SurgOnAir: Hierarchy-Aware Real-Time Surgical Video Commentary

Jingyi He*[1], Yue Zhou*[1,2], Long Bai[4],
Kun Yuan[1,2,3], Nassir Navab[1,2], and Yuan Bi[1,2]

[1] Computer Aided Medical Procedures (CAMP),
TU Munich, Germany
yue.zhou@tum.com
[2] Munich Center for Machine Learning (MCML), Munich, Germany
[3] University of Strasbourg, France
[4] The Chinese University of Hong Kong, Hong Kong

Abstract. Understanding surgical workflow in real time is fundamental for intelligent surgical embodiment, where AI systems continuously perceive and respond as surgery proceeds. In the operating room, critical decisions depend on subtle, moment-to-moment changes, such as fine instrument movements and evolving tissue states, where even slight perceptual delays can limit assistance or compromise safety. Yet existing methods remain offline or operate at coarse temporal scales, generating descriptions only after processing clips, preventing immediate reaction. We address this by proposing SurgOnAir, a streaming vision-language model that processes frames sequentially without future access and progressively generates narration tokens as visual input arrives. SurgOnAir achieves fine-grained frame-to-token generation, enabling instant responsiveness to evolving surgical dynamics. Built upon our curated hierarchical dataset SurgOnAir-11k spanning action-, step-, and phase-level supervision, the model is trained to produce multi-level textual responses that reflect the inherent hierarchy of surgical procedures. Furthermore, special transition tokens are generated to explicitly mark state changes, allowing SurgOnAir to capture and signal key workflow transitions as they occur. Experiments show that SurgOnAir enables real-time understanding through a single vision-language model that unifies streaming across multiple hierarchies of the surgical workflow, generating superior and hierarchy-aware narrations. Code and dataset will be public.

1 Introduction

Surgical procedures are inherently dynamic, evolving through continuous and multi-stage transitions. As surgical robotics advances toward higher autonomy, the surgical AI systems must develop temporally grounded understanding to perceive the environment and make informed real-time decisions [12, 8]. Recent multimodal large language models (MLLMs) [2, 6] have significantly enhanced multimodal perception and reasoning across both general and surgical

* Equal contribution.

Table 1. Comparison of surgical workflow analysis models in hierarchical and streaming capability.

	Hierarchy			Caption	
Dataset	Phase	Step	Action	Offline	Stream
EndoNet [16]	✓	✗	✗	✗	✗
Gastric Bypass [9]	✓	✓	✗	✗	✗
Rendezvous [14]	✗	✗	✓	✗	✗
DAISI [15]	✗	✗	✗	✓	✗
HecVL [20]	✓	✓	✓	✗	✗
LLaVA-Surg [10]	✗	✗	✗	✓	✗
SurgVidLM [17]	✓	✓	✓	✗	✗
Ours	✓	✓	✓	✓	✓

domains. However, their computationally heavy backbones and offline processing paradigms fundamentally constrain their use in real-time, temporally grounded surgical workflow understanding scenarios.

In the natural video domain, MLLMs have achieved remarkable success in comprehending complex spatial-temporal sequences, as evidenced by pioneering frameworks like Video-ChatGPT [13], Chat-UniVi [7], and Video-LLaVA [11]. Driven by these breakthroughs, researchers have rapidly migrated such paradigms into the surgical domain for understanding the dynamics of surgical scenes. LLaVA-Surg [10] curates a large-scale dataset to train a multimodal conversational assistant for surgical video question-answering. Later, SurgVidLM [17] addresses both holistic and fine-grained surgical video comprehension by proposing a two-stage mechanism that fuses global procedural context with local details. Despite their strong performance, most existing surgical video understanding methods are designed under an offline setting. They assume access to the entire video sequence and perform captioning or question answering retrospectively, limiting their applicability in real-time surgical scenarios.

Recent efforts have begun exploring online video understanding. VideoLLM-Online [4] introduces a Learning-In-Video-Stream framework that models video understanding as continuous streaming dialogue, enabling temporally aligned real-time interaction. Building on this direction, LiveCC [5] proposes a timestamp-aligned streaming training scheme that densely interleaves Automatic Speech Recognition (ASR) words with video frames, facilitating fine-grained speech-vision alignment. More recently, StreamingVLM [18] focuses on memory-efficient inference for near-infinite video streams via a compact KV-cache with short-term vision and long-term text windows. However, modeling them as flat sequences overlooks the inherent hierarchical organization which is essential for surgical procedures understanding since surgical workflow is inherently hierarchical, spanning phases, steps, and actions.

