SurgΣ: A Spectrum of Large-Scale Multimodal Data and Foundation Models for Surgical Intelligence

Zhitao Zeng[1*] **Mengya Xu**[2*] **Jian Jiang**[3*] **Pengfei Guo**[4*] **Yunqiu Xu**[1] **Zhu Zhuo**[1]
Chang Han Low[1] **Yufan He**[4] **Dong Yang**[4] **Chenxi Lin**[3] **Yiming Gu**[3] **Jiaxin Guo**[2]
Yutong Ban[3†] **Daguang Xu**[4†] **Qi Dou**[2†] **Yueming Jin**[1†]
[1]NUS [2]CUHK [3]SJTU [4]NVIDIA
Project page: `https://SurgSigma.github.io`

Abstract

Surgical intelligence has the potential to improve the safety and consistency of surgical care, yet most existing surgical AI frameworks remain task-specific and struggle to generalize across procedures and institutions. Although multimodal foundation models, particularly multimodal large language models, have demonstrated strong cross-task capabilities across various medical domains, their advancement in surgery remains constrained by the lack of large-scale, systematically curated multimodal data. To address this challenge, we introduce SurgΣ, a spectrum of large-scale multimodal data and foundation models for surgical intelligence. At the core of this framework lies SurgΣ-DB, a large-scale multimodal data foundation designed to support diverse surgical tasks. SurgΣ-DB consolidates heterogeneous surgical data sources (including open-source datasets, curated in-house clinical collections and web-source data) into a unified schema, aiming to improve label consistency and data standardization across heterogeneous datasets. SurgΣ-DB spans 6 clinical specialties and diverse surgical types, providing rich image- and video-level annotations across 18 practical surgical tasks covering understanding, reasoning, planning, and generation, at an unprecedented scale (over 5.98M conversations). Beyond conventional multimodal conversations, SurgΣ-DB incorporates hierarchical reasoning annotations, providing richer semantic cues to support deeper contextual understanding in complex surgical scenarios. We further provide empirical evidence through recently developed surgical foundation models built upon SurgΣ-DB, illustrating the practical benefits of large-scale multimodal annotations, unified semantic design, and structured reasoning annotations for improving cross-task generalization and interpretability.

1 Introduction

According to estimates from the Lancet Commission, more than 300 million surgical procedures are performed worldwide each year [49], underscoring the urgent demand for safer and more accessible surgical care. Despite advances in minimally invasive [26, 50] and robotic techniques [25, 32], surgery remains inherently complex, requiring continuous interpretation of dynamic anatomy and high-stakes decision-making under uncertainty. Surgical AI is therefore emerging as a transformative paradigm, acting as an intelligent collaborator that enhances perception, understanding, and reasoning. By leveraging multimodal intraoperative signals (*e.g.*, visual streams, textual instructions, robotic kinematics, and preoperative imaging), AI systems promise to improve safety, reduce variability, and broaden access to high-quality surgical expertise. However, most prior surgical AI systems remain narrowly designed for isolated tasks, including phase recognition [71, 40], tool or tissue

[*]These authors contributed equally to this work. [†]Corresponding authors.

Preprint.

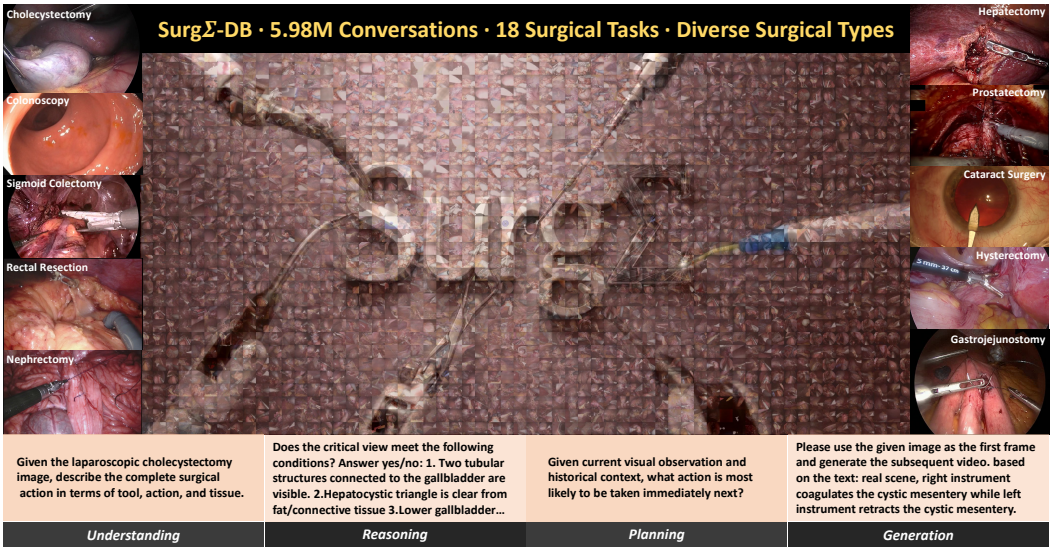


Figure 1: SurgΣ-DB is a large-scale multimodal data foundation for surgical intelligence.

segmentation [6, 5], and action classification [57, 7], often within a tailored task or a single surgical type. This task-specific paradigm limits knowledge transfer and leads to brittle generalization, where models degrade under distribution shifts caused by differences in imaging systems, anatomy, or surgical styles.

Foundation models, particularly multimodal large language models [70, 68, 9, 8], have recently emerged as a promising paradigm for unified visual perception, language understanding and multimodal reasoning, enabling models to jointly interpret visual content and reason with natural language. In principle, such models offer a unified framework capable of supporting a broad spectrum of surgical tasks, ranging from describing anatomical structures and instrument states to answering intraoperative queries, summarizing procedural context, and generating interpretable decision-support rationales. While foundation models have achieved remarkable success across domains such as radiology [84, 69, 12], pathology [17, 23, 73], and molecular biology [38, 1], their application to the surgical domain remains comparatively underexplored. Surgery poses fundamentally distinct challenges for training multimodal foundation models. Intraoperative scenes are not only visually complex (*e.g.*, severe occlusion, tissue deformation, and rapid camera motion), but also exhibit strong spatiotemporal structure and causal interdependence, where subtle instrument–tissue interactions can induce irreversible anatomical changes. Clinically relevant cues are often fine-grained, transient, and context-dependent, demanding long-horizon temporal reasoning and precise spatial grounding beyond static image understanding. Furthermore, variability across institutions, surgeons, devices, and patient anatomies introduces substantial distribution shifts that hinder generalization.

