SD-LLaVA: Surgical Subspace Decoupling for Continual Instruction Tuning in LLaVA

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Continual instruction tuning for multimodal large language models is essential for enabling progressive acquisition of diverse capabilities without access to historical data. In endoscopic scene understanding, this challenge is amplified by the need to support diverse tasks with increasing reasoning complexity. In this work, we first introduce a continual instruction tuning benchmark built upon Surg-396K, covering five tasks with increasing reasoning complexity. Then, we propose a unified continual instruction tuning framework that effectively balances knowledge retention and task adaptability. Our approach introduces an orthogonal gradient constraint on LoRA updates in shallow layers to prevent interference with previously learned knowledge, while preserving shared representations. For deep layers, we employ a mixture-of-experts mechanism with instruction-aware gating to dynamically combine task-specific adapters for adaptive multi-task response. Extensive experiments demonstrate that our method outperforms existing approaches, achieving superior knowledge retention and consistent gains across all tasks.

Keywords: Continual learning · Multimodal large language model · Instruction tuning · Endoscopic surgery.

1 Introduction

Multimodal large language models (MLLMs) have recently enabled flexible vision-language reasoning by following natural language instructions [9, 11]. However, in real-world scenarios, models are required to continuously acquire new capabilities over time while preserving previously learned knowledge, which remains a fundamental challenge for continual learning (CL) [19]. In endoscopic surgical scene understanding [1, 16], visual perception involves multiple cognitive levels, including instrument recognition, relation reasoning, region understanding, target localization, and detailed description. These capabilities form a natural progression from simple to complex reasoning and typically emerge sequentially in practice. Therefore, models must incrementally learn new tasks while maintaining prior knowledge and supporting diverse instruction formats. Yet, this process often leads to catastrophic forgetting [13, 18].

This problem is further complicated by clinical constraints. Due to strict data privacy regulations, historical data may not be retained [8], making joint training infeasible. Meanwhile, deploying separate models for different tasks leads to

excessive memory usage and latency. To address these challenges, we construct a multi-task continual instruction tuning benchmark for endoscopic scene understanding based on the Surg-396K dataset [16]. The benchmark includes five tasks with increasing reasoning complexity, i.e., Single Phrase, Visual Question Answering, Region-based QA, Grounding QA, and Detailed Description.

In parallel, we propose a unified continual instruction tuning (CIT) framework that enables a single model to learn these tasks sequentially. Specifically, we impose orthogonal gradient constraints on LoRA [6] adapters to prevent interference with previously learned knowledge while encouraging feature reuse. In addition, we introduce a dynamic task-aware fusion mechanism via an large language model (LLM) to predict task relevance from input instructions and then adaptively combine task-specific adapters, enabling adaptive multi-task response to diverse instructions. Our main contributions are summarized as follows:

- We present a multi-task CIT benchmark for endoscopic scene understanding with progressively increasing task complexity.
- We propose a CIT framework with orthogonal gradient constraints and dynamic adapter fusion to mitigate forgetting while enabling flexible response.
- Extensive experiments demonstrate improved knowledge retention and positive transfer across tasks.

2 Methodology

2.1 Preliminaries

Problem Setting. Continual instruction tuning aims to enable MLLMs to sequentially learn new tasks while preserving previously acquired capabilities. Consider a sequence of T tasks For each task, the dataset $\mathcal{D}^t = \{(x_v^{t,i}, x_{ins}^{t,i}, x_{ans}^{t,i})\}_{i=1}^{N_t}$ comprises image tokens x_v, instruction tokens x_{ins}, and answer tokens x_{ans} of length L. During training, the model sees only $\mathcal{D}^t$ without access to past and future data. Given parameters θ, MLLM minimizes the autoregressive loss:

$$\mathcal{L}^t(\theta; \mathcal{D}^t) = -\sum_{i=1}^{N_t} \sum_{l=1}^{L} \log p(x_{ans}^l \mid x_v, x_{ins}, x_{ans}^{<l}; \theta). \tag{1}$$

2.2 Orthogonal Subspace Decoupling for Shallow Layers

In CL, catastrophic forgetting primarily arises because parameter updates for new tasks alter previously learned feature representations and disrupt old decision boundaries [12]. Neuroscience suggests that new memories are encoded in neural subspaces approximately orthogonal to existing ones, minimizing interference between different memories [3, 15]. Inspired by this, we impose orthogonal constraints between tasks during the optimization of shallow model parameters, restricting new task weight updates to the orthogonal complement of the subspace spanned by previous task updates.