Hierarchical modeling is crucial in surgical workflows, as hierarchical structure provides contextual grounding that improves cross-context disambiguation and reduces hallucination. HecVL [20] introduces a hierarchical video-language

pretraining framework that pairs surgical videos with multi-level textual supervision, and learns disentangled representations across fine-to-coarse hierarchies. Similarly, PeskaVLP [19] proposes a hierarchical knowledge-augmented pretraining strategy that integrates refined textual supervision with dynamic temporal alignment for workflow-aware representation learning. However, these approaches focus on hierarchical representation learning in offline pretraining settings and are not formulated as generative, streaming models for surgical video narration.

Therefore, it is essential to design a generative live-streaming narration model that explicitly incorporates hierarchical structure inherent in surgical procedures for narration. As illustrated in Tab. 1, existing surgical understanding models lack either the support to complete Phase-Step-Action hierarchy or for both offline and streaming captioning. To this end, we introduce SurgOnAir, a hierarchy-aware streaming surgical narration model capable of real-time generation. To achieve this, we propose the SurgOnAir-11K, a hierarchically paired video-language dataset, where surgical clips are aligned with temporally grounded narration at the word level and enriched with temporally annotated hierarchical phase and step labels. This design enables temporally aligned hierarchical supervision for live-streaming surgical narration. By incorporating hierarchical supervision as structured state constraints, SurgOnAir learns to capture workflow transitions and cross-granularity dynamics across phases, steps, and actions.

Our contributions are threefold: (1) We design a hierarchical temporal grounding pipeline that converts raw surgical videos and ASR transcripts into multi-level, temporally aligned video-language supervision, yielding SurgOnAir-11K. (2) We introduce SurgOnAir, a hierarchical streaming framework which interleaves multi-modal visual-text tokens for real-time processing, while leveraging specialized transition tokens to explicitly predict state changes for hierarchical-aware narration. (3) Extensive experiments demonstrate that our approach outperforms both existing offline and streaming models, achieving superior real-time narration quality and enhanced hierarchical awareness at critical procedural transitions.

2 Method

2.1 Hierarchical Temporal Grounding Data Curation Pipeline

Overview. To generate visually-grounded video-language supervision, we introduce a carefully designed automatic pipeline to process online surgical videos into an interleaved hierarchically paired vision-language training corpus. Our pipeline (Fig. 1) operates sequentially: it first utilizes audio-to-text transcription with precise timestamps, then corrects domain-specific errors to ensure descriptive accuracy. Since online surgical videos inherently contain unaligned meta-discourse, we explicitly filter out this commentary so the text strictly captures observable content. Finally, we leverage metadata to construct a structured hierarchical organization. We detail each module of this pipeline below.

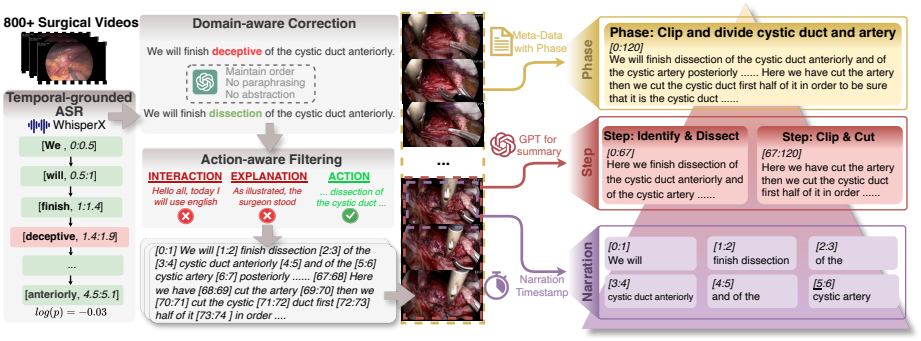

Fig. 1. Overview of the hierarchical temporal grounding pipeline for narration data curation. Raw videos and ASR transcripts are refined through domain-aware correction and action-aware filtering to retain visually relevant narration while preserving temporal alignment. Phase and step annotations are constructed to produce a hierarchically-structured and temporally aligned dataset.

