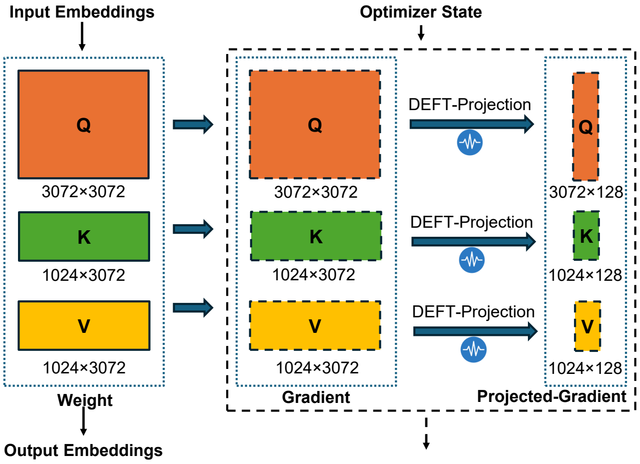


Figure 3: Example of constructing DEFT-GaLore projection gradient for LLaMA 3.2 model.

Deterministic Energy-based Fourier Transform Gradient Low-rank Projection To overcome the bottleneck caused by aforementioned gradient low-rank projection techniques, we introduce Deterministic Energy-Based Fourier Transform GaLore (DEFT-GaLore), a novel gradient low-rank projection method that replaces SVD with an efficient frequency-domain analysis. Energy is defined as the sum of squared magnitudes, which measures the informational content of a matrix and is numerically equivalent to the square of its Frobenius norm (Reza and Tripathi 2019). The proposed method rests on two theoretical pillars: (i) Projection matrix construction based on Frobenius norms minimises information loss due to projection (Modoranu et al. 2025), and (ii) Parseval's theorem guarantees that the energy computed in the frequency domain equals that in the time (or spatial) domain (Fein-Ashley 2025).

DEFT-GaLore is proposed to compress full-rank gradients into a low-rank subspace. To this end, the method first transforms into the frequency domain using a Fast Fourier Transform (FFT), computes the spectral energy of the matrix, and then extracts a sketch matrix corresponding to the dominant frequencies. Finally, a real-valued orthogonal projection matrix is constructed from this sketch. Specifically, given an input matrix $G_t \in \mathbb{R}^{m\times n}$ at step t and target rank k, the DEFT-GaLore algorithm computes a low-rank orthogonal projection matrix $M_t \in \mathbb{R}^{m\times k}$ through the following steps:

1. **Fast Fourier Transform (FFT)**: Apply FFT along the column dimension to G_t:
$$G_{t,f} = \mathcal{F}(G_t) \in \mathbb{R}^{m\times n} \qquad (3)$$
where G_t is real-valued input matrix at step t (time/spatial domain); $G_{t,f}$ is complex frequency-domain representation after FFT; $\mathcal{F}(\cdot)$ is the Fast Fourier Transform applied column-wise.
2. **Energy Spectrum Calculation**: Compute the energy for each frequency component:
$$s_j = \sum_{i=1}^{m} \left|G_{t,f}(i,j)\right|^2, \quad j = 1, 2, \ldots, n \qquad (4)$$
where s_j is energy value at frequency index j; $G_{t,f}(i,j)$is complex element at row i, column j; $|\cdot|$is Complex modulus operator.
3. **Frequency Selection**: Identify the top-k frequency indices with maximum energy:
$$\mathcal{I} = \{j_1, j_2, \ldots, j_k\} \subseteq \{1, 2, \ldots, n\} \qquad (5)$$
where $\mathcal{I}$ is the index set sorted in ascending order.
4. **Sketch Matrix Construction**: Extract columns corresponding to $\mathcal{I}$:
$$Y_c = G_{t,f}(:, \mathcal{I}) \in \mathbb{C}^{m\times k} \qquad (6)$$
where Y_c is the complex sketch matrix.
5. **Complex QR Decomposition**: Perform QR decomposition on the sketch matrix:
$$Q_c, R_c = \text{QR}(Y_c), \quad Q_c \in \mathbb{C}^{m\times k},\ R_c \in \mathbb{C}^{k\times k} \qquad (7)$$
where Q_c is the complex orthogonal basis matrix.
6. **Real-Imaginary Separation**: Decompose Q_c into real components:
$$Y_r = \text{reshape}\left(\text{RealImag}(Q_c)\right) \in \mathbb{R}^{m\times 2k} \qquad (8)$$
where $\text{RealImag}(Q_c)$ splits each complex element $q = a + bi$ into (a, b); reshape converts the $\mathbb{C}^{m\times k}$ matrix to $\mathbb{R}^{m\times 2k}$
7. **Real QR Decomposition**: Perform QR decomposition on the extended real matrix:
$$Q_r, R_r = \text{QR}(Y_r), \quad Q_r \in \mathbb{R}^{m\times 2k},\ R_r \in \mathbb{R}^{2k\times 2k} \qquad (9)$$
8. **Final Projection Matrix**: Extract the first k columns:
$$M_t = Q_r(:, 1:k) \in \mathbb{R}^{m\times k} \qquad (10)$$