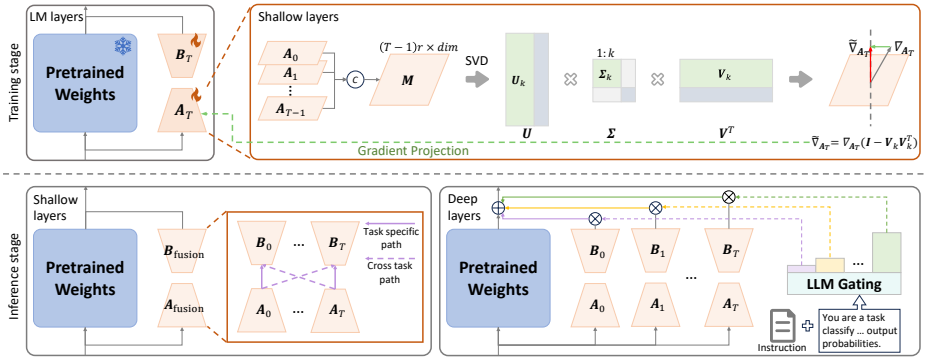

Fig. 1: Overview of the proposed method. (a) **Training phase**: For shallow layers, orthogonal projection constrains the gradient update direction of the input projection matrix $\mathbf{A}_T$ to be orthogonal to the parameter update directions of old tasks (Eq. 6). For deep layers, each task independently trains its LoRA experts without orthogonal constraints. (b) **Inference phase**: For shallow layers, the LoRA modules of all tasks are merged via parameter addition (Eq. 7). For deep layers, an LLM-based task gating mechanism computes task weights based on the user instruction, which are then used to perform weighted fusion of expert outputs (Eq. 8).

Consider continual learning with $T-1$ learned tasks, each task obtains its optimal parameter increment $\Delta\mathbf{W}_t$. To mitigate interference between tasks, for a new task T, we enforce orthogonality in the parameter space:

$$\langle \Delta\mathbf{W}_t, \Delta\mathbf{W}_T \rangle_F = \mathrm{tr}\big(\Delta\mathbf{W}_t^\top \Delta\mathbf{W}_T\big) = 0, \quad \forall t < T, \tag{2}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. However, directly imposing this in full dimension $\mathbb{R}^{d\times d}$ is costly. Parameter-efficient fine-tuning methods like LoRA can represent updates as $\Delta\mathbf{W} = \mathbf{BA}$, enabling constraints in low-rank subspaces.

Theorem 1 (Sufficient Condition for LoRA Subspace Orthogonality). *If the row space of the new parameters $\mathbf{A}_T$ is orthogonal to that of the historical parameters $\mathbf{A}_t$, i.e., $\mathbf{A}_t\mathbf{A}_T^\top = \mathbf{0}$, then for any $\mathbf{B}_t$ and $\mathbf{B}_T$, $\langle \Delta\mathbf{W}_t, \Delta\mathbf{W}_T \rangle_F = 0$.*

Proof. Substituting LoRA decomposition into Frobenius inner product:

$$\langle \Delta\mathbf{W}_t, \Delta\mathbf{W}_T \rangle_F = \mathrm{tr}\big((\mathbf{B}_t\mathbf{A}_t)^\top(\mathbf{B}_T\mathbf{A}_T)\big) = \mathrm{tr}\big(\mathbf{A}_t^\top\mathbf{B}_t^\top\mathbf{B}_T\mathbf{A}_T\big)\,.$$

Using the cyclic property of the trace, the above expression can be rewritten as:

$$\mathrm{tr}\big(\mathbf{A}_t^\top\mathbf{B}_t^\top\mathbf{B}_T\mathbf{A}_T\big) = \mathrm{tr}\big(\mathbf{A}_T\mathbf{A}_t^\top\mathbf{B}_t^\top\mathbf{B}_T\big)\,. \tag{3}$$

If $\mathbf{A}_t\mathbf{A}_T^\top = \mathbf{0}$, then its transpose $\mathbf{A}_T\mathbf{A}_t^\top = \mathbf{0}$. Consequently, the trace is identically zero, proving the theorem.

