This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Dynamic Fusion of Multi-Layer Visual Features for Endoscopic Image Understanding

Changfei Wang[1 [0009−0007−7941−2591]] Xuguang Zhu[1 [0009−0003−9716−3593]] Xiankang Yu[1 [0009−0009−1886−5372]] Yuichiro Hayashi[1] Masahiro Oda[2,1 [0000−0001−7714−422X]] Kensaku Mori[1,2,3 [0000−0002−0100−4797]]

[1] Graduate School of Informatics, Nagoya University
kensaku@is.nagoya-u.ac.jp
[2] Information Technology Center, Nagoya University, Japan
[3] Research Center of Medical Bigdata, National Institute of Informatics, Japan

Abstract. Recent advances in multimodal large language models have demonstrated strong performance across diverse domains; however, medical image understanding remains challenging due to the fine-grained and domain-specific characteristics of surgical data. We propose a plug-and-play dynamic visual feature fusion module to enhance endoscopic scene understanding by improving the interaction between Vision Transformers (ViTs) and language models. The proposed approach dynamically modulates multi-layer visual features conditioned on textual queries and refines shallow and deep representations prior to projection into the language model. In quantitative evaluations, the proposed fusion modules achieved the highest BLEU-4 scores among the compared methods for scene description tasks. Furthermore, the fine-tuned model based on EndoViT outperforms the current state-of-the-art method EndoChat on the EndoVis-VQLA benchmark, while enabling a lighter-weight architecture to achieve performance comparable to existing state-of-the-art models.

Keywords: Vision-and-language · Surgical scene understanding · Vision Language Models.

1 Intoduction

Minimally invasive endoscopic surgery is widely adopted due to reduced trauma and faster recovery, yet the intraoperative environment remains highly complex, involving dynamic instrument–tissue interactions, occlusions, and subtle anatomical structures[8, 22]. Reliable scene understanding is therefore critical for Computer-assisted Intervention (CAI) system[19], which aim to improve surgical safety and workflow efficiency through accurate perception of instruments, procedural stages, and contextual cues.

Recent multimodal large language models (MLLMs) enable joint visual–text reasoning and have shown potential for surgical visual question answering and context-aware assistance[13]. In such systems, Vision Transformers (ViTs) serve

as the primary visual encoder, and the quality of their extracted representations directly affects downstream reasoning performance. However, current MLLMs typically rely on features from a single deep ViT layer, potentially overlooking fine-grained details that are clinically relevant in surgical scenarios[18].

In classical vision models, multi-level feature fusion is known to enhance both local detail sensitivity and global semantic understanding[14, 7, 21]. While hierarchical Transformers partially adopt this idea, standard ViT remains the dominant backbone in MLLMs due to its pretraining compatibility and stable cross-modal alignment[6]. Consequently, effective multi-layer feature fusion in ViT-based MLLMs remains underexplored, particularly for medical applications.

Given that early ViT layers preserve texture-rich structural information and deeper layers encode global semantics[11], we hypothesize that integrating multi-level representations can improve semantic completeness and fine-grained sensitivity in surgical scene understanding, thereby better supporting CAI systems. Motivated by these observations, we introduce a query-aware dynamic visual feature fusion module that adaptively reweights multi-layer Transformer representations via text-conditioned routing and residual refinement, enabling parameter-efficient enhancement of medical multimodal large language models. The main contributions are summarized as follows:

1. **Query-Conditioned Multi-Layer Dynamic Fusion with Residual Refinement**: We propose a lightweight, plug-and-play text-guided fusion module that dynamically modulates multi-layer Vision Transformer features conditioned on textual embeddings. To preserve high-level semantic integrity, the module adopts a residual refinement design that retains the deepest-layer representation while softly injecting modulated lower-level features.
2. **Systematic Analysis of Vision Backbones**: We conduct comprehensive experiments comparing multiple pretrained vision encoders in surgical scenarios and systematically analyze their impact on downstream performance.
3. **Empirical Validation of Multi-Level Visual Fusion**: Extensive experiments across multiple public endoscopic datasets demonstrate that multi-level visual feature fusion enhances semantic richness in visual representations, leading to improved holistic understanding of complex surgical scenes and better downstream reasoning performance.