The complete DEFT-GaLore projection is summarized as:
$$M_t = \text{QR}\left[\text{reshape}\left(\text{RealImag}\left(\text{QR}\left(G_{t,f}(:, \mathcal{I})\right)\right)\right)\right]_{(:,1:k)} \qquad (11)$$
A similar process can be applied to the transpose of the gradient matrix to obtain the right projection matrix $N_t \in \mathbb{R}^{n\times k}$. Finally, given the pretrained weight $W_0 \in \mathbb{R}^{m\times n}$ the gradient update using the projected matrices for our DEFT-GaLore can be derived as:
$$W_T = W_0 + \eta \textstyle\sum_{t=0}^{T-1} \left[M_t\, \rho_t\Big(M_t^\top\, G_t\, N_t\Big)\, N_t^\top\right] \qquad (12)$$

DCT-GaLore is constrained by its reliance on a fixed global orthogonal basis composed exclusively of real cosine functions with even symmetry. When this basis is truncated to a low rank, it can only adjust the amplitude, while the phase information, implicitly encoded in higher-frequency components, is lost during truncation. In comparison, our approach enables both amplitude and phase information to be preserved in the low-rank projection, and it selects the columns with the highest energy, allowing for more accurate maintenance of gradient directionality in low-rank settings.

As shown in Figure 3, we integrated DEFT-GaLore into every transformer block of the LLaMA 3.2 models, specifically applying it to the query projection (Q), key projection (K), value projection (V), and output projection (O) layers, while maintaining a projection rank of 128 for a fair comparison with the original GaLore.

Experiments and Results

Dataset

In addition to our PitAgent dataset, we also test our agent generated question prompts using a publicly available dataset of open-ended PitVQA (He et al. 2025). The dataset, derived from 25 videos of pituitary surgery (He et al. 2024), consists of 59 unique questions spanning 6 key aspects, including surgical phases, steps, instruments, tool-tissue interaction, position and quantity. To evaluate our agent-generated prompts, we retained the original validation sets (He et al. 2025), which include 24,767 frames and 182,720 question-answer pairs.

Implementation Details

The experiments are based on the LLaMA-3.2-3B-Instruct (Grattafiori et al. 2024) and Qwen-2.5-1.5B-Instruct (Zhu et al. 2025) models provided by HuggingFace [1]. We first test the zero-shot performance of task plan-

[1] https://huggingface.co/meta-llama

Table 2: Comparison of the performance of various PEFT methods in terms of Prompt Generation, Planning, and Training Overhead. Results cover PEFT methods for the zero-shot LLMs and two PEFT-fine-tune models, with bold being the best in the same segment.

Method	**Prompt Generation**				**Planning**		**Training Overhead**
	BLEU-3 (%)	BLEU-4 (%)	ROUGE-L (%)	METEOR (%)	F1 (%)	ACC (%)	Time (second)
Zero-shot							
LLaMA 3.2 (Grattafiori et al. 2024)	10.88	8.21	44.89	52.05	48.17	32.5	—
Qwen 2.5 (Zhu et al. 2025)	3.41	2.19	21.25	36.64	45.33	31.65	—
Gemma 2 (Team et al. 2024)	3.03	2.21	12.25	27.55	26.67	18.42	—
Deepseek llm (Bi et al. 2024)	16.30	12.40	43.64	49.49	42.67	30.22	—
LLaMA-3.2-3B PEFT							
LoRA (Hu et al. 2022)	29.47	25.14	63.78	70.02	70.5	61.46	—
DoRA (Liu et al. 2024)	51.57	46.32	71.76	75.48	77.17	64.65	—
MoRA (Jiang et al. 2024)	33.52	28.88	65.12	69.43	69.33	58.16	—
GaLore (Zhao et al. 2024a)	74.11	70.68	87.81	89.12	97.67	94.94	8762.43
GaLore 2 (Su et al. 2025)	73.18	70.28	88.38	90.54	98.17	94.85	6891.79
DCT-GaLore (Modoranu et al. 2025)	74.15	71.21	87.24	89.96	97.17	92.76	3535.69
DEFT-GaLore (ours)	**74.27**	**71.88**	**89.87**	**92.40**	**99.00**	**96.25**	3355.75
Qwen-2.5-1.5B PEFT							
LoRA (Hu et al. 2022)	12.48	8.39	35.97	41.92	25.83	15.81	—
DoRA (Liu et al. 2024)	27.54	22.87	57.15	67.32	78.33	69.86	—
MoRA (Jiang et al. 2024)	8.94	5.47	33.94	40.85	26.5	16.3	—
GaLore (Zhao et al. 2024a)	52.24	47.68	73.15	77.09	88.67	79.07	2733.62
GaLore 2 (Su et al. 2025)	51.11	46.29	72.70	76.42	88.83	75.15	2033.09
DCT-GaLore (Modoranu et al. 2025)	44.81	40.39	67.68	72.88	85.17	67.29	1775.41
DEFT-GaLore (ours)	**59.3**	**55.53**	**78.72**	**84.1**	**95.33**	**91.96**	1764.88