We observe that a fundamental obstacle to advancing surgical multimodal foundation models lies in the lack of large-scale, high-quality, and systematically curated multimodal data. Conventional surgical datasets [72, 7, 81] are predominantly vision-centric and designed under a closed-set paradigm, providing only predefined categorical annotations (*e.g.*, surgical phase or instrument tags) while lacking practical natural language instruction–visual pairs that better reflect real-world clinical usage and flexible interaction. In addition, these datasets are typically limited in scale and surgical-type diversity, as they are often confined to a small number of procedures or institutions. Consequently, **they remain fragmented across tasks and modalities, with inconsistent annotation standards and heterogeneous label spaces that hinder cross-dataset integration and large-scale training**. Although some recent works [20, 56, 58] have introduced datasets for surgical foundation models, they still exhibit notable limitations in scale, diversity, and task coverage, as summarized in Table 1. On the other hand, annotation quality and granularity remain insufficient: the absence of a unified label space can mislead training and weaken generalization, while existing datasets largely lack high-quality multi-step reasoning traces.

To facilitate research, we present SurgΣ, a spectrum of large-scale multimodal data and foundation models for surgical intelligence. At its core, SurgΣ-DB serves as a unified multimodal data foundation

designed to enable large-scale training of surgical foundation models. Rather than releasing isolated task-specific datasets, our SurgΣ-DB systematically consolidates diverse surgical data sources into a unified and well-structured foundation. We curate data spanning multiple surgical specialties and procedures, covering diverse clinical departments and operative types to ensure broad anatomical and procedural variability. **SurgΣ-DB combines open-source resources with web-collected surgical videos, and employs a semi-automated annotation pipeline integrating expert human labeling and controlled synthesis** to ensure both real-world representativeness and scalable clinical fidelity. Within the same surgical scenes, **SurgΣ-DB provides rich annotations with hierarchical reasoning traces**, enabling multi-grained spatial understanding as well as temporal modeling. To the best of our knowledge, **SurgΣ-DB provides one of the most comprehensive task coverages in surgical intelligence**, spanning understanding, reasoning, planning, and generative capabilities through richly structured and multi-level annotations, and **is constructed at an unprecedented scale** (*i.e.*, $\sim$5.98M) across diverse surgical types from 6 clinical specialties. Importantly, all data are organized under a unified format, facilitating interoperable training, cross-task integration, and future extensibility, while promoting more consistent label spaces across heterogeneous datasets.

Building upon SurgΣ-DB, a family of surgical foundation models (*i.e.*, BSA [87], SurgVLM [95], Surg-R1 [33], and Cosmos-H-Surgical [31]) are developed, which empirically demonstrate the effectiveness of the proposed unified data spectrum. BSA [87] demonstrates that fundamental surgical actions exhibit consistent, recognizable patterns across anatomically diverse procedures, enabling cross-specialty generalization without domain-specific adaptation and supporting clinically meaningful downstream applications including skill assessment and procedural planning. SurgVLM [95] demonstrates that large-scale multimodal instruction-tuning data can enhance cross-task generalization, enabling a single model to effectively handle diverse surgical understanding tasks through shared vision–language training. Surg-R1 [33] further illustrates the critical role of structured reasoning annotation, where multi-step inference traces significantly strengthen grounded surgical understanding. Cosmos-H-Surgical [31] demonstrates that surgical world models can transform large-scale unlabeled surgical video into actionable training data for robot policy learning by synthesizing realistic surgical scenes and recovering pseudo-kinematics through inverse dynamics inference, thereby enabling scalable vision–language–action training with limited real demonstrations and significantly improving policy performance and sample efficiency. Together, these models provide complementary evidence that scale, semantic unification, and chain-of-thought reasoning annotations are key ingredients for advancing surgical foundation models within a coherent data-centric framework.

We hope that our data foundation and preliminary findings will inspire the research community to further explore and unlock the untapped potential of multimodal foundation models in surgical intelligence, ultimately advancing clinically reliable and generalizable surgical AI systems. In future work, we will continually expand SurgΣ-DB in data scale and diversity, and progressively enrich each surgical scene with comprehensive and holistic annotations toward full task coverage within unified surgical contexts. In summary, our contributions can be summarized as follows:

- **A large-scale multimodal surgical data foundation.** We introduce SurgΣ-DB, a large-scale multimodal dataset spanning multiple surgical specialties and procedures. The dataset integrates image- and video-level annotations across understanding, reasoning, planning, and generation tasks, providing the most comprehensive task coverages in surgical intelligence.
- **Comprehensive and unified multi-granular annotations.** We consolidate heterogeneous data sources into a unified data schema with a consistent label space. A semi-automated annotation pipeline combining human labeling and controlled synthesis with hierarchical reasoning annotations enables semantic coherence and scalable foundation model training.
- **Empirical validation through foundation models.** A family of surgical foundation models is developed upon SurgΣ-DB, providing empirical validation of key data design principles and demonstrating the impact of large-scale multimodal data foundation.

2 Related Work

2.1 Surgical Datasets and Benchmarks

Surgical AI has long been supported by a rich ecosystem of public datasets that enable the development and evaluation of data-driven models. Conventional datasets [72, 7, 81, 18] typically provide closed-

Table 1: Comparison with existing multimodal surgical datasets and benchmarks.

Datasets	Visual Modality		Conversation Type			Data Source			Reasoning	#Sample	#Task
	Video	Image	VQA	Caption	Generation	In-House	Open-Source	Internet			
Cholec80-VQA [65]	✘	✔	✔	✘	✘	✘	✔	✘	✘	43K	3
EndoVis2018-VQA [65]	✘	✔	✔	✘	✘	✘	✔	✘	✘	11.78K	3
PSI-AVA-VQA [64]	✘	✔	✔	✘	✘	✘	✔	✘	✘	10.29K	3
PitVQA [30]	✘	✔	✔	✘	✘	✘	✔	✘	✘	884.24K	5
LRSP-VQA [16]	✘	✔	✔	✘	✘	✘	✘	✔	✘	1.13K	4
CoPESD [78]	✘	✔	✔	✘	✘	✘	✔	✘	✘	121.09K	3
EndoVQA-Instruct [44]	✘	✔	✔	✘	✘	✔	✔	✘	✘	446.54K	12
Surg-396K [76]	✘	✔	✔	✔	✘	✘	✔	✘	✘	396K	7
SurgPub-Video [42]	✔	✘	✔	✘	✘	✘	✔	✔	✘	48.52K	3
SurgMLLMBench [21]	✘	✔	✔	✘	✘	✘	✔	✘	✘	893.15K	5
SurgLaVi [55]	✘	✔	✔	✘	✘	✔	✔	✘	✘	239.8K	4
SurgVeo [19]	✔	✔	✘	✘	✔	✘	✔	✘	✘	50	1
SVU-31K [77]	✔	✘	✔	✘	✘	✘	✔	✔	✘	31K	4
SurgCoTBench [45]	✘	✔	✔	✘	✘	✘	✘	✔	✔	14.25K	5
SUREON [56]	✔	✘	✔	✘	✘	✘	✔	✘	✔	206.8K	12
SurgΣ-DB (ours)	✔	✔	✔	✔	✔	✔	✔	✔	✔	5.98M	18

set categorical labels designed for supervised learning of scene understanding (*e.g.*, instrument recognition) and workflow understanding (*e.g.*, phase recognition). In parallel, large-scale pre-training datasets [15, 83, 7, 22, 85] offer abundant unlabeled or weakly labeled data for representation learning, but their lack of structured annotations limits their effectiveness for fine-grained surgical understanding and multimodal reasoning. While effective for benchmarking perception and workflow analysis, these datasets remain limited to predefined category labels and fail to capture complex interactions and reasoning over surgical activities, leaving a considerable gap between these datasets and real-world surgical applications. Their procedure-centric design also leads to limited cross-procedure generalization and heterogeneous label spaces, posing challenges for unified model training.