2 Method

2.1 Layer-wise Visual Representation Analysis in surgical scene with pretrained ViTs

To design an effective multi-layer fusion strategy for surgical understanding, we first conduct a systematic layer-wise analysis of pretrained Vision Transformer (ViT) backbones in medical scenarios.

Recent work on multi-layer visual feature fusion systematically investigated layer selection strategies for multimodal large language models and suggested that selecting representative layers from shallow, middle, and deep stages yields

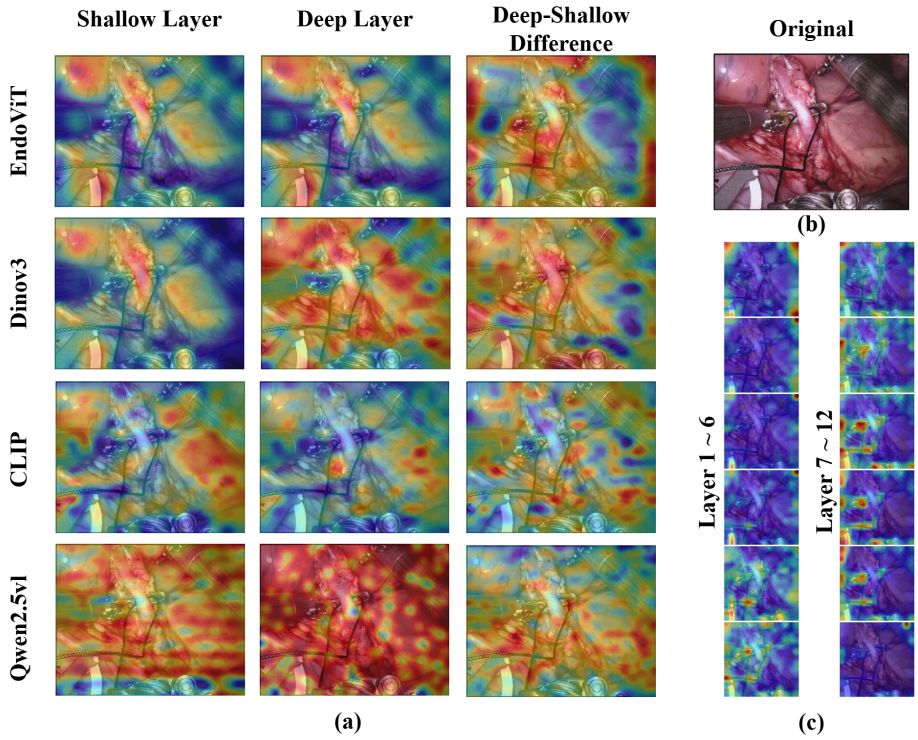


Fig. 1. Layer-wise Visual Representation Analysis. (a) Visualization of feature activation intensity for shallow and deep layers across various backbones. Red regions indicate higher feature response weights. The "Deep-Shallow Difference" column highlights the spatial cues lost in deeper stages. (b) The original surgical frame. (c) Visualization of the Attention Rollout across 12 EndoViT layers, illustrating the gradual loss of fine-grained geometric details (e.g., instrument contours) as the network depth increases.

strong generalization performance. To this end, we analyze four representative pretrained vision encoders on EndoVis2017 [2] and EndoVis2018 [1] datasets: CLIP ViT-L [16] (24 layers), Qwen2.5-VL [4] (32 layers), DINOv3 ViT-L [17] (24 layers), EndoViT [5] (12 layers, domain-specific pretrained). For each model, we extract intermediate features from shallow layers (layer 1 or 2) and deep layers (penultimate layer).

Patch-wise Feature Activation Visualization. We analyze feature activation distributions directly from the output token embeddings. Given the visual feature tensor from layer l:

$$M_i^{(l)} = \left\| F_i^{(l)} \right\|_2 \qquad F^{(l)} \in \mathbb{R}^{N \times C} \tag{1}$$

where i is the index of patches and N is the number of image patches. C is the channel dimension, and M is the patch-wise feature magnitude. This visualization reflects where the feature responses are strong in spatial regions, rather than attention weights.