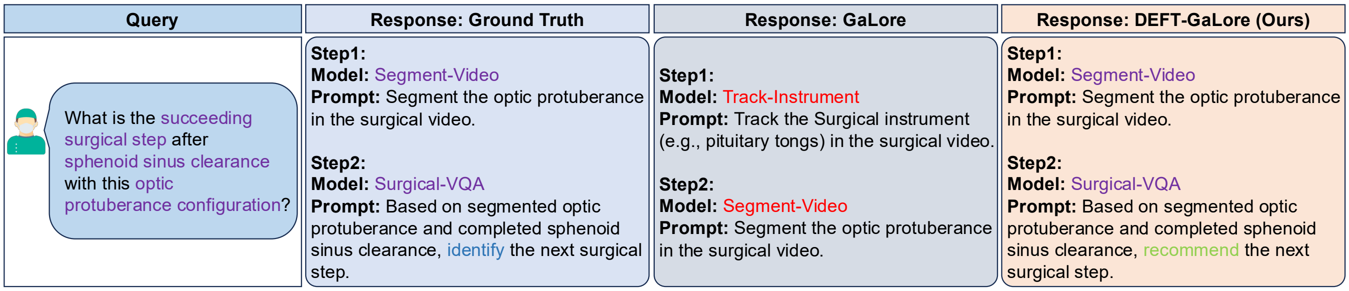
ning and prompt generation abilities on above-mentioned models and two extra models, they are Gemma-2-2b (Team et al. 2024) and Deepseek-llm-7b-chat (Bi et al. 2024). Then We fine-tune the LLaMA and Qwen models with a cross-entropy loss and the AdamW optimizer, setting the learning rate and rank to 3×10^{-7} and 128 respectively. We compare our DEFT-GaLore with various parameter-efficient fine-tuning (PEFT) methods including LoRA (Hu et al. 2022), MoRA (Jiang et al. 2024), DoRA (Liu et al. 2024), GaLore (Zhao et al. 2024a), GaLore2 (Su et al. 2025), and DCT-GaLore (Modoranu et al. 2025). We also use the generated prompts from our agent as queries to the open-source PitVQA++ network (He et al. 2025) to further evaluate its performance on visual question answering tasks. All experiments are conducted within the PyTorch framework on an NVIDIA RTX A100_80 GPU. More details such as *System Message* used, *external validation* and *Pseudo-code* are included in the technical appendix.

Results

***Planner*'s Performance** The performance of the *Planner* is evaluated from two aspects: task planning and prompt generation. We evaluate prompt generation quality using BLEU, ROUGE, and METEOR, and assess its task planning accuracy using accuracy (ACC) and F1 scores. Table 2 presents the results of performance of four zero-shot models. It also presents the performance of *Planner* fine-tuned by our DEFT-GaLore compared with that fine-tuned by other PEFT methods. The zero-shot performance reveals that existing LLMs exhibit disappointing capabilities in task planning and prompt generation within the domain of pituitary surgery. With LLaMA as the agent *Planner*, our method improved BLEU-4, METEOR, and F1 scores by 1.20%, 3.28%, and 1.33% over the GaLore baseline on the PitAgent dataset. When using Qwen, the corresponding improvements reached 8.85%, 7.01%, and 12.89%. Similar trends are observed in other metrics of ROUGE and ACC. Building upon this, our DEFT-GaLore achieves best-in-class results on all metrics across both PEFT model lines. It outperforms other state-of-the-art PEFT methods in both prompt generation and planning accuracy. Figure 4 visualizes the task-planning results of GaLore and DEFT-GaLore. Compared with the ground truth, the *Planner* fine-tuned with GaLore misinterprets the query and selects the wrong model in both steps; the prompt it generates also differs greatly in meaning from the correct prompt. In contrast, DEFT-GaLore not only performs task planning accurately but also produces a prompt whose structure closely mirrors the reference prompt and conveys the same semantics.

Table 2 also presents the average time consumption of GaLore, DCT-GaLore and our DEFT-GaLore on LLaMA and Qwen models for fine-tuning one epoch. For LLaMA model, GaLore 2, DCT-GaLore and our DEFT-GaLore reduce the training time by 21.35% 59.65% and 61.7% respectively compared to GaLore. For the Qwen model, the training time for both is reduced by 25.63%, 35.04% and 35.44%, respectively.

***Worker*'s Performance** The performance of the *Worker* is evaluated by assessing the downstream model's output when given prompts generated by the planner as input. Table 3 compares *Worker*'s performance using agent-generated



Query

What is the succeeding surgical step after sphenoid sinus clearance with this optic protuberance configuration?