Theorem 1 suggests that enforcing orthogonality on the row spaces of $\mathbf{A}$ is sufficient to decouple task update directions, without directly constraining $\mathbf{B}$. This property significantly reduces the computational complexity of subspace decoupling. To this end, we first concatenate the $\mathbf{A}_t$ matrices from all previous tasks to construct their joint row space $\mathbf{M}$:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_{T-1} \end{bmatrix} \in \mathbb{R}^{(T-1)r \times d}, \tag{4}$$

and then perform Singular Value Decomposition (SVD) on $\mathbf{M}$:

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top. \tag{5}$$

The first k columns of $\mathbf{V}$, corresponding to the significant singular values, form an orthonormal basis $\mathbf{P} = \mathbf{V}_k^\top \in \mathbb{R}^{k \times d}$ of the subspace, satisfying $\mathbf{P}\mathbf{P}^\top = \mathbf{I}_k$ and $\text{span}(\mathbf{P}^\top) = \text{rowspan}(\mathbf{M})$.

When training the new task T, we constrain $\mathbf{A}_T$'s update gradient to the orthogonal complement of this space, i.e., $\mathbf{I} - \mathbf{P}^\top\mathbf{P}$, by orthogonally projecting the gradient direction. Given the original gradient $\nabla_{\mathbf{A}_T}\mathcal{L}$, the projected effective gradient is:

$$\tilde{\nabla}_{\mathbf{A}_T} = \nabla_{\mathbf{A}_T}\mathcal{L} - (\nabla_{\mathbf{A}_T}\mathcal{L}\mathbf{P}^\top)\mathbf{P} = \nabla_{\mathbf{A}_T}\mathcal{L}(\mathbf{I} - \mathbf{P}^\top\mathbf{P}). \tag{6}$$

Notably, since the randomly initialized $\mathbf{A}_T$ is approximately orthogonal to the low-dimensional historical subspace in the high-dimensional space, and all subsequent updates lie in the orthogonal complement, the update directions of new and old tasks become decoupled during optimization.

During inference, shallow layers retain some cross-task shareability. Thus we adopt parameter merging to integrate LoRA modules. Specifically, we sum all learned $\mathbf{A}_i$ and $\mathbf{B}_i$ to obtain the fused increment:

$$\Delta\overline{\mathbf{W}}_T = \left(\sum_{i=1}^{T}\mathbf{B}_i\right)\left(\sum_{i=1}^{T}\mathbf{A}_i\right) = \underbrace{\sum_{i=1}^{T}\mathbf{B}_i\mathbf{A}_i}_{\text{task-specific features}} + \underbrace{\sum_{i\neq j}\mathbf{B}_i\mathbf{A}_j}_{\text{shared features}}, \tag{7}$$

where the first term corresponds to task-specific components. Thanks to the enforced row-space orthogonality, $\mathbf{A}_i\mathbf{A}_j^\top \approx \mathbf{0}$, minimizing interference between $\mathbf{B}_i\mathbf{A}_i$ and $\mathbf{B}_j\mathbf{A}_j$ and ensuring independence of task-specific knowledge. The second term captures shared features across task. Since shallow networks typically learn general low-level semantics, these cross-terms preserve cross-task shared knowledge. This merging strategy decouples and unifies task-specific knowledge without extra parameters, thereby enabling synergistic representation of task knowledge within a unified model structure.

Table 1: Summary of the five tasks in the continual instruction tuning.

Task	Input Format	Expected Output	Metric
Single Phrase (SP)	Question requiring a concise term	Single word or short phrase	Accuracy
Visual QA (VQA)	Open-ended question about the scene	Complete sentence with context	ROUGE-L
Region-based QA (RQA)	Question with bounding box coordinates	Textual answer specific to the region	ROUGE-L
Grounding QA (GQA)	Question requesting object location	Bounding box coordinates	mIoU
Detailed Description (DD)	Prompt for comprehensive analysis	Multi-attribute paragraph	LLM Score

2.3 Mixture-of-Experts Isolation for Deep Layers

In deep layers, feature representations are typically more task-specific and exhibit weaker cross-task shareability [4]. Therefore, we adopt a Mixture of Experts (MoE) [14] mechanism to isolate tasks and eliminate parameter interference. During training, each task t independently trains its $\mathbf{A}_t$ and $\mathbf{B}_t$ without any orthogonal projection constraints. This design physically separates expert parameters across tasks, avoiding mutual interference during parameter updates and allowing each expert to fully learn task-specific high-level features.