To quantify inter-layer differences, we compute:

$$\Delta^{(l_s,l_d)} = \left\| \mathcal{M}^{(l_s)} - \mathcal{M}^{(l_d)} \right\| \tag{2}$$

where l_s denotes a shallow layer and l_d a deep layer.

Across Fig. 1, we observe a consistent pattern: 1. Shallow layers exhibit concentrated high-magnitude responses around surgical instruments and anatomical boundaries, indicating strong sensitivity to local textures and object contours. 2. Deep layers show more spatially diffused activation, expanding toward surrounding contextual regions.

This suggests that shallow layers encode high-detail structural cues, while deep layers emphasize global semantic aggregation.

Attention Rollout Analysis. To further analyze progressive attention propagation, we compute attention rollout:

$$\tilde{A}^{(l)} = \prod_{k=1}^{l} \bar{A}^{(k)} \tag{3}$$

where $\bar{A}^{(k)}$ denotes the normalized attention matrix of layer k integrated with identity connections to account for residual flow.

The rollout results confirm that shallow layers preserve localized attention patterns. Attention progressively spreads globally through self-attention propagation. This behavior is an inherent structural property of vanilla ViT due to global self-attention.

Special Case. Qwen2.5-VL exhibits a different activation pattern. Due to window-based attention mechanisms, hierarchical encoder design and optimization for high-resolution grounding tasks. Its shallow layers demonstrate extremely strong local discrimination ability. For grounding tasks, features must maintain high spatial separability to enable precise localization.

2.2 MLLM Framework with Text-Guided Fusion

Our framework is built upon the LLaVA architecture[15], which consists of a vision encoder, a projector, and a large language model. We experiment with multiple vision backbones: CLIP, Qwen2.5-VL, DINOv3, EndoViT, while keeping Vicuna-7B as the language model. The proposed MLLM architecture integrates a dynamic feature modulation mechanism to align visual representations with textual instructions as shown in Fig. 2.

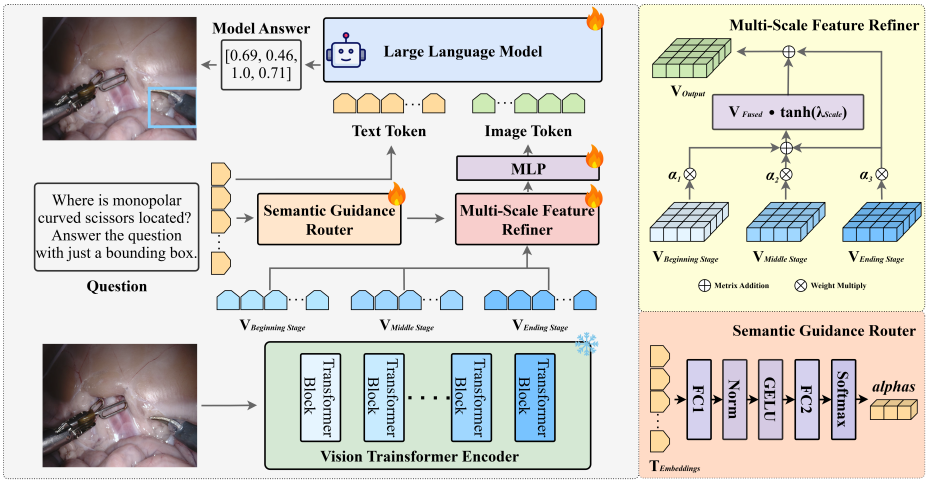


Fig. 2. The framework consists of a frozen Vision Transformer (ViT) Encoder, a Semantic Guidance Router, and a Multi-Scale Feature Refiner. Given a surgical image and a text-based question, the ViT extracts multi-level visual features $(V_{Beginning}, V_{Middle}, V_{Ending})$. The Router processes text embeddings to generate layer-specific weights (α), which guide the Refiner in adaptively aggregating multi-stage features to produce refined visual tokens for the Large Language Model (LLM).