Response: Ground Truth

Step1:
Model: Segment-Video
Prompt: Segment the optic protuberance in the surgical video.

Step2:
Model: Surgical-VQA
Prompt: Based on segmented optic protuberance and completed sphenoid sinus clearance, identify the next surgical step.

Response: GaLore

Step1:
Model: Track-Instrument
Prompt: Track the Surgical instrument (e.g., pituitary tongs) in the surgical video.

Step2:
Model: Segment-Video
Prompt: Segment the optic protuberance in the surgical video.

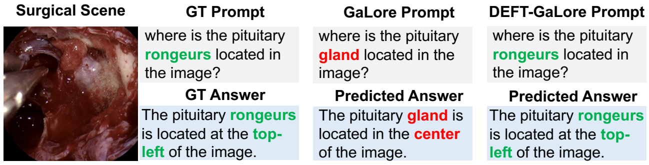
Response: DEFT-GaLore (Ours)

Step1:
Model: Segment-Video
Prompt: Segment the optic protuberance in the surgical video.

Step2:
Model: Surgical-VQA
Prompt: Based on segmented optic protuberance and completed sphenoid sinus clearance, recommend the next surgical step.

Figure 4: Qualitative results analysis of agent planning. The responses are generated by Ground Truth, GaLore fine-tuned *Planner* and our DEFT-GaLore fine-tuned *Planner*.

prompts versus ground-truth prompts on a publicly available open-ended PitVQA dataset (He et al. 2025). We evaluated our agent's response quality in surgical VQA using two publicly available VLMs: LLaMA-3.2-Vision-11B (Grattafiori et al. 2024) and PitVQA++ (He et al. 2025), where the performance metrics for *Planner*-generated prompts are obtained from zero-shot PitVQA++ with its pretrained weights. The evaluation used standard metrics such as BLEU, ROUGE, and METEOR scores, measuring the impact of agent-generated prompts on model performance. While agent-generated prompts showed slightly lower performance than GT prompts (on average decreases of 19.44% and 16.24% in BLEU-4 and METEOR), our DEFT-GaLore outperformed other methods, improving BLEU-4 and METEOR scores by 3.42% and 3.48% over the GaLore baseline. Figure 5 shows qualitative results for surgical VQA. Prompt generation errors can affect answer accuracy, for instance, when the agent confused ”rongeurs” with ”gland”, the downstream PitVQA++ model failed to locate the pituitary rongeurs.



Surgical Scene

GT Prompt: where is the pituitary rongeurs located in the image?

GaLore Prompt: where is the pituitary gland located in the image?

DEFT-GaLore Prompt: where is the pituitary rongeurs located in the image?

GT Answer: The pituitary rongeurs is located at the top-left of the image.

Predicted Answer: The pituitary gland is located in the center of the image.

Predicted Answer: The pituitary rongeurs is located at the top-left of the image.

Figure 5: Qualitative result analysis of *Worker* on *Planner*-generated prompt vs GT prompt using publicly available zero-shot PitVQA++ model (He et al. 2025).

Table 3: *Worker*'s performance with agent-generated vs. ground-truth prompts (zero-shot PitVQA++). Metrics: BLEU, ROUGE-L and METEOR.

Model		B-3 (%)	B-4 (%)	R-L (%)	MET. (%)
Ground Truth Prompt	PitVQA++ (He et al. 2025)	78.55	76.37	84.64	84.39
	LLaMA-Vision	24.52	18.44	53.86	56.96
Agent-generated Prompt (LLaMA 3.2)	LoRA (Hu et al. 2022)	36.39	31.74	46.41	41.90
	DoRA (Liu et al. 2024)	52.01	47.56	64.63	65.37
	MoRA (Jiang et al. 2024)	47.72	42.19	57.27	53.38
	GaLore (Zhao et al. 2024a)	69.57	67.31	77.76	76.82
	GaLore 2 (Su et al. 2025)	71.51	68.98	80.10	79.60
	DCT-GaLore (Mod. et al. 2025)	72.52	70.01	80.23	79.69
	DEFT-GaLore (ours)	**73.31**	**70.73**	**80.77**	**80.3**

Ablation Study Table 4 illustrates that rank-128 performs the best on LLaMA 3.2 model among all ranks tested. Both excessively low and high ranks slightly degrade the performance metrics. This indicates that the DEFT-GaLore method is sensitive to the rank hyperparameter: a rank that is too low may lead to underfitting due to the loss of optimization information, while a rank that is too high may introduce noise and cause overfitting.

Table 4: Comparison of performance of LLaMA-based *Planner* fine-tuned with DEFT-GaLore on different rank. The rank is set to 64, 128, and 256, respectively.