With the emergence of multimodal foundation models, recent efforts have shifted toward instruction-following and multi-task datasets tailored to generalizable multimodal surgical modeling. Early VQA-style datasets [65, 65, 64, 91] are typically constructed by converting single-task annotations into question–answer pairs, but remain limited in annotation richness and task diversity. Subsequent works [44, 55] improve coverage by aggregating multiple open-source datasets, while more recent efforts [42, 45] leverage web-sourced videos to scale multimodal supervision. In addition, various benchmarks have been developed to evaluate multimodal comprehension in surgical scenarios [76, 21, 59, 45], surgical scene generation [19], interpretability [20], and surgical quality assessment [3, 10]. Despite this progress, existing datasets remain limited in diversity and annotation quality, are biased toward VQA-style conversations, and lack support for dense prediction, spatiotemporal reasoning, planning, and generative tasks. Moreover, heterogeneous annotation schemas hinder unified multi-task training, motivating a unified, large-scale dataset for multimodal surgical intelligence.

2.2 Foundation Models for Surgical Intelligence

Surgical AI has traditionally relied on task-specific models for instrument/tissue detection and segmentation [90, 43], workflow analysis [71, 35, 36] and triplet recognition [18, 52, 53, 54], which are typically trained with procedure-specific supervision and are sensitive to domain shifts across clinical environments. With the increasing availability of large-scale surgical video data, recent efforts have moved toward surgical foundation models that learn transferable visual representations via self-supervised [79, 62, 94, 93, 89] or weakly supervised pre-training [39]. These models demonstrate improved robustness and cross-domain generalization, and can be adapted to diverse downstream tasks through lightweight fine-tuning. However, such approaches remain primarily perception-centric and lack the ability to support open-ended reasoning, interactive understanding, and high-level decision-making.

Building upon these advances, multimodal foundation models extend foundation modeling by enabling natural-language interaction and reasoning over surgical scenes. Early surgical vision-language models focus on VQA-style formulations [65, 30], representing surgical elements such as instruments, tissues, and spatial relationships through textual descriptions and generating answers. More recently, instruction-tuned surgical multimodal large language models [75, 34, 64, 76, 86, 41,

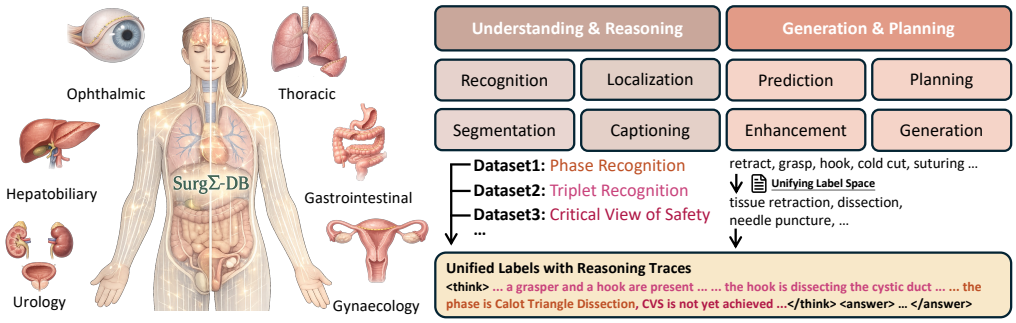


Figure 2: SurgΣ-DB integrates heterogeneous surgical data across 6 clinical specialties into a unified multimodal data foundation. It supports diverse tasks through standardized annotations enriched with hierarchical reasoning traces.

95] and video-level surgical understanding models [77, 80, 92] demonstrate strong performance across diverse image- and video-level surgical tasks. However, deploying multimodal foundation models in surgical settings remains challenging due to fragmented data resources that lack scale, diversity, unified label spaces, and high-quality reasoning annotations. This highlights the need for large-scale, consistently annotated, and unified multimodal datasets for training and evaluation.

3 SurgΣ-DB: A Large-Scale Multimodal Data Foundation for Surgical AI

SurgΣ-DB is a large-scale multi-grained dataset constructed for multimodal foundation models in surgical intelligence. It contains ∼5.98M annotated samples spanning 6 clinical specialties, as shown in Figure 2. SurgΣ-DB integrates rich annotations for both static video frames and video clips, and aligned natural language conversations including instruction–response pairs and reasoning traces. Data are sourced from both publicly available surgical datasets and curated in-house clinical collections. Annotations are produced through a combination of expert human labeling and controlled synthesis pipelines, which further introduce hierarchical reasoning annotations to capture contextual relationships within surgical scenes, ensuring annotation quality, semantic consistency, and scalable coverage. All subsets are organized under a unified data schema with harmonized label spaces and standardized formats to support multi-task training and benchmarking.

3.1 Data Curation

3.1.1 Multi-Source Surgical Data Collection

Our goal is to construct a highly diverse surgical data foundation spanning multiple clinical specialties and surgical types. Guided by this objective, we collect raw data from a wide range of sources, including publicly available surgical datasets, online surgical videos, and curated in-house clinical collections developed in collaboration with medical partners. As summarized in Table 2, we collect 16 surgical types across 6 major clinical specialties (*i.e.*, gynecologic [81, 63], ophthalmic [27, 46], hepatobiliary [61, 2, 53, 48, 71, 74], gastrointestinal [40, 37, 47, 13], urologic [6, 5, 51, 14, 7, 57, 11] and thoracic [28] surgeries), encompassing both robotic and manual operations under diverse imaging modalities, including laparoscopic, endoscopic, OphScope, and thoracoscopic settings. This breadth ensures substantial variability in anatomy, instrumentation, and workflow dynamics, providing a highly diverse foundation for surgical intelligence modeling.

3.1.2 Holistic and Multi-Granular Surgical Task Design

We design a diverse and multi-granular set of tasks in SurgΣ-DB to comprehensively cover the key objectives of surgical intelligence, as demonstrated in Figure 3. These tasks are organized into two complementary groups: (1) **Understanding and Reasoning** and (2) **Planning and Generation**, collectively reflecting the fundamental capabilities required for surgical multimodal foundation models, spanning perception, reasoning, predictive modeling, and controllable content generation.