Temporally Grounded ASR. We employ WhisperX-large [3] to extract word-level transcripts from surgical videos. WhisperX performs forced alignment to produce temporally precise word-level timestamps from the audio stream. Each word is represented as a triplet $w_i = \langle \text{word}_i,\ t_i^{\text{start}},\ t_i^{\text{end}} \rangle$, where t_i^{start} and t_i^{end} denote the start and end times of the word_i, and an avarage probability $log(p)$ is reported for every sentence. This temporally grounded representation enables fine-grained synchronization between spoken narration and video frames.

Domain-aware ASR Correction We flag sentences with $log(p) < -0.15$ (equivalent to a probability threshold of 0.865) as low-confidence, and correct them using a strictly constrained rewriting procedure via GPT-4o [1]. The system is restricted to terminology normalization without inserting or deleting content. Modifications are limited strictly to correcting recognition errors and medical terminology to minimize unnecessary intervention and maintain data integrity.

Action-aware Filtering. Building upon the data curation setup in [5], we ensure that the textual descriptions are strictly grounded in the visual content. Since surgical narration contains heterogeneous discourse, we prompt GPT-4o [1] to classify each sentence into one of three categories: *ACTION*, *EXPLANATION*, and *INTERACTION*. We preserve only the *ACTION* sentences for supervision, since *EXPLANATION*, and *INTERACTION* lack direct semantic alignment with the video frames (see the example in Fig. 1).

Hierarchical Data Construction To facilitate hierarchical surgical narration training, we enriched the dataset with an explicit three-level structure: Phase → Step → Word. We began with fine-grained *ACTION* sentences which provide word-level temporal grounding. These groups were then processed by a large language model to generate 1–3 high-level procedural steps per segment, ensuring a temporally monotonic and non-overlapping partition of the narration. Finally,

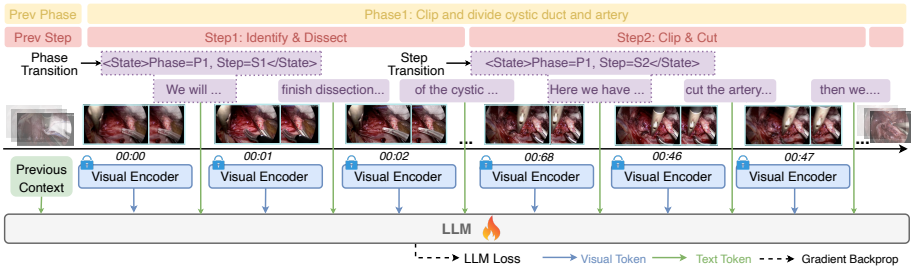

Fig. 2. Overview of our hierarchical streaming training architecture. Visual tokens extracted from temporally aligned frames are interleaved with narration tokens in a causal LLM. Structured `<State>` tokens explicitly encode Phase and Step states, marking hierarchical state transitions along the timeline.

we integrated these steps with phase definitions and temporal boundaries from video metadata to complete the hierarchical structure.

2.2 SurgOnAir Streams Surgical Workflows Hierarchically

As illustrated in Fig. 2, our proposed model leverages a multimodal Large Language Model (MLLM) to perform real-time surgical video narration. To effectively synchronize the stream of visual frames with real-time speech, we adopt the dense temporal interleaving strategy introduced in [5]. Instead of relying on traditional post-hoc captioning, the interleaved input sequence is formulated as:

$$\begin{aligned}[\text{Con}]\ \langle F_{t:t+k}\rangle\ \langle W_{t:t+k}\rangle\ \langle F_{t+k:t+2k}\rangle\ \langle W_{t+k:t+2k}\rangle \\ \dots\ \langle F_{t+nk:t+(n+1)k}\rangle\ \langle W_{t+nk:t+(n+1)k}\rangle\ ,\end{aligned} \tag{1}$$

where [Con] denotes the preceding contextual metadata (e.g., initial prompts or historical context shown as "Previous Context" in Fig. 2), $\langle F\rangle$ denotes visual frames extracted by the visual encoder, $\langle W\rangle$ represents ASR and generated text tokens, t is the temporal index, and k is the stride.