Rank	Prompt Generation				Planning	
	BLEU-3 (%)	BLEU-4 (%)	ROUGE-L (%)	METEOR (%)	F1 (%)	ACC (%)
64	68.00	65.71	87.37	90.63	98.17	95.62
128	**74.27**	**71.88**	**89.87**	**92.40**	**99.00**	**96.25**
256	71.80	69.04	87.88	90.23	97.33	95.51

Discussion and Conclusion

We presented Surgical AI Copilot, a novel *Planner–Worker* LLM agent framework for pituitary surgery that combines multi-step reasoning, tool invocation, and language-driven prompting through two core innovations: the PitAgent dataset, a richly annotated benchmark for surgical planning, and DEFT-GaLore, a Fourier-based low-rank adaptation method for efficient and effective LLM fine-tuning. Key findings highlight that this combination consistently improves the agent's planning accuracy and prompt quality compared to existing PEFT baselines, demonstrating its effectiveness in surgical decision support. The Copilot effectively decomposes complex surgical queries, selects appropriate visual tools, and generates high-quality prompts, even in zero-shot settings, demonstrating its utility in real-time intraoperative workflows. However, the current system is limited to a single surgical domain. In this work, we validate the system only on a single visual model focused on surgical VQA. Future work will focus on expanding to other text-prompted visual models, such as segmentation, instrument detection, and overlaying, and evaluating the agent's end-to-end performance across broader surgical scenarios, with the goal of developing a more generalizable, memory-augmented, and robust AI assistant for real-time surgical decision-making.

Acknowledgements

This work was supported by the EPSRC under grant [EP/W00805X/1, EP/Z534754/1]. HJM was supported by the NIHR UCLH/UCL Biomedical Research Centre. DZK was supported by Cleveland Clinic London for this work. Thanks to Medtronic for access to the *Touch Surgery^{TM} Ecosystem* for video recording, annotation, and storage. HJM is employed by and holds shares in Panda Surgical Ltd.

References

Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Chen, J.; Zhang, A.; Shi, X.; Li, M.; Smola, A.; and Yang, D. 2023. Parameter-efficient fine-tuning design spaces. *arXiv preprint arXiv:2301.01821*.

Chen, Z.; Luo, X.; Wu, J.; Chan, D.; Lei, Z.; Wang, J.; Ourselin, S.; and Liu, H. 2024. VS-assistant: versatile surgery assistant on the demand of surgeons. *arXiv preprint arXiv:2405.08272*.

Dutta, A.; and Hsiao, Y.-C. 2024. Adaptive reasoning and acting in medical language agents. *arXiv preprint arXiv:2410.10020*.

Enkaoua, A.; Islam, M.; Ramalhinho, J.; Dowrick, T.; Booker, J.; Khan, D. Z.; Marcus, H. J.; and Clarkson, M. J. 2023. Image-guidance in endoscopic pituitary surgery: an in-silico study of errors involved in tracker-based techniques. *Frontiers in Surgery*.

Fein-Ashley, J. 2025. The FFT Strikes Back: An Efficient Alternative to Self-Attention. *arXiv e-prints*, arXiv–2502.

Gao, S.; Fang, A.; Huang, Y.; Giunchiglia, V.; Noori, A.; Schwarz, J. R.; Ektefaie, Y.; Kondic, J.; and Zitnik, M. 2024. Empowering biomedical discovery with AI agents. *Cell*, 187(22): 6125–6151.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2.

Gao, Z.; Zhang, B.; Li, P.; Ma, X.; Yuan, T.; Fan, Y.; Wu, Y.; Jia, Y.; Zhu, S.-C.; and Li, Q. 2025. Multi-modal Agent Tuning: Building a VLM-Driven Agent for Efficient Tool Usage. In *The Thirteenth International Conference on Learning Representations*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; . . . ; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint*, 2407.21783.

Han, Z.; Gao, C.; Liu, J.; Zhang, J.; and Zhang, S. Q. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

He, R.; Khan, D. Z.; Mazomenos, E. B.; Marcus, H. J.; Stoyanov, D.; Clarkson, M. J.; and Islam, M. 2025. Pitvqa++: Vector matrix-low-rank adaptation for open-ended visual question answering in pituitary surgery. *arXiv preprint arXiv:2502.14149*.

He, R.; Xu, M.; Das, A.; Khan, D. Z.; Bano, S.; Marcus, H. J.; Stoyanov, D.; Clarkson, M. J.; and Islam, M. 2024. Pitvqa: Image-grounded text embedding llm for visual question answering in pituitary surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 488–498. Springer.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Islam, M.; VS, V.; and Ren, H. 2020. AP-MTL: Attention Pruned Multi-task Learning Model for Real-time Instrument Detection and Segmentation in Robot-assisted Surgery. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Jiang, T.; Huang, S.; Luo, S.; Zhang, Z.; Huang, H.; Wei, F.; Deng, W.; Sun, F.; Zhang, Q.; Wang, S.; Wang, D.; and Zhuang, F. 2024. MoRA: High-Rank Updating for Parameter-Efficient Fine-Tuning. *arXiv preprint arXiv:2405.12130*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1601–1611.