During inference, however, the task identity is often unknown, posing a challenge for expert selection. A learnable gating network would itself suffer from catastrophic forgetting. Thus we need an expert selection mechanism that adapts to dynamic task changes. Leveraging the powerful semantic understanding of large language models (LLMs) [20], we introduce an LLM as a task resolver that infers the task from the user instruction's semantic content. Concretely, we design a prompt-based task gating mechanism. For each test sample, we construct a prompt template containing task descriptions and the instruction, feed it into an LLM, and let the model predict a probability distribution over tasks. The LLM produces a probability vector $\mathbf{w} = [w_1, w_2, \ldots, w_T]$ with $\sum_{i=1}^{T} w_i = 1$, where w_i denotes the probability that the sample belongs to task i. These probabilities serve as gating weights to dynamically fuse the LoRA expert parameters from all tasks, yielding the final model output:

$$\mathbf{y} = \left(\mathbf{W}_0 + \sum_{i=1}^{T} w_i \cdot (\mathbf{B}_i \mathbf{A}_i) \right) \mathbf{x}, \tag{8}$$

where $\mathbf{W}_0$ is the pretrained weight, $\mathbf{A}_i, \mathbf{B}_i$ are the LoRA expert parameters for task i, and $\mathbf{x}$ is the input feature. This approach preserves the specificity of each task expert while enabling adaptive task selection via LLM semantic understanding, requiring no extra training of a task classifier and aligning naturally with instruction tuning scenarios.

3 Experiments

3.1 Experiment Setting

Datasets. Visual data are drawn from two public endoscopic surgery datasets, EndoVis-VQLA [1] and CoPESD [17], covering laparoscopic cholecystectomy,

Table 2: Comparison of different CL methods on the EndoVis-VQLA part of the Surg-396K dataset in terms of *Average* and *Last* scores. The best and second-best results are highlighted in **bold** and underline, respectively.

	Method	SP	VQA	RQA	GQA	DD	**Average**
	Zero-shot	0.00	43.38	31.61	0.00	4.39	15.87
Average	EWC [7]	25.13	67.53	42.65	80.67	53.97	53.99
	LwF [10]	25.61	69.10	44.95	80.50	**54.30**	54.89
	MoELoRA [2]	25.40	65.75	37.56	**81.07**	53.84	52.72
	HiDe-LLaVA [4]	35.04	69.34	80.26	78.30	36.44	59.88
	Ours	**60.14**	**74.75**	**81.73**	78.73	53.97	**69.86**
Last	EWC [7]	0.00	48.47	22.46	77.10	53.97	40.40
	LwF [10]	0.00	48.53	24.92	77.00	**54.30**	40.95
	MoELoRA [2]	0.00	47.63	20.53	77.34	53.84	39.87
	HiDe-LLaVA [4]	6.36	50.51	77.62	77.31	36.44	49.65
	Ours	**58.41**	**69.11**	**80.73**	**78.02**	53.97	**68.05**

nephrectomy, and endoscopic submucosal dissection. Instruction pairs are sourced from Surg-396k [16]. Table 1 summarizes the five tasks arranged sequentially (SP → VQA → RQA → GQA → DD).

Evaluation Metrics. We adopt task-specific metrics as detailed in Table 1 to accommodate heterogeneous output types. To facilitate unified analysis, all raw scores are linearly normalized to $[0, 100]$ and denoted as $A_{i,j}$, representing the model's performance on task i after completing training on task j. Given T total tasks, we follow the protocol [4, 5] and evaluate the model's proficiency and knowledge retention using the average performance $Average_i$, final retention $Last_i$, stage-wise average accuracy A_t, and forgetting measure F_t.

3.2 Comparison with State-of-the-art Methods

Table 2 reports the *Average* and *Last* scores on EndoVis-VQLA. Our method achieves the best results on both metrics (69.86% and 68.05%), outperforming HiDe-LLaVA by 9.98% and 18.41%, respectively. Notably, severe forgetting is observed in prior methods on early tasks. For example, EWC, LwF, and MoELoRA drop to 0.00% on the SP task, while HiDe-LLaVA achieves only 6.36%. In contrast, our method maintains 58.41%, demonstrating strong knowledge retention. On the final DD task, our method achieves performance comparable to the best baseline while significantly outperforming HiDe-LLaVA, indicating a favorable stability–plasticity trade-off. Table 3 shows that our method achieves the lowest or near-lowest forgetting across all stages. Unlike prior methods where forgetting accumulates, ours decreases, indicating better knowledge consolidation and forward transfer. Futhermore, results on CoPESD (Table 4) confirm the superiority

Table 3: Comparison on EndoVis-VQLA in terms of Accuracy (A↑) and Forgetting (F↓) after each task.