Text-Guided Router. A lightweight neural network that processes text embeddings to generate layer-specific gate logits, determining the importance of intermediate visual features. The Router serves as a task-aware controller. It maps the text embedding $T \in \mathbb{R}^{D_{text}}$ to a set of weighting coefficients α via a bottleneck MLP structure.

$$\alpha = \text{Softmax}\left(\mathbf{W}_2 \cdot \text{GELU}\left(\text{LN}\left(\mathbf{W}_1 T + b_1\right)\right) + b_2\right) \tag{4}$$

Here, Softmax is employed to generate a normalized weight distribution across different layers, ensuring that $\sum \alpha_i = 1$. $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable linear projection layers. This set of weights, $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_{L-1}\}$, dynamically reflects the model's relative preference for different visual stages based on the textual context. This process identifies which hierarchical levels of the vision encoder (e.g., low-level textures vs. mid-level shapes) are most critical to the provided question through a competitive selection mechanism.

Multi-Scale Feature Refiner. A fusion module that adaptively aggregates weighted intermediate features and refines the final visual tokens through a learnable residual connection. The features from the first $L-1$ layers are linearly weighted and summed to produce the fused feature V_{Fused}:

$$V_{Fused} = \sum_{i=0}^{L-2} \alpha_i \cdot V_i \tag{5}$$

V_i represents the intermediate feature maps extracted from the Vision Transformer. Guided by the weights α_i, the model can selectively extract low-level textural information or mid-level geometric cues as needed.

To prevent the loss of fine-grained information in deep layers, the Refiner aggregates the first $L-1$ layers of the ViT. By employing a tanh-gated scale parameter λ_{Scale}, the module ensures a stable training trajectory.

$$V_{Output} = \tanh(\lambda_{Scale}) \odot V_{Fused} + V_{L-1} \tag{6}$$

At initialization, the module effectively acts as an identity mapping of the final layer, gradually incorporating multi-scale guidance as λ_{Scale} updates. The final output feature V_{Output} is generated by superimposing the weighted fused features onto the last-layer feature V_{L-1} after a non-linear scaling.

3 Experiments and Results

3.1 Datasets and Evaluation Metrics

The evaluation of our proposed framework is conducted across two benchmark datasets specialized for the surgical domain: Cholec80[9] and the EndoVis-VQLA[3], which includes both the 2017 and 2018 versions. For the Cholec80 dataset, we utilize a substantial collection of 21,591 surgical frames annotated with instrument labels and surgical phase descriptions to facilitate robust visual-text alignment. The EndoVis-VQLA dataset, sourced from the EndoVis-VQLA challenge, serves as the primary benchmark for fine-tuning and validation. This dataset is partitioned into EndoVis-18-VQLA (2,007 frames) and EndoVis-17-VQLA (97 frames), collectively providing 11.8k annotations.

To rigorously assess the model's comprehensive understanding of surgical scenes, the dataset is further categorized into four specialized tasks: single-phase recognition for stage identification, detection grounding for instrument localization, and both Region-Based and Visual-QA for assessing instrument states, surgical actions, and complex inter-object relationships. To measure performance across these diverse tasks, we employ a multi-faceted set of metrics: Accuracy and F1-score are utilized for recognition; Average Precision at 50% IoU (AP@50) and mean Intersection over Union (mIoU) are used for grounding; and BLEU-4, METEOR, ROUGE-L, and CIDEr are adopted to evaluate the quality of language generation in the QA tasks.

3.2 Implementation Details

The training procedure is executed in a decoupled two-stage paradigm to optimize both cross-modal alignment and task-specific reasoning.

In the first stage, we perform joint pre-training of the MLP projector, the Semantic Guidance Router, and the Multi-Scale Feature Refiner using the Cholec80 dataset. To bridge the gap between visual features and textual descriptions, we generate descriptive VQA pairs using a fixed template (e.g., "Please describe the

Table 1. Quantitative comparison of Visual QA and Region-based QA on the EndoVis-18 and EndoVis-17 datasets. Underlined model names indicate evaluation in a zero-shot setting. The bold values represent the best performance for each metric within each dataset.