Kelly, C.; Hu, L.; Yang, B.; Tian, Y.; Yang, D.; Yang, C.; Huang, Z.; Li, Z.; Hu, J.; and Zou, Y. 2024. Visiongpt: Vision-language understanding agent using generalized multimodal framework. *arXiv preprint arXiv:2403.09027*.

Khan, D. Z.; Hanrahan, J. G.; Baldeweg, S. E.; Dorward, N. L.; Stoyanov, D.; and Marcus, H. J. 2023. Current and future advances in surgical therapy for pituitary adenoma. *Endocrine Reviews*, 44(5): 947–959.

Liu, S.-Y.; Wang, C.-Y.; Yin, H.; Molchanov, P.; Wang, Y.-C. F.; Cheng, K.-T.; and Chen, M.-H. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Maier-Hein, L.; Eisenmann, M.; Sarikaya, D.; März, K.; Collins, T.; Malpani, A.; Fallert, J.; Feussner, H.; Giannarou, S.; Mascagni, P.; et al. 2022. Surgical data science–from concepts toward clinical translation. *Medical image analysis*, 76: 102306.

Mao, Z.; Das, A.; Islam, M.; Khan, D. Z.; Williams, S. C.; Hanrahan, J. G.; Borg, A.; Dorward, N. L.; Clarkson, M. J.; Stoyanov, D.; Marcus, H. J.; and Bano, S. 2024. PitSurgRT: real-time localization of critical anatomical structures in endoscopic pituitary surgery. *International Journal of Computer Assisted Radiology and Surgery*.

Modoranu, I.-V.; Safaryan, M.; Schultheis, E.; and Alistarh, D. 2025. SVD-Free Low-Rank Adaptive Gradient Optimization for Large Language Models. *arXiv preprint arXiv:2505.17967*.

Moghani, M.; Doorenbos, L.; Panitch, W. C.-H.; Huver, S.; Azizian, M.; Goldberg, K.; and Garg, A. 2024. SuFIA: language-guided augmented dexterity for robotic surgical assistants. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6969–6976. IEEE.

Reza, A.; and Tripathi, R. 2019. A class of randomized Subset Selection Methods for large complex networks. *arXiv preprint arXiv:1905.04452*.

Su, D.; Gu, A.; Xu, J.; Tian, Y.; and Zhao, J. 2025. Galore 2: Large-scale llm pre-training by gradient low-rank projection. *arXiv preprint arXiv:2504.20437*.

Tang, X.; Zou, A.; Zhang, Z.; Li, Z.; Zhao, Y.; Zhang, X.; Cohan, A.; and Gerstein, M. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Wei, J.; Yang, D.; Li, Y.; Xu, Q.; Chen, Z.; Li, M.; Jiang, Y.; Hou, X.; and Zhang, L. 2024. Medaide: Towards an omni medical aide via specialized llm-based multi-agent collaboration. *arXiv preprint arXiv:2410.12532*.

Wu, J.; Liang, X.; Bai, X.; and Chen, Z. 2024. SurgBox: Agent-Driven Operating Room Sandbox with Surgery Copilot. In *2024 IEEE International Conference on Big Data (BigData)*, 2041–2048. IEEE.

Xu, B.; Peng, Z.; Lei, B.; Mukherjee, S.; Liu, Y.; and Xu, D. 2023. Rewoo: Decoupling reasoning from observations for efficient augmented language models. *arXiv preprint arXiv:2305.18323*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380.

Zeng, A.; Liu, M.; Lu, R.; Wang, B.; Liu, X.; Dong, Y.; and Tang, J. 2024. AgentTuning: Enabling Generalized Agent Abilities for LLMs. In *Findings of the Association for Computational Linguistics ACL 2024*, 3053–3077.

Zhang, D.; Feng, T.; Xue, L.; Wang, Y.; Dong, Y.; and Tang, J. 2025. Parameter-efficient fine-tuning for foundation models. *arXiv preprint arXiv:2501.13787*.

Zhao, J.; Zhang, Z.; Chen, B.; Wang, Z.; Anandkumar, A.; and Tian, Y. 2024a. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*.

Zhao, Y.; Bai, L.; Zhang, Z.; Wu, Y.; Islam, M.; and Ren, H. 2024b. Transferring Knowledge from High-Quality to Low-Quality MRI for Adult Glioma Diagnosis. In *BraTS-SSA Challenge, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) Workshop*.

Zhu, S.; Hu, W.; Yang, Z.; Yan, J.; and Zhang, F. 2025. Qwen-2.5 outperforms other large language models in the chinese national nursing licensing examination: Retrospective cross-sectional comparative study. *JMIR Medical Informatics*, 13: e63731.