Understanding and Reasoning Tasks. These tasks encompass a diverse set of perception and reasoning problems designed to capture multi-granular spatio-temporal understanding of surgical scenes. They involve grounding surgical instruments, anatomical structures, and procedural dynamics

Table 2: Data sources integrated into SurgΣ-DB, categorized by clinical specialty and surgical types.

Clinical Specialty	Surgical Type	Data Source	Surgical Platform	Protocol
Gynecologic	Hysterectomy	AutoLaparo [81]	Non-Robotic	Laparoscopy
		SurgicalActions160 [63]	Non-Robotic	Laparoscopy
		Web-Collected Data	Both	Laparoscopy
Ophthalmic	Cataract Surgery	Cataract-1K [27]	Non-Robotic	OphScope
		CaDISv2 [46]	Non-Robotic	OphScope
Hepatobiliary	Cholecystectomy	Cholec80 [72]	Non-Robotic	Laparoscopy
		Cholec80-CVS [61]	Non-Robotic	Laparoscopy
		CholecInstanceSeg [2]	Non-Robotic	Laparoscopy
		CholecT50 [53]	Non-Robotic	Laparoscopy
		Endoscapes [48]	Non-Robotic	Laparoscopy
		M2CAI16 [71]	Non-Robotic	Laparoscopy
		HeiChole [74]	Non-Robotic	Laparoscopy
		Web-Collected Data	Both	Laparoscopy
		In-House Data	Non-Robotic	Laparoscopy
	Hepatectomy	In-House Data	Non-Robotic	Laparoscopy
Gastrointestinal	Gastrectomy	Web-Collected Data	Both	Laparoscopy
		In-House Data	Non-Robotic	Laparoscopy
	Gastrojejunostomy	MultiBypass140 [40]	Non-Robotic	Laparoscopy
	Colonoscopy	SegCol [37]	Non-Robotic	Endoscopy
	Proctocolectomy	HeiCo [47]	Non-Robotic	Laparoscopy
	Sigmoid Colectomy	HeiCo [47]	Non-Robotic	Laparoscopy
	Rectal Resection	HeiCo [47]	Non-Robotic	Laparoscopy
	Ladd's Procedure	Web-Collected Data	Non-Robotic	Laparoscopy
	Appendectomy	Web-Collected Data	Non-Robotic	Laparoscopy
	Rectal Resection/Extirpation	DSAD [13]	Robotic-Assisted	Laparoscopy
Urologic	Nephrectomy	EndoVis2017 [6]	Robotic-Assisted	Laparoscopy
		EndoVis2018 [5]	Robotic-Assisted	Laparoscopy
		Nephrec9 [51]	Non-Robotic	Laparoscopy
		SurgT [14]	Robotic-Assisted	Laparoscopy
		Web-Collected Data	Both	Laparoscopy
		In-House Data	Both	Laparoscopy
	Prostatectomy	GraSP [7]	Robotic-Assisted	Laparoscopy
		SAR-RARP [57]	Robotic-Assisted	Laparoscopy
		MESAD-Real [11]	Robotic-Assisted	Laparoscopy
		Web-Collected Data	Robotic-Assisted	Laparoscopy
Thoracic	Lobectomy	Lobectomy Dataset [28]	Robotic-Assisted	Thoracoscopic

from both image and video inputs, enabling detailed analysis of tool–tissue interactions and surgical workflows. By covering capabilities such as geometric modeling, semantic interpretation, safety verification, and contextual workflow understanding, these tasks reflect both foundational perception challenges and clinically relevant reasoning problems in surgical environments.

- **Instrument Recognition:** Identify surgical instruments present in a video frame, serving as a fundamental perceptual capability for understanding tool usage and surgical workflow.
- **Instrument Localization:** Predict spatial regions of surgical instruments using either bounding boxes or image patches, enabling precise spatial grounding of tool positions.
- **Instrument Segmentation:** Generate pixel-wise masks of surgical instruments to capture fine-grained shapes and tool–tissue occlusion relationships for precise spatial modeling.
- **Tissue and Organ Recognition:** Classify visible anatomical entities (*i.e.*, tissues or organs) to establish semantic awareness of operative regions and contextual surgical states.
- **Tissue and Organ Localization:** Localize anatomical entities (*i.e.*, tissues and organs) via bounding boxes to provide spatial awareness of anatomical structures during surgery.
- **Phase Recognition:** Identify the current surgical phase from either video frames or video clips, recognizing the highest-level procedural stage of the surgical workflow.
- **Step Recognition:** Identify finer-grained surgical steps from either video frames or video clips, modeling intermediate-level workflow progression within each phase.