To enforce hierarchy-aware surgical narration, we incorporate hierarchical procedural supervision via special `<State>` tokens. Specifically, the text sequence $\langle W_{t:t+k}\rangle$ is not merely a flat stream of words, instead it dynamically encapsulates both the continuous procedure narration and the discrete structural workflow boundaries. We formulate $\langle W_{t:t+k}\rangle$ as a conditional concatenation:

$$\langle W_{t:t+k}\rangle = \begin{cases} \langle S_t\rangle \oplus \langle N_{t:t+k}\rangle, & \text{if a state transition occurs within } [t, t+k) \\ \langle N_{t:t+k}\rangle, & \text{otherwise} \end{cases} \tag{2}$$

where $\oplus$ denotes sequence concatenation, and $\langle N_{t:t+k}\rangle$ represents the continuous text narration. The state sequence $\langle S_t\rangle$ acts as a discrete workflow anchor (e.g., `<State>Phase=P1, Step=S1</State>`), injecting critical hierarchical context exactly at the moment of phase or step transitions. This dual-level formulation enables the model to simultaneously act as a continuous narrator and a strict state tracker.

Training Details We adopt the pre-trained LiveCC-7B-Base [5] with streaming capability as the backbone. During fine-tuning, the LLM is updated while the vision encoder remains frozen. To enable fine-grained alignment and reduce redundancy, we set the stride $k = 1$ and the frame rate to 2 FPS. The model is trained with a learning rate of $2e - 5$. To enhance visual grounding and reduce over-reliance on textual context, we adopt a staged conditioning strategy: in the first 2 epochs, the model is conditioned on previous ASR text to stabilize training and maintain temporal coherence. In the final epoch, the ASR context is removed and only the video title is provided, forcing the model to rely more directly on visual evidence.

3 Experiment

Dataset We partition the SurgOnAir-11K dataset into training and test sets with an 80/20 split, employing stratification by surgical meta-type to ensure a representative distribution of diverse procedural categories and maintain balanced complexity across both sets.
Baseline We evaluate our method against two categories of baselines: offline Video LLMs, including LLaVA-Video-7B [11] and Qwen2.5-VL-7B [2], and the streaming-based LiveCC-7B [5]. While the offline models follow a conventional captioning protocol by processing all frames simultaneously, LiveCC-7B serves as our primary competitive baseline for real-time, frame-by-frame generation. Notably, during evaluation, we provide only the video title as context to force the model to rely on salient visual cues rather than textual continuation from prior ASR transcripts.
Evaluation Protocal We formulate evaluation as a pairwise choice task for GPT-4o [1], acting as an LLM-as-a-judge. Conditioned on the ground-truth ASR transcripts, the judge is prompted to explicitly select the superior narration between two models. We report the win rate [5], defined as the percentage of wins against the other model. To align with the practical requirements of streaming surgical narration, the evaluation paradigm prioritizes semantic correctness, specifically focusing on the precise grounding of anatomical and procedural content rather than syntactic perfection.

3.1 Real-time Narration Results

We adopt Hulu-Med [6] as a fixed comparison model, as its extensive medical pretraining has demonstrated strong effectiveness in surgical applications [6]. Therefore, all pairwise evaluations are conducted between the evaluated model and Hulu-Med [6].
Discussion. Shown in Tab. 2, offline video-language models show limited effectiveness in surgical captioning. Despite having access to the entire video offline, LLaVA-Video-7B [11] and Qwen2.5-VL-7B [2] achieve win-rate scores of only 11.3% and 6.2%, respectively, indicating a significant lack of domain-specific

Table 2. Comparison of offline and streaming models for captioning

Model	Live?	Win Rate
Hulu-Med-7B [6]	✗	❀
LLaVA-Video-7B [11]	✗	11.3
Qwen2.5-VL-7B [2]	✗	6.2
LiveCC-7B [5]	✓	16.7
SurgOnAir-base	✓	60.4
SurgOnAir	✓	**66.1**

Table 3. Pairwise win-rate ablation of hierarchy, version, and phase correctness.