Appendix A: System Message for the Agent Planning

A system message is an instruction given to a LLM to guide its behavior, tone, or role throughout the conversation. In our work, the system message specifies the role of LLM, the overall task, the available text-promptable downstream models and examples of response. We tested 20 different system messages on the zero-shot LLaMA 3.2 and Qwen 2.5 models to evaluate their optimal task planning and prompt generation performance. After testing, the system message we selected is as follows:

> You are a surgical AI agent assisting in pituitary surgery. Your job is to handle surgeons' queries efficiently by choosing appropriate text-promptable AI models and generating corresponding prompts.
> Available models: Segment-Video, Segment-MRI, Track-Instrument, Surgical-VQA, Overlaying.
> - Use one model if query focuses on a single, simple aspect:
> Example (single-model):
>
> > Model: Segment-Video
> > Prompt: Segment the sella in the video.
>
> - Use MULTIPLE models if query requires several types of information:
> Example (multi-model):
>
> > Step1:
> > Model: Segment-MRI
> > Prompt: Segment the pituitary tumor from MRI.
> > Step2:
> > Model: Segment-Video
> > Prompt: Segment the sella in the video.
>
> Now, follow the same format to answer the provided question-no extra text, labels, or formatting.

Appendix B: Algorithm for DEFT-GaLore

DEFT-GaLore is a memory-efficient training method that significantly reduces optimizer state memory. Instead of using costly SVD, it employs a highly efficient energy-based FFT projection to identify the gradient's most important directions. This projection is updated periodically to amortize computational cost. The full gradient is projected into a low-rank subspace, where the compact AdamW optimizer states (moments) are maintained and updated. This low-rank update is then projected back to the original dimension to update the model's weights, followed by a standard decoupled weight decay step. This approach effectively balances substantial memory savings with high computational performance.

Algorithm 1: AdamW with DEFT-GaLore

```
Require: Layer weights $\mathbf{W} \in \mathbb{R}^{m \times n}$, learning rate $\eta$, betas $\beta_1, \beta_2$, epsilon $\epsilon$, weight decay $\lambda$, rank $k$, subspace update frequency $T$.
Initialize $\mathbf{M}_0 \leftarrow 0$, $\mathbf{V}_0 \leftarrow 0$, step $t \leftarrow 0$.
Initialize projection matrix $P_t \leftarrow$ None.
repeat
    $G_t \leftarrow \nabla_{\mathbf{W}_t} \mathcal{L}(\mathbf{W}_t)$    ▷ Compute gradient
    if $t \pmod{T} = 0$ or $P_t$ is None then
        $P_t \leftarrow$ DEFT-Projection$(G_t, r)$ Update projector (see Alg. 2)
    else
        $P_t \leftarrow P_{t-1}$    ▷ Reuse previous projector
    end if
    $\tilde{G}_t \leftarrow P_t^T G_t$ Project gradient into low-rank space
    $t \leftarrow t + 1$
    $\mathbf{M}_t \leftarrow \beta_1 \cdot \mathbf{M}_{t-1} + (1 - \beta_1) \cdot \tilde{G}_t$
    $\mathbf{V}_t \leftarrow \beta_2 \cdot \mathbf{V}_{t-1} + (1 - \beta_2) \cdot \tilde{G}_t^2$
    $\hat{\mathbf{M}}_t \leftarrow \mathbf{M}_t / (1 - \beta_1^t)$    ▷ Bias correction
    $\hat{\mathbf{V}}_t \leftarrow \mathbf{V}_t / (1 - \beta_2^t)$
    $N_t \leftarrow \hat{\mathbf{M}}_t / (\sqrt{\hat{\mathbf{V}}_t} + \epsilon)$ ▷ Compute low-rank update
    $\tilde{G}_t \leftarrow P_t N_t$ ▷ Project update back to original space
    $\mathbf{W}_t \leftarrow \mathbf{W}_{t-1} - \eta \cdot \tilde{G}_t$    ▷ Update weights
    $\mathbf{W}_t \leftarrow \mathbf{W}_t - \eta \cdot \lambda \cdot \mathbf{W}_t$ ▷ Apply decoupled weight decay
until convergence criteria met
return $\mathbf{W}_T$
```

Algorithm 2: DEFT-Projection

```
Require: Matrix $G_t \in \mathbb{R}^{m \times n}$, target rank $k$.
Ensure: Orthogonal basis $M \in \mathbb{R}^{m \times k}$.
$G_{t,f} \leftarrow \mathrm{FFT}(G_t)$    ▷ Apply FFT along columns
score $\leftarrow \sum_{i=0}^{m-1} |G_{t,f}|^2$    ▷ Calculate energy score per frequency
idx $\leftarrow$ indices of top $k$ values in score
$Y_c \leftarrow G_{t,f}[:, \mathrm{idx}]$    ▷ Sketch with top-k complex components
$Q_c, \_ \leftarrow \mathrm{QR}(Y_c)$    ▷ Complex QR decomposition
$Y_r \leftarrow \mathrm{Reshape}(\mathrm{RealImag}(Q_c))$ to matrix in $\mathbb{R}^{m \times 2k}$
$Q_r, \_ \leftarrow \mathrm{QR}(Y_r)$    ▷ Real QR decomposition
return $Q_r[:, : k]$    ▷ Return first k orthogonal vectors
```

In the main text, we named the projection matrix M_t. Here, to distinguish it from the AdamW momentum, we name the projection matrix P_t.