Method	**SP**	**VQA**		**RQA**		**GQA**		**DD**	
	A_1↑	A_2↑	F_2↓	A_3↑	F_3↓	A_4↑	F_4↓	A_5↑	F_5↓
EWC [7]	67.17	59.95	27.58	57.75	28.18	44.19	45.70	40.40	41.45
LwF [10]	67.17	60.38	26.72	59.89	25.17	45.56	43.91	40.95	40.89
MoELoRA [2]	67.17	59.72	28.05	59.87	25.15	38.27	53.88	39.87	42.30
HiDe-LLaVA [4]	67.17	66.64	13.48	61.11	21.77	63.59	18.26	49.65	24.33
Ours	67.17	**69.03**	**10.27**	**70.88**	**9.49**	**74.56**	**4.27**	**68.05**	**6.19**

Table 4: Comparison on the CoPESD part of Surg-396K in terms of Accuracy (A↑) and Forgetting (F↓).

Method	**SP**	**VQA**		**RQA**		**GQA**		**DD**	
	A_1↑	A_2↑	F_2↓	A_3↑	F_3↓	A_4↑	F_4↓	A_5↑	F_5↓
EWC [7]	78.44	65.90	12.26	53.63	25.34	64.91	14.99	59.47	19.78
LwF [10]	78.44	66.27	11.40	53.37	25.72	55.17	27.95	57.02	22.79
MoELoRA [2]	78.44	60.12	23.70	47.78	33.92	64.23	15.76	56.34	23.68
HiDe-LLaVA [4]	78.44	63.41	12.97	42.60	36.10	61.54	14.87	46.42	27.46
Ours	78.44	**68.06**	**7.29**	**59.18**	**14.88**	**71.35**	**4.89**	**66.29**	**9.44**

of our method, which achieves the highest accuracy and lowest forgetting across all tasks, demonstrating strong robustness.

Figure 2 presents qualitative comparisons after training on all tasks. Existing methods either fail to follow instructions or suffer from semantic errors. In contrast, our method produces accurate outputs across tasks. These results highlight our method's superior instruction following and knowledge retention.

3.3 Ablation Study

Table 5 ablates key components by dividing the 32-layer LLM into shallow (layers 1–16) and deep (layers 17–32) parts.

Table 5: Ablation studies on the contribution of each component.

Shallow layers	Deep layers	Average ↑	Last ↑	F ↓
–	MoE	68.31	66.50	7.16
MoE	MoE	69.69	68.54	5.07
Gradient Projection	MoE	69.86	68.05	6.19

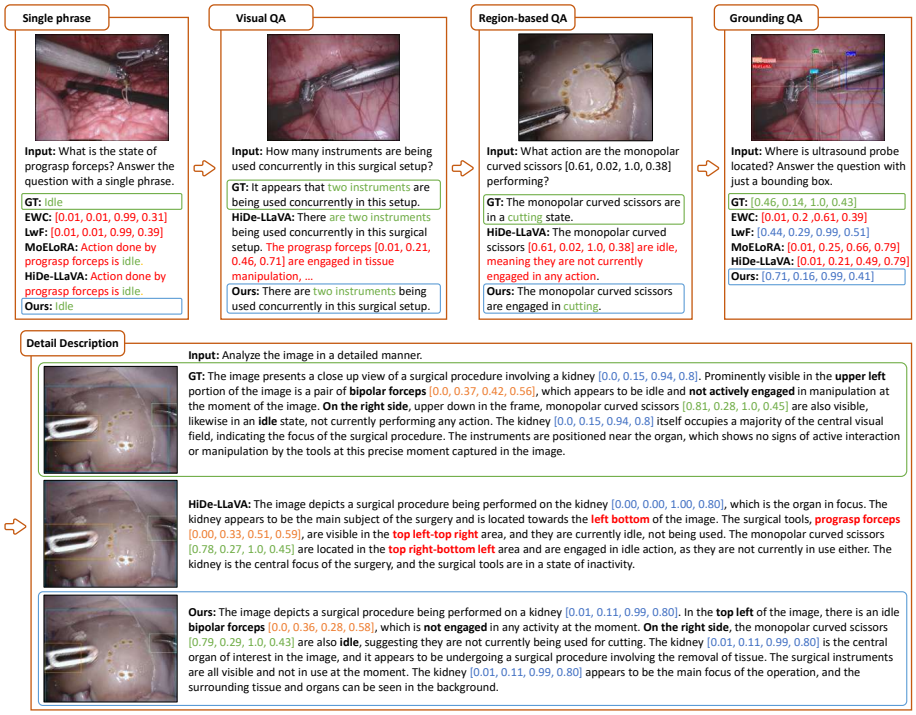

Fig. 2: Qualitative comparison on EndoVis-VQLA after learning all tasks.