Dataset	Model	Visual QA				Region Based QA			
		BLEU-4	CIDEr	METEOR	ROUGE-L	BLEU-4	CIDEr	METEOR	ROUGE-L
EndoVis-18	LLaVA-1.5[15]	0.0389	0.3421	0.1521	0.3085	0.0283	0.1528	0.1510	0.4027
	LLaVA-Med[12]	0.1402	1.3387	0.2032	0.3592	0.0394	0.1536	0.1665	0.3397
	InternVL-3[23]	0.0167	0.0824	0.1233	0.2909	0.0144	0.0428	0.0994	0.1472
	Qwen2.5VL[4]	0.0596	0.6327	0.1881	0.2995	0.0738	0.7154	0.1866	0.4436
	EndoChat[20]	0.5113	5.9802	0.3997	0.7874	0.5885	5.5939	0.4120	0.8019
	EndoViT-Fusion	**0.5144**	5.9236	**0.4049**	**0.7935**	**0.6267**	**6.3504**	**0.4443**	**0.8355**
	DINOv3-Fusion	0.5105	**6.1394**	0.4030	0.7928	0.6053	6.0757	0.4267	0.8135
EndoVis-17	LLaVA-1.5[15]	0.0431	0.3328	0.1693	0.2935	0.0642	0.1829	0.2167	0.4692
	LLaVA-Med[12]	0.1633	1.5221	0.2312	0.3583	0.0844	0.1725	0.1823	0.4821
	InternVL-3[23]	0.0451	0.5291	0.1482	0.2901	0.0152	0.1437	0.1021	0.3127
	Qwen2.5VL[4]	0.1142	1.0526	0.2015	0.3208	0.0792	0.1830	0.2089	0.4902
	EndoChat[20]	0.2579	2.9993	0.2734	0.5682	0.2531	2.9213	0.2544	**0.5421**
	EndoViT-Fusion	0.3077	3.2771	0.2925	0.5920	**0.2760**	**3.1715**	**0.2610**	0.5419
	DINOv3-Fusion	**0.3231**	**3.6859**	**0.2942**	**0.6046**	0.2690	2.8720	0.2596	0.5361

surgical scene") based on existing annotations, allowing the router to learn preliminary weight selection strategies. This pre-training stage is conducted with a learning rate of 1×10^{-3}, a batch size of 32, and is trained for a single epoch.

In the second stage, we conduct joint fine-tuning of the Large Language Model using Low-Rank Adaptation (LoRA)[10] on the EndoVis-VQLA dataset. This stage utilizes 26.2k training pairs to refine the model across various surgical tasks, followed by evaluation on an 8.6k testing set. The LoRA hyperparameters are configured with a rank r of 128 and an alpha α of 256. For optimization, we apply a fine-tuning learning rate of 2×10^{-4} for the LLM and a dedicated projector learning rate of 2×10^{-5}. All experiments are implemented on two NVIDIA A100 (80GB PCIe) GPUs with a batch size of 16 for the second stage, ensuring the model achieves convergence within a single epoch.

3.3 Comparison Evaluation

We further evaluate the proposed method on the EndoVis-VQLA benchmark under zero-shot settings using several general-purpose multimodal large language models, including Qwen2.5-VL, LLaVA-1.5, LLaVA-Med, and InternVL-3. We compare their performance with the proposed fusion-enhanced models and the current state-of-the-art surgical MLLM, EndoChat, on Region-Based QA and Visual-QA tasks. Table 1 shows that the EndoViT-Fusion model, built upon a comparatively lightweight vision backbone and language model, achieves performance comparable to or exceeding EndoChat. These findings indicate that effective multi-layer visual feature fusion can compensate for model scale differences and improve surgical scene understanding without relying on large-scale architectures.

Table 2. Ablation study on the effect of the proposed module (w/ and w/o) across different backbones. (w/) and (w/o) denote the model with and without our specific design, respectively. The bold values indicate the better performance within each experimental group (EndoViT or DINOv3).