Appendix C: Supplementary Implementation Details of The *Planner* Fine-tuning Experiment

Fo the PEFT methods based on learning parameters (LoRA (Hu et al. 2022), DoRA (Liu et al. 2024), MoRA (Jiang et al. 2024)), We set their learning rate (LR) to 3×10^{-7}, batch size to 6, number of epochs to 10, rank to 8, dropout to 0.3, and scaling factor to 16.

For PEFT methods based on gradient low-rank projection (GaLore (Zhao et al. 2024a), GaLore 2 (Su et al. 2025), DCT-GaLore (Modoranu et al. 2025), and our DEFT-GaLore), we also set their learning rate to 3×10^{-7}, batch size to 6, number of epochs to 5, rank to 128, projection matrix update interval to 50, scaling factor to 1, and projection strategy to *reverse_std*. All other hyperparameters use the default settings of Galore[2].

In the GaLore design, matrix projection has multiple modes. The standard mode (*std*) adapts the projection direction based on the shape of the gradient matrix. When the height of the matrix is greater than or equal to the width, right projection $G_t Q_t$ is performed, and otherwise, left projection $P_t^T G_t$ is performed. This strategy selects the smaller of the matrix dimensions for projection. The reverse std method performs the inverse operation and left projection when the height of the matrix is greater than or equal to the width. This strategy selects the larger of the matrix dimensions for projection. We use the same projection design when reproducing the DCT-GaLore and designing the DEFT-GaLore method.

Appendix D: Validation on Rewoo Planner Benchmark

We also validate our DEFT-GaLore on a benchmark agent planning dataset: the ReWOO Planner Instruction Tuning dataset. This dataset, released by the ReWOO team [3] and referred to as "rewoo/planner_instruction_tuning_2k" on Hugging Face (Xu et al. 2023), contains 2,000 high-quality examples designed to fine-tune the planning component of large language models. It emphasizes explicit multi-step reasoning by providing planning trajectories for complex question-answering tasks, particularly from HotpotQA (Yang et al. 2018) and TriviaQA (Joshi et al. 2017). Each example includes an instruction, an input query, and a sequence of structured planning steps that incorporate external tool calls (e.g., Wikipedia[input]) and reference intermediate evidence variables (e.g., #E1, #E2). This setup allows models to learn how to decompose a task into executable steps and gather supporting evidence before generating a final answer. In our experiments, we used 1,500 samples for training and held out the remaining 500 for testing, following the setup in (Zeng et al. 2024). This dataset serves as a valuable resource for evaluating and improving the planning and reasoning capabilities of ReWOO-style agents.

To implement this experiment, we keep the same configuration as the Planner fine-tuning experiment except for changing learning rate, seed, and number of epochs to 5×10^{-5}, 2025, and 15 respectively.

[2] https://github.com/jiaweizzhao/GaLore

[3] https://huggingface.co/datasets/rewoo/planner_instruction_tuning_2k

Table 5: Comparison of PEFT methods on the ReWOO Planner Instruction Tuning dataset for multi-step planning performance. Our method, DEFT-GaLore, achieves the best results across BLEU-4, ROUGE-L, and METEOR, demonstrating state-of-the-art performance in structured multi-step planning for LLM-based agents.

Method	BLEU-4 (%)	ROUGE-L (%)	METEOR (%)
Zero-Shot	2.1	15.3	21.4
LoRA (Hu et al. 2022)	52.5	68.7	76.1
DoRA (Liu et al. 2024)	53.5	69.2	76.1
MoRA (Jiang et al. 2024)	50.9	67.6	75.7
GaLore (Zhao et al. 2024a)	58.5	71.8	78.7
GaLore 2 (Su et al. 2025)	62.3	73.9	80.1
DCT-GaLore (Modoranu et al. 2025)	62.7	74.2	80.7
DEFT-GaLore (ours)	**64.9**	**75.5**	**81.8**

Table 1 showcases the multi-step planning performance of our method, DEFT-GaLore, compared to other state-of-the-art PEFT approaches on the benchmark ReWOO Planner Instruction Tuning dataset. DEFT-GaLore consistently outperforms prior methods, achieving up to 3.5% absolute improvement in BLEU-4, 1.3% in ROUGE-L, and 1.1% in METEOR over the best previous SOTA (DCT-GaLore). These gains demonstrate the effectiveness of our FFT-based low-rank adaptation in enhancing the planning capability of LLM agents for structured multi-step reasoning tasks.