(1) *No constraint on shallow layers, MoE only on deep*: *Last*=66.50%, *F*=7.16%, showing shallow representations are vulnerable without explicit regularization. (2) *MoE on all layers*: achieves lowest forgetting (5.07%) and *Last*=68.54%, but at the cost of linear parameter growth and slightly lower *Average*. (3) *Ours (gradient projection on shallow, MoE on deep)*: attains the highest *Average* (69.86%), comparable *Last*, and reduced forgetting (6.19%). This design protects shared shallow features without parameter explosion while enabling task-specific adaptation in deep layers.

4 Conclusion

We present a unified framework for CIT of MLLMs in endoscopic scene understanding. By combining orthogonal subspace constraints in shallow layers with instruction-aware mixture-of-experts in deep layers, our method effectively mitigates catastrophic forgetting while preserving task-specific expressivity. Extensive experiments demonstrate consistent improvements in both knowledge retention and overall performance across tasks. These results suggest a promising direction for building continually evolving MLLMs in real-world clinical settings.

References

1. Bai, L., Islam, M., Seenivasan, L., Ren, H.: Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. arXiv preprint arXiv:2305.11692 (2023)
2. Chen, C., Zhu, J., Luo, X., Shen, H.T., Song, J., Gao, L.: Coin: A benchmark of continual instruction tuning for multimodel large language models. Advances in neural information processing systems **37**, 57817–57840 (2024)
3. Flesch, T., Balaguer, J., Dekker, R., Nili, H., Summerfield, C.: Comparing continual task learning in minds and machines. Proceedings of the National Academy of Sciences **115**(44), E10313–E10322 (2018)
4. Guo, H., Zeng, F., Xiang, Z., Zhu, F., Wang, D.H., Zhang, X.Y., Liu, C.L.: Hidellava: Hierarchical decoupling for continual instruction tuning of multimodal large language model. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13572–13586 (2025)
5. He, J., Guo, H., Zhu, K., Tang, M., Wang, J.: Continual instruction tuning for large multimodal models. IEEE Transactions on Image Processing (2026)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
7. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
8. Kumari, P., Chauhan, J., Bozorgpour, A., Huang, B., Azad, R., Merhof, D.: Continual learning in medical image analysis: A comprehensive review of recent advancements and future prospects. Medical Image Analysis p. 103730 (2025)
9. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
10. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
11. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
12. Mirzadeh, S.I., Farajtabar, M., Pascanu, R., Ghasemzadeh, H.: Understanding the role of training regimes in continual learning. Advances in neural information processing systems **33**, 7308–7320 (2020)
13. Serra, J., Suris, D., Miron, M., Karatzoglou, A.: Overcoming catastrophic forgetting with hard attention to the task. In: International conference on machine learning. pp. 4548–4557. PMLR (2018)
14. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
15. Tian, K., Zhao, Z., Chen, Y., Ge, N., Cao, S., Han, X., Gu, J., Yu, S.: Domain-specific schema reuse supports flexible learning to learn in the primate brain. Nature Communications (2026)
16. Wang, G., Bai, L., Wang, J., Yuan, K., Li, Z., Jiang, T., He, X., Wu, J., Chen, Z., Lei, Z., et al.: Endochat: Grounded multimodal large language model for endoscopic surgery. Medical Image Analysis p. 103789 (2025)

17. Wang, G., Xiao, H., Zhang, R., Gao, H., Bai, L., Yang, X., Li, Z., Li, H., Ren, H.: Copesd: A multi-level surgical motion dataset for training large vision-language models to co-pilot endoscopic submucosal dissection. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 12636–12643 (2025)
18. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: Theory, method and application. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5362–5383 (2024)
19. Zhao, H., Zhu, F., Guo, H., Wang, M., Wang, R., Meng, G., Zhang, Z.: Mllm-cl: Continual learning for multimodal large language models. arXiv preprint arXiv:2506.05453 (2025)
20. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)