Study	Model	Win Rate
Hierarchy	SurgOnAir-base	39.4
	SurgOnAir	**60.6**
$\langle W \rangle$ Formulation	SurgOnAir-v1	42.5
	SurgOnAir	**57.5**
Phase Correctness	SurgOnAir-base	34.2
	SurgOnAir	**65.8**

surgical understanding. In contrast, streaming models perform substantially better: LiveCC-7B [5] already surpasses all offline baselines with a score of 16.7%. However, generic streaming pretraining remains insufficient for complex surgical scenarios. Training exclusively on in-domain surgical data using standard flat narration *SurgOnAir-base* markedly improves the performance to 60.4%. Building upon this, our final model *SurgOnAir*, which incorporates hierarchical-aware modeling to capture both overarching procedural states and granular narrations, further elevates the score to 66.1%. These results clearly demonstrate the crucial importance of both domain-specific data and hierarchical-structured supervision for accurate, real-time surgical narration.

3.2 Ablation Results

For our ablation study, we perform pairwise comparisons to isolate the contribution of each designed component. The win rate reflects the relative preference between two competing variants via GPT-4o. Tab. 3 presents a structured analysis focusing on three key aspects: hierarchy modeling, $\langle W \rangle$ formulations, and phase correctness.

Hierarchy. We evaluate the impact of hierarchical modeling by comparing our full model (*SurgOnAir*) against *SurgOnAir-base*, a baseline trained exclusively on flat narration from the surgical dataset without any hierarchical structure. The inclusion of our hierarchical design yields a substantial improvement, elevating the win rate from 39.4% to 60.6%. This gap indicates that hierarchical modeling provides essential structural guidance, enabling more coherent narratives compared to standard flat modeling.

$\langle W \rangle$ **Formulations.** To validate our specific design for the state formulation, we compare our final model against an earlier variant, *SurgOnAir-v1*. Unlike our proposed formulation that properly models state transitions, the v1 variant strictly predicts the state during every single word token generation step $\langle W_{t:t+k} \rangle = \langle S_t \rangle \oplus \langle N_{t:t+k} \rangle$. Forcing the model to generate state information for every short interval introduces excessive redundant overhead. Consequently, this modeling causes the model to over-rely on state prediction, inadvertently neglecting the most critical textual narration. Our *SurgOnAir* model achieves a

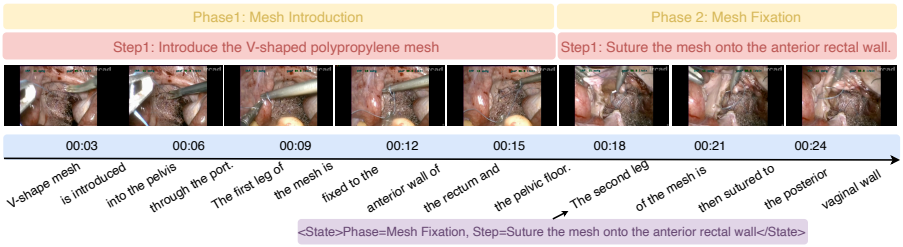

Fig. 3. Qualitative result of SurgOnAir.

57.5% win rate against the 42.5% of the *SurgOnAir-v1* variant, confirming that a transition-aware $\langle W \rangle$ formulation is crucial for generating fluent and accurate narration.

Phase Correctness. We study the specific contribution of accurate phase grounding. For this evaluation, we specifically select the test cases where *SurgOnAir* correctly predicts the surgical phase, and compare its generated narrations against those produced by *SurgOnAir-base* on the exact same videos. Under this strictly controlled setting, *SurgOnAir* decisively outperforms the base model with a 65.8% to 34.2% win rate. This large margin explicitly demonstrates that successfully predicting the procedural phase provides strong, direct support for generating accurate and contextually aligned narration.