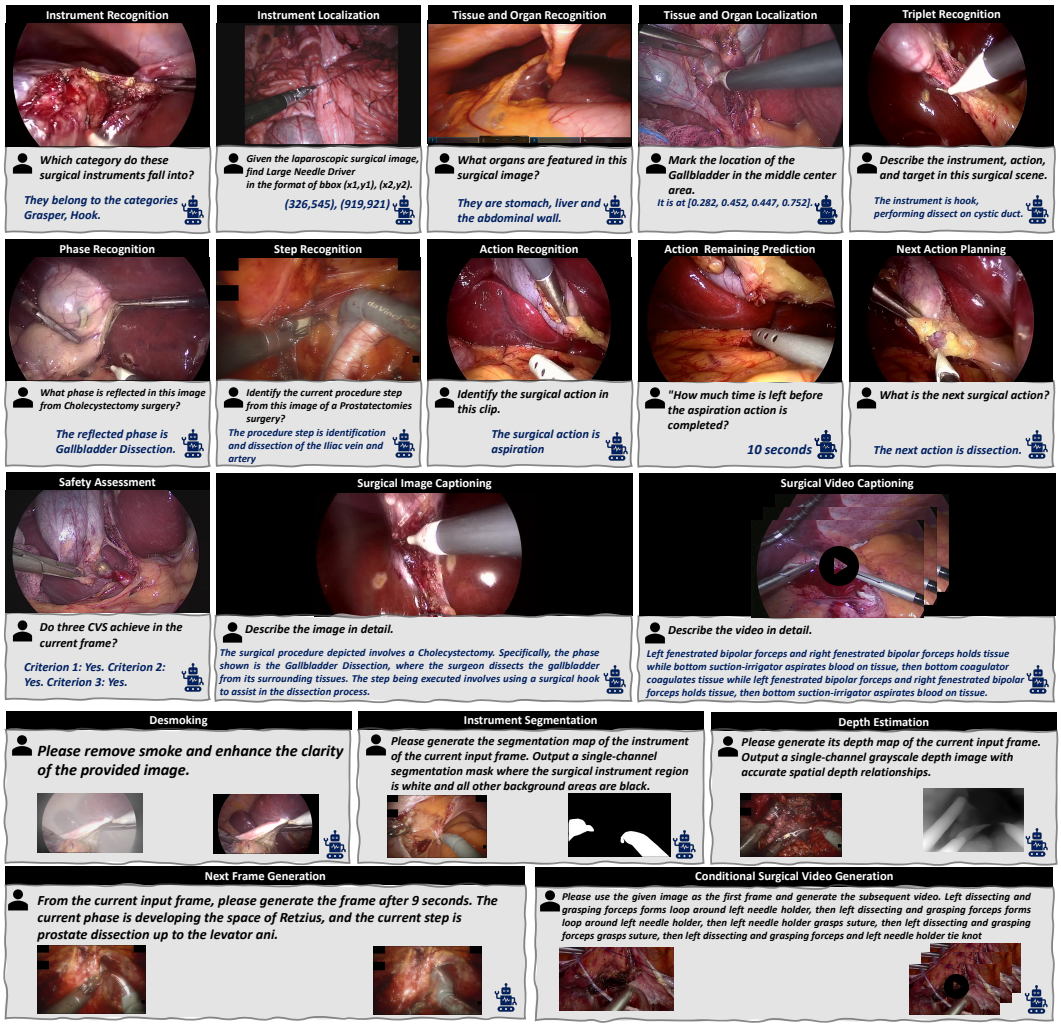


Figure 3: SurgΣ-DB contains diverse multimodal conversations spanning 13 understanding and reasoning tasks as well as 5 planning and generation tasks, supporting a wide range of perception, reasoning, simulation, and decision-oriented capabilities for surgical intelligence.

- **Action Recognition:** Classify atomic surgical actions from either video frames or video clips, encompassing both basic surgical actions and procedure-specific operations, and capturing the lowest-level dynamics of the surgical workflow.
- **Triplet Recognition:** Predict instrument–action–target triplets to explicitly represent structured surgical interactions between instruments and targets (*i.e.*, tissues or other instruments), enabling higher-level understanding and reasoning about surgical activities.
- **Depth Estimation:** Infer relative depth maps for surgical video frames, facilitating geometry-aware downstream tasks such as instrument navigation and spatial perception during surgery.
- **Safety Assessment:** Determine whether safety-critical anatomical structures have been sufficiently exposed and correctly identified (*e.g.*, critical view of safety [61, 3, 10] in cholecystectomy), ensuring surgical actions such as clipping or transection can be performed safely.
- **Surgical Image Captioning:** Generate natural language descriptions of static surgical scenes, summarizing instruments, anatomy, and contextual information visible in inputs.
- **Surgical Video Captioning:** Produce temporally coherent textual descriptions of surgical videos, capturing the spatio-temporal evolution of surgical scenes and activities.

Planning and Generation Tasks. Beyond scene comprehension, these tasks focus on predictive modeling and controllable generation that are closely aligned with practical clinical needs. They

Table 3: Unified annotation format of SurgΣ-DB.

```
{
  "meta data":
    {
      "info": "dataset information",
    },

  "images": [
      {
        "id": 0,
        "source dataset": "dataset name",
        "source url": "url to data source",
        "clinical specialty": "name of the specialty",
        "surgical type": "name of the surgical procedure",
        "image path": "path to image"
      }
    ],

  "videos": [
      {
        "id": 0,
        "source dataset": "dataset name",
        "source url": "url to data source",
        "clinical specialty": "name of the specialty",
        "surgical type": "name of the surgical procedure",
        "video path": "path to video"
      }
    ],

  "annos": [
      {
        "id": 0,
        "videos": [video id],
        "time step": [frame idx],
        "images": [image id],
        "question": "user instructions",
        "thinking": "model thinking response (optional)",
        "answer": "model answer response",
        "dense prediction": ["path to dense prediction"],
        "task type": ["task name"],
        "gt label": ["ground-truth label"],
        "annotator": {"NUS", "CUHK", "SJTU", "NVIDIA"}
      },
    ]
}
```

involve forecasting future procedural developments, estimating surgical progress, and generating coherent multimodal descriptions of surgical activities. Such tasks provide a basis for applications including intraoperative assistance, workflow anticipation, and simulation-driven learning, highlighting the potential of surgical AI systems to move from passive observation toward proactive support.

- **Action Remaining Prediction:** Estimate the remaining duration of the current action, enabling workflow progress assessment and anticipation of upcoming procedural transitions.
- **Next Action Planning:** Predict the most probable next surgical action given current visual observation and historical context, enabling anticipation of upcoming workflow transitions.
- **Desmoking:** Generate smoke-free surgical video frames from smoke-degraded observations, improving visual clarity and perceptual robustness under real-world imaging conditions.
- **Next Frame Prediction:** Generate a future video frame at a specified time horizon to predict the temporal evolution of surgical scenes, supporting workflow anticipation and simulation.

Table 4: Three typical annotation samples for MLLM training.

```
[
  {
    "id": 0,
    "images": ["image path"],
    "conversations": [
      {
        "from": "human",
        "value": "<image>user instruction for image tasks"
      },
      {
        "from": "gpt",
        "value": "<thinking>model thinking response</thinking>\n
                  <answer>model answer response</answer>"
      }
    ]
  },

  {
    "id": 1,
    "videos": ["video path"],
    "conversations": [
      {
        "from": "human",
        "value": "<image>user instruction for video tasks"
      },
      {
        "from": "gpt",
        "value": "<thinking>model thinking response</thinking>\n
                  <answer>model answer response</answer>"
      }
    ]
  },

  {
    "id": 2,
    "images": ["image path (input)", "image path (prediction)"],
    "conversations": [
      {
        "from": "human",
        "value": "<image>user instruction for dense prediction tasks"
      },
      {
        "from": "gpt",
        "value": "<image>"
      }
    ]
  }
]
```

- **Conditional Surgical Video Generation:** Generate surgical video clips conditioned on textual instructions and visual context, including text-to-video, image-to-video, and text-guided image-to-video generation, enabling controllable simulation and data augmentation.

3.1.3 From Heterogeneous Labels to Unified Annotations

Surgical datasets collected from diverse sources often exhibit inconsistent terminology, mismatched category definitions, and varying annotation formats, especially for fine-grained units such as atomic surgical actions. These discrepancies arise from divergent naming conventions, variations in semantic granularity, and heterogeneous representation forms (*e.g.*, categorical labels, dense masks, or question–answer pairs with multi-step reasoning), leading to semantic drift and instability in large-scale joint training. To address this, we reorganize heterogeneous labels into a unified framework that aims

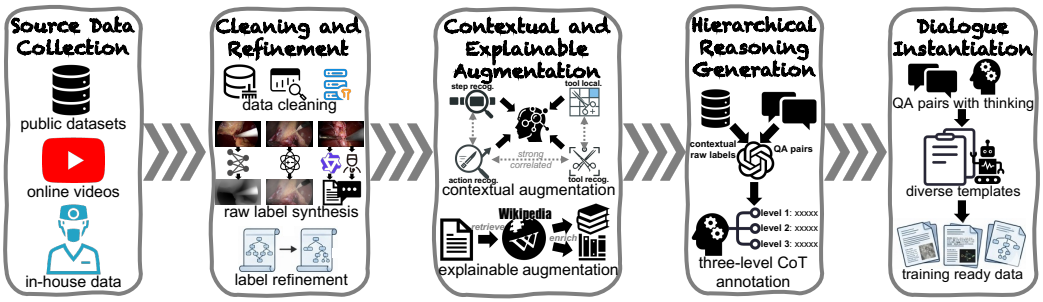


Figure 4: Overview of the data curation and annotation pipeline for SurgΣ-DB.

to standardize fine-grained semantic definitions, reduce cross-dataset inconsistencies, and support scalable, interoperable training.