Dataset	Model	Single Phrase		Grounding		Visual		Region Based	
		Acc	F-score	AP@50	mIoU	BLEU-4	ROUGE-L	BLEU-4	ROUGE-L
EndoVis-18	EndoViT(w/o)	**0.6496**	**0.4207**	**0.8263**	**0.6677**	0.4969	0.7635	0.5995	0.8089
	EndoViT(w/)	0.5511	0.3400	0.8055	0.6491	**0.5144**	**0.7935**	**0.6267**	**0.8355**
	DINOv3(w/o)	**0.5912**	**0.3012**	**0.8112**	**0.6675**	0.4977	0.7632	0.5927	0.7966
	DINOv3(w/)	0.5163	0.2325	0.7999	0.6398	**0.5105**	**0.7928**	**0.6053**	**0.8135**
EndoVis-17	EndoViT(w/o)	**0.4663**	**0.3881**	0.7098	0.6079	0.2922	0.5852	0.2490	0.5342
	EndoViT(w/)	0.4385	0.3355	**0.7372**	**0.6103**	**0.3077**	**0.5920**	**0.2760**	**0.5419**
	DINOv3(w/o)	**0.3912**	**0.2631**	0.7128	0.5972	0.3020	0.5819	0.2612	0.5211
	DINOv3(w/)	0.3730	0.2102	**0.7137**	**0.5995**	**0.3231**	**0.6046**	**0.2690**	**0.5361**

3.4 Ablation Study

We conduct ablation studies on both DINOv3 and EndoViT backbones to evaluate the effect of the proposed fusion module. Specifically, we compare models trained with and without the fusion module under identical settings. The results in Table 2 shows that incorporating the fusion module improves performance on Region-Based QA and Visual-QA tasks, indicating enhanced semantic reasoning and contextual representation. However, a slight decrease is observed in grounding and single-phase recognition performance. This phenomenon suggests that while multi-layer feature fusion enriches semantic attributes of visual representations, the integration of shallow features may introduce boundary-level noise, potentially affecting precise object localization and stage discrimination. These findings highlight the trade-off between semantic enrichment and spatial precision in multi-layer feature integration.

4 Discussion and Conclusion

Results on multiple endoscopic benchmarks show that the proposed multi-layer fusion improves semantic representation and contextual understanding across different ViT backbones. By integrating hierarchical features, the model captures complementary information beyond single-layer tokens. However, shallow feature integration may introduce minor boundary noise, indicating a trade-off between semantic enrichment and spatial precision.

The effectiveness of fusion also depends on the representational gap between shallow and deep layers in pretrained ViTs. In summary, we present a text-guided fusion framework that enhances visual token representation for surgical MLLMs. Future work will explore finer-grained text–patch interactions to further improve spatial accuracy in clinical scenarios.

Acknowledgments. Thanks to the Make New Standards (MNS) program, and the Mori Laboratory for their help and advice. This work was partially supported by the MEXT/JSPS KAKENHI (24H00720, 24K03262, 24K15206), JST CREST (Grant Number JPMJCR20D5) and JST Moonshot (Grant Number JPMJMS2214).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
2. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
3. 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)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
5. Batić, D., Holm, F., Özsoy, E., Czempiel, T., Navab, N.: Endovit: pretraining vision transformers on a large collection of endoscopic images. International Journal of Computer Assisted Radiology and Surgery **19**(6), 1085–1091 (2024)
6. Carolan, K., Fennelly, L., Smeaton, A.F.: A review of multi-modal large language and vision models. arXiv preprint arXiv:2404.01322 (2024)
7. Chen, G., Shen, L., Shao, R., Deng, X., Nie, L.: Lion: Empowering multimodal large language model with dual-level visual knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26540–26550 (2024)
8. Fuchs, K.: Minimally invasive surgery. Endoscopy **34**(02), 154–159 (2002)
9. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
11. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022)
12. 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)
13. Liang, C.X., Tian, P., Yin, C.H., Yua, Y., An-Hou, W., Ming, L., Song, X., Wang, T., Bi, Z., Liu, M.: A comprehensive survey and guide to multimodal large language models in vision-language tasks. arXiv preprint arXiv:2411.06284 (2024)
14. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
15. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)

16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
17. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
18. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9568–9578 (2024)
19. Vercauteren, T., Unberath, M., Padoy, N., Navab, N.: Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions. Proceedings of the IEEE **108**(1), 198–214 (2019)
20. 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. arXiv preprint arXiv:2501.11347 (2025)
21. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint arXiv:2408.04840 (2024)
22. Zaffino, P., Moccia, S., De Momi, E., Spadea, M.F.: A review on advances in intraoperative imaging for surgery and therapy: imagining the operating room of the future. Annals of Biomedical Engineering **48**(8), 2171–2191 (2020)
23. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv preprint arXiv:2504.10479 (2025)