3.3 Qualitative Results

Fig. 3 presents a qualitative example of our hierarchical streaming narration. Benefiting from explicit hierarchical modeling, the narration transitions coherently across procedural boundaries. As illustrated, after detailing the actions within the *Mesh Introduction* phase (e.g., "V-shape mesh is introduced into the pelvis..."), the model accurately predicts the state transition at exactly 00:18. This structural anchor naturally guides the continuous word stream into the subsequent *Mesh Fixation* phase (e.g., "The second leg of the mesh is then sutured..."), ensuring structured temporal continuity. Beyond tracking macro-level phase transitions, this hierarchical conditioning enables strict temporal alignment at the micro-level. The model accurately recognizes and describes key instruments (e.g., the V-shaped mesh) and relevant anatomical structures (e.g., the pelvic floor) precisely at the moment of interaction. Such comprehensive behavior demonstrates genuine temporally aligned narration, effectively moving away from the limitations of isolated, frame-level captioning.

4 Conclusion

In this work, we present the first real-time surgical narration framework built upon hierarchical modeling. Our approach produces temporally aligned and

hierarchical-aware descriptions, advancing real-time surgical understanding beyond conventional video captioning. This work lays an important foundation for future surgical automation and intelligent surgical teaching systems. Limitations include dataset scale and the absence of future action prediction, both of which offer promising directions for further exploration.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bain, M., Huh, J., Han, T., Zisserman, A.: Whisperx: Time-accurate speech transcription of long-form audio. arXiv preprint arXiv:2303.00747 (2023)
4. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: Videollm-online: Online video large language model for streaming video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18407–18418 (2024)
5. Chen, J., Zeng, Z., Lin, Y., Li, W., Ma, Z., Shou, M.Z.: Livecc: Learning video llm with streaming speech transcription at scale. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29083–29095 (2025)
6. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
7. Jin, P., Takanobu, R., Zhang, W., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13700–13710 (2024)
8. Kim, J.W., Chen, J.T., Hansen, P., Shi, L.X., Goldenberg, A., Schmidgall, S., Scheikl, P.M., Deguet, A., White, B.M., Tsai, D.R., et al.: Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning. Science robotics **10**(104), eadt5254 (2025)
9. Lavanchy, J.L., Ramesh, S., Dall'Alba, D., Gonzalez, C., Fiorini, P., Müller-Stich, B.P., Nett, P.C., Marescaux, J., Mutter, D., Padoy, N.: Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. International Journal of Computer Assisted Radiology and Surgery (May 2024). https://doi.org/10.1007/s11548-024-03166-3, http://dx.doi.org/10.1007/s11548-024-03166-3
10. Li, J., Skinner, G., Yang, G., Quaranto, B.R., Schwaitzberg, S.D., Kim, P.C., Xiong, J.: Llava-surg: towards multimodal surgical assistant via structured surgical video learning. arXiv preprint arXiv:2408.07981 (2024)
11. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
12. Long, Y., Lin, A., Kwok, D.H.C., Zhang, L., Yang, Z., Shi, K., Song, L., Fu, J., Lin, H., Wei, W., Chen, K., Chu, X., Hu, Y., Yip, H.C., Chiu,

P.W.Y., Kazanzides, P., Taylor, R.H., Liu, Y., Chen, Z., Wang, Z., Au, S.K.W., Dou, Q.: Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. Science Robotics **10**(104), eadt3093 (2025). https://doi.org/10.1126/scirobotics.adt3093
13. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
14. Nwoye, C.I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T., Jia, F., Yang, Y., Wang, H., et al.: Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. Medical Image Analysis **86**, 102803 (2023)
15. Rojas-Muñoz, E., Couperus, K., Wachs, J.: Daisi: Database for ai surgical instruction (2020), https://arxiv.org/abs/2004.02809
16. Twinanda, A., Shehata, S., Mutter, D., Marescaux, J., Mathelin, M.D., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging **36** (02 2016). https://doi.org/10.1109/TMI.2016.2593957
17. Wang, G., Wang, J., Mo, W., Bai, L., Yuan, K., Hu, M., Wu, J., He, J., Huang, Y., Padoy, N., et al.: Surgvidlm: Towards multi-grained surgical video understanding with large language model. arXiv preprint arXiv:2506.17873 (2025)
18. Xu, R., Xiao, G., Chen, Y., He, L., Peng, K., Lu, Y., Han, S.: Streamingvlm: Real-time understanding for infinite video streams. arXiv preprint arXiv:2510.09608 (2025)
19. Yuan, K., Navab, N., Padoy, N., et al.: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. Advances in Neural Information Processing Systems **37**, 122952–122983 (2024)
20. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 306–316. Springer (2024)