Taking action recognition as a representative example, we consolidate diverse atomic actions into a unified taxonomy of ten basic surgical actions with explicit semantic boundaries and inclusion criteria. Each action is grounded in clinically interpretable descriptions and aligned with related attributes (*e.g.*, instruments and target tissues), forming a coherent and structured label space that supports consistent supervision across heterogeneous surgical data.

Beyond semantic normalization, we unify heterogeneous annotation formats under a consistent structural schema. Different annotation types and tasks are reorganized into a standardized representation that aligns multi-granular supervision across image- and video-level data, as illustrated in Table 3. This unified structure can be readily converted into training-ready multimodal format (see Table 4) compatible with large language models, enabling direct integration into foundation model pipelines. Such standardized annotation format ensures structural consistency, facilitates large-scale joint training, and supports future dataset scaling and extensibility.

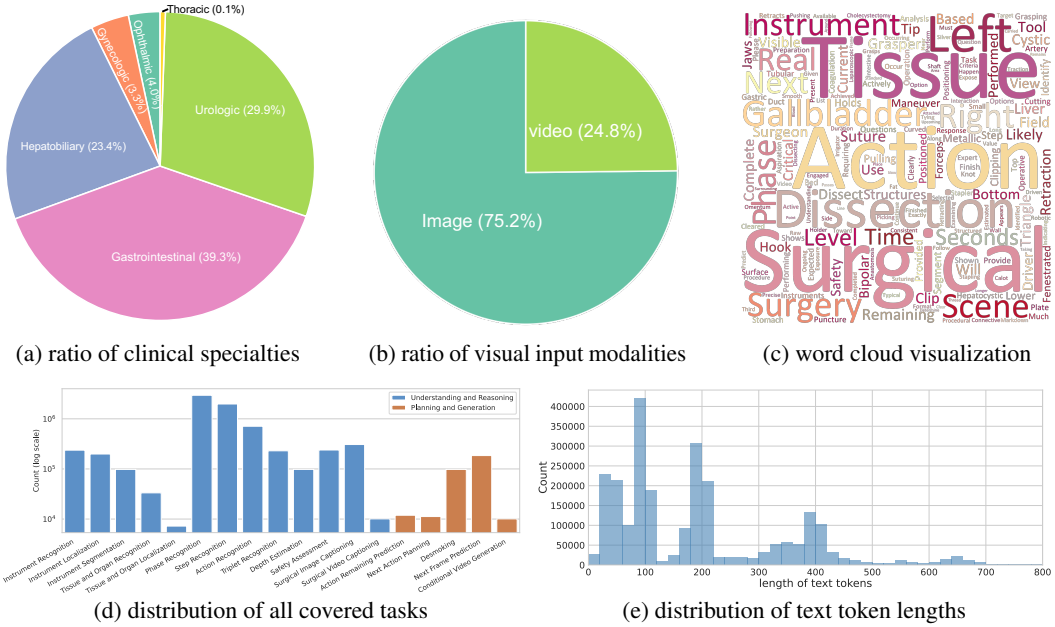
3.1.4 Semi-Automated Annotation Pipeline

Data Pre-Processing and Label Refinement. Following aforementioned unified annotation standards, we systematically refine raw labels from heterogeneous sources to ensure semantic and structural consistency. Coarse categories, institution-specific shorthand, and inconsistent terminology are replaced with explicit, context-complete descriptions and normalized under standardized medical vocabularies. For instance, ambiguous placeholders (*e.g.*, "other") are reformulated into precise statements, and synonymous anatomical terms (*e.g.*, "Calot's triangle" *vs.* "Hepatocystic triangle") are consolidated into canonical forms. This standard-driven refinement reduces cross-dataset ambiguity and improves stability in large-scale joint training.

For the raw annotations that are already provided by the original source datasets, we directly adopt the official labels to preserve dataset fidelity. For incomplete textual annotations (*e.g.*, image and video captioning), we leverage Qwen3-VL-235B [8] to generate enriched and context-consistent descriptions. For missing dense predictions, such as smoke masks, segmentation maps, and depth annotations, we employ off-the-shelf methods [29, 60, 88] to automatically generate the corresponding information. For some noise-prone labels (*e.g.*, temporal boundaries of surgical actions), we perform manual verification and label refinement to ensure precise and reliable annotations.

Contextual and Explainable Augmentation. Surgical attributes such as phase, step, instrument, and action are inherently interdependent. To enhance structural learning, we optionally consolidate correlated attributes into unified prompts that explicitly encode hierarchical and relational dependencies (*e.g.*, instrument–action pairs), transforming isolated labels into structured multi-attribute supervision. In addition, to strengthen vision-language alignment, we augment original categorical labels with knowledge-grounded descriptions derived from authoritative medical sources (*e.g.*, Wikipedia), linking visual evidence to surgical intent and anatomical context. Together, these contextual and explanation-aware augmentations promote compositional reasoning, improve fine-grained alignment, enhance robustness across diverse scenes, and increase interpretability for real-world deployment.

Hierarchical Reasoning Trajectory Generation. To explicitly align visual evidence with structured reasoning processes, a three-level chain-of-thought [82] annotation strategy is constructed to decompose surgical inference into perceptual grounding, relational understanding, and contextual reasoning,



(a) ratio of clinical specialties (b) ratio of visual input modalities (c) word cloud visualization

(d) distribution of all covered tasks (e) distribution of text token lengths

Figure 5: Statistical analysis of the constructed SurgΣ-DB.

as shown in Figure 4. Specifically, the first level emphasizes describing fundamental visual elements present in the scene, focusing on the appearance and spatial properties of tools and tissues. The second level focuses on characterizing interactions among different elements, capturing how tools, tissues, and actions relate to and influence one another through contact, motion, and functional associations. The third level abstracts perceptual and relational cues into high-level procedural inference, enabling structured reasoning about the global surgical context.

These structured reasoning trajectories are synthesized using GPT-5.1 [68] conditioned on all verified raw labels (from existing datasets or annotated by ourselves) under explicit forward-reasoning constraints, ensuring that intermediate steps remain grounded in observable visual evidence and preventing hallucinated interactions beyond the annotated scene. Consequently, the hierarchical CoT annotations transform flat supervision into structured reasoning guidance, promoting compositional inference, mitigating shortcut learning, and enhancing interpretability.

Conversational Diversity Expansion. To prepare multimodal foundation model training data, structured annotations are transformed into diverse conversational formats. Rigid template design may lead to overfitting in MLLMs, as the model can memorize superficial linguistic patterns rather than learning underlying visual–semantic alignment. To mitigate this issue, we construct 100–200 dialogue templates with controlled variations in phrasing, ordering, and information density. This strategy increases linguistic diversity while preserving logical consistency, thereby improving instruction-following robustness and cross-task generalization.

3.2 Dataset Analysis and Discussions

3.2.1 Data Statistics

SurgΣ-DB contains 5.98M multimodal conversations spanning both image- and video-based surgical tasks, extracted from 1.59K unique surgical video sources. Among them, 4.49M conversations are associated with images, while 1.48M correspond to video clips, forming a large-scale multimodal corpus for surgical intelligence. In total, SurgΣ-DB comprises 471.29M text tokens, providing rich linguistic corpus for multimodal learning. It covers 6 clinical specialties and 16 surgical procedure types, capturing diverse surgical scenes and procedural contexts across a wide range of operative domains, as illustrated in Figure 5a.

In terms of visual data, SurgΣ-DB comprises 1.58M unique images paired with 4.21M image-based conversations. This large-scale image corpus supports a wide range of perception and reasoning

tasks, including recognition, localization, segmentation, and captioning. For the video component, SurgΣ-DB contains 1.35M video clips paired with 1.45M conversations. These clips are typically short segments capturing fine-grained surgical activities and procedural context, supporting tasks such as action recognition and conditional surgical video generation.

The scale, diversity and multimodal richness of SurgΣ-DB make it a strong foundation for training multimodal foundation models for surgical intelligence. As shown in Figure 5d, SurgΣ-DB exhibits broad coverage across diverse tasks, while Figure 5e reveals a wide distribution of text token lengths, indicating rich linguistic diversity spanning both concise perception queries and complex multi-step reasoning.

3.2.2 Dataset License and Accessibility

SurgΣ-DB is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0). The license applies to all annotations to which we have directly contributed. SurgΣ-DB also incorporates surgical videos and images sourced from pre-existing collections. For these data, the original licensing terms are respected and remain applicable.

Our first public release, SurgΣ-DB v0.1, will be made publicly accessible through the official project page, where users can obtain the annotation files, metadata, and documentation necessary to reproduce the structure of the dataset. SurgΣ-DB is intended for non-commercial research purposes, and users are expected to properly cite both SurgΣ-DB and the original datasets in any resulting publications.

4 Advanced Foundation Models Built upon SurgΣ-DB

Building upon SurgΣ-DB, a spectrum of foundation models are developed, which cover complementary dimensions of surgical intelligence, from action-centric understanding and multimodal surgical scene understanding to structured reasoning and embodied policy learning. In the following, we briefly describe each model, highlighting its model and training designs, as well as empirical findings.

4.1 BSA: A Cross-Specialty Foundation Model for Basic Surgical Action Recognition

BSA [87] is a cross-specialty foundation model for recognizing basic surgical actions as a shared semantic unit across diverse procedures. Rather than modeling each procedure in isolation, BSA treats surgical workflow as compositions of reusable primitive actions (*e.g.*, dissection, coagulation, clipping, knot-tying), enabling a unified representation that transfers across anatomical sites, institutions, and recording conditions. Given short surgical video clips as input, it outputs probability distributions over ten predefined surgical action categories. Specifically, BSA builds upon a Video Transformer backbone [24] with two key design considerations for surgical video analysis: (1) sequential temporal and spatial attention mechanisms are utilized to effectively capture spatiotemporal dependencies inherent in surgical procedures, such as instrument movements and tissue interactions; (2) a dual-head prediction module is specifically designed to address the class imbalance issue. With video-action samples in SurgΣ-DB, the training pipeline utilized standard video preprocessing with temporal downsampling and uniform frame sampling to balance computational efficiency with visual information preservation. The Evidential loss [66] is employed to encourage well-calibrated uncertainty estimates, preventing both overconfident and excessively uncertain predictions. Please refer to BSA [87] for more implementation details.

Experimental results show that BSA learns stable and transferable representations of basic surgical actions across heterogeneous procedures, institutions, and imaging conditions (see Figure 6). Beyond recognition, BSA provides structured and uncertainty-aware action semantics that naturally support downstream applications, including surgical skill assessment and surgical action planning. These findings indicate that BSA functions not only as a standalone recognition model, but also as a foundational perception module that bridges low-level visual understanding and higher-level reasoning systems and embodied policy-learning pipelines. From a data-centric perspective, BSA operationalizes three principles for scalable surgical intelligence. First, ontology-first supervision: defining a compact, clinically meaningful action vocabulary improves semantic consistency across datasets and specialties. Second, cross-specialty alignment: harmonized labels and standardized preprocessing reduce dataset-specific shortcuts and encourage representations that generalize beyond a single procedure type. Third, uncertainty-aware recognition: modeling confidence is essential

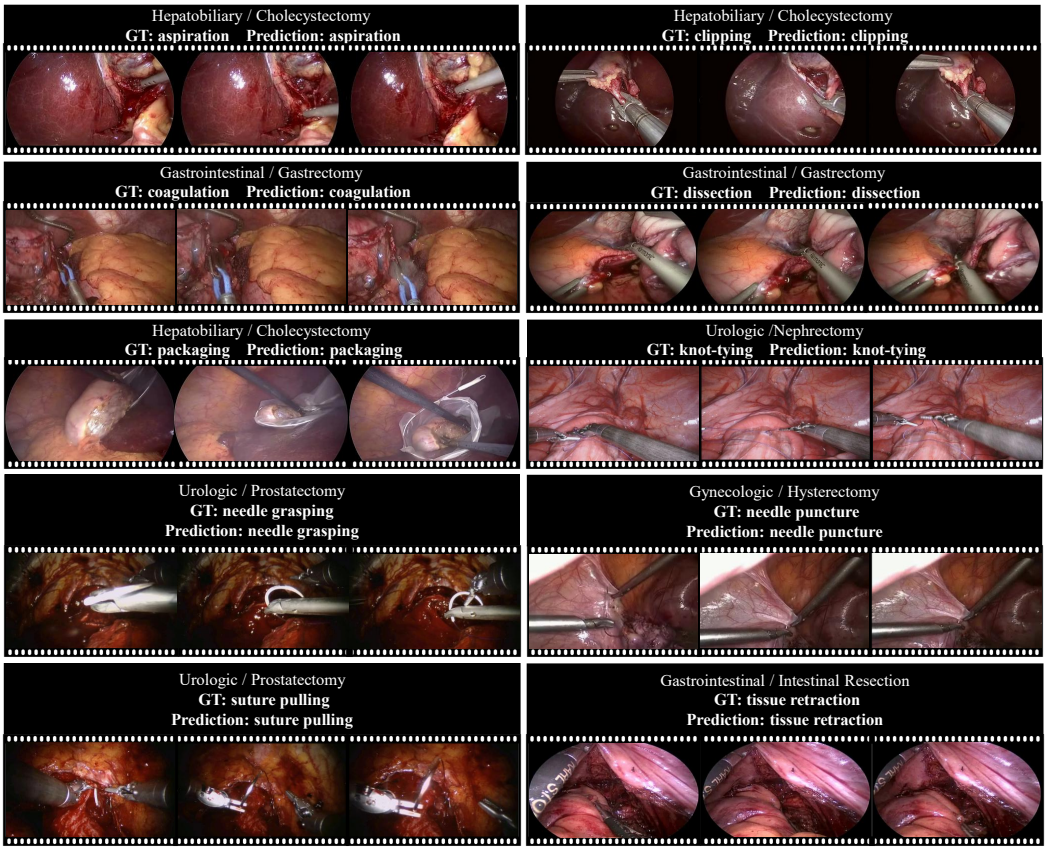


Figure 6: Qualitative visualization of BSA foundation model predictions across diverse surgical procedures.

for safe handoff to reasoning and control modules in high-stakes clinical settings. Together with SurgVLM (large-scale perception and reasoning alignment), Surg-R1 (hierarchical structured surgical reasoning), and Cosmos-H-Surgical (world-model-driven action-data synthesis for robot policy learning), BSA forms the front-end perceptual anchor of a unified pipeline toward embodied surgical intelligence: perception $\rightarrow$ reasoning $\rightarrow$ action. This systems view suggests that scalable surgical AI depends not on isolated model gains, but on tight co-design of action ontology, multimodal reasoning, and physically grounded world models.

4.2 SurgVLM: A Multimodal Foundation Model for Surgical Intelligence

Surgical intelligence presents unique challenges, requiring surgical visual perception, temporal analysis, and reasoning. Existing general-purpose vision-language models fail to address these needs due to insufficient domain-specific supervision and the lack of a large-scale high-quality surgical database. General-purpose vision-language models, trained predominantly on natural images and text, often exhibit inefficiency by generating excessive, clinically irrelevant outputs. Additionally, their outputs tend to be ambiguous, presenting multiple plausible scenarios rather than definitive, medically meaningful answers. Such ambiguity and verbosity undermine their alignment with surgeons' professional standards and real-world clinical requirements, significantly limiting their reliability and applicability in surgical practice. To address this challenge, SurgVLM is built upon SurgΣ-DB, adapted to ten surgical tasks through a unified sequence-to-sequence formulation optimized with a single autoregressive language modeling loss. To enhance generalization and mitigate biases, it develops an effective database construction pipeline, including data cleaning and refinement, cross-task correlation enrichment, explainable answer generation, and conversational diversity expansion. As a multimodal foundation model available in multiple scales (*i.e.*, 7B, 32B, and 72B), SurgVLM is designed to support a wide range of surgical understanding tasks, spanning both spatial and temporal analysis of surgical scenes, covering capabilities from visual perception to high-level reasoning. To

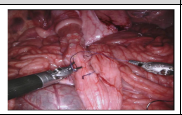
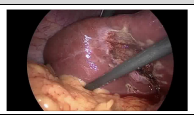
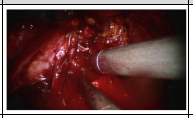
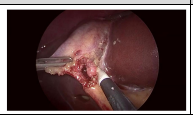
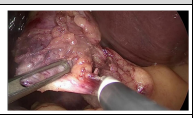
Task	Instrument Localization	Phase Recognition	Action Recognition	Triplet Recognition	Critical View of Safety Assessment
Image					
Question	Identify the location of the large needle driver in this image, using 3x3 grids to describe the location.	In the Cholecystectomy surgical image, what is the current phase? The available phase options are ...	What action related to the needle and suture is the surgeon focusing on right now? The available action options are ...	What tasks are the instrument accomplishing with the target in this surgical image? The available instrument, action, and target options are ...	For each Critical View of Safety criterion, answer yesor no. 1.Only two tubular structures connect to the gallbladder. 2.Hepatocystic triangle cleared for visibility. 3.Lower gallbladder detached from liver bed.
Answer	The large needle driver is at bottom-left area.	F. Cleaning Coagulation	D. pushing the needle through the tissue	Triplet1: C. grasper B. retract H. gallbladder Triplet2: E. hook C. dissect M.cystic_artery	1.no 2.no 3.no

Figure 7: Qualitative results of SurgVLM-72B including five typical examples from visual perception to temporal analysis to safety reasoning. For triplet recognition, the output format is triplet list with <Instrument,Verb,Target>.

build an effective training pipeline, SurgVLM models follow Qwen2.5-VL [9] architecture, consisting of a vision encoder, a projector, and a LLM decoder. The vision encoder is a transformer-based image backbone that processes images and video frames at their native resolution, while the LLM serves as the decoder for generating outputs. Please refer to SurgVLM [95] for more implementation details.

To systematically evaluate multimodal surgical intelligence, SurgVLM is evaluated on SurgVLM-Bench, a comprehensive benchmark designed to assess vision-language models across clinically relevant dimensions of surgical understanding. SurgVLM-Bench integrates six widely used surgical datasets spanning three hierarchical levels of task complexity: visual perception, temporal workflow analysis, and safety reasoning. These task categories reflect increasing contextual and temporal dependency and align with the requirements of real-world surgical assistance. The qualitative results with typical examples shown in Figure 7 generated by SurgVLM-72B, including instrument localization, phase recognition, action recognition, triplet recognition, and CVS assessment. The experiments of SurgVLM indicate that fine-tuning general VLMs on SurgΣ-DB provides an efficient and reliable pathway for surgical adaptation with following two important insights: (1) It indicates one of core problems in surgical foundation modeling lies in balancing diversity across surgical types with the exploitation of shared cross-procedure structure. Related procedures often exhibit substantial overlap in anatomical appearance, tissue characteristics, and instrument-associated visual cues, and joint training on multiple categories in SurgΣ-DB allows the model to leverage these synergies to learn richer and more transferable representations. At the same time, substantial variation remains across procedures in anatomy, workflow, and visual distribution; accordingly, SurgΣ-DB incorporates the surgical type as explicit contextual information during instruction tuning, reducing ambiguity and improving procedure-specific conditioning. (2) It supports a hierarchical view of surgical intelligence in which low-level perception, temporal understanding, and high-level reasoning are tightly coupled. Accurate recognition of instruments and tissues provides the basis for phase and action understanding, while robust temporal modeling underpins reliable intraoperative reasoning. This structure motivates multi-task co-training and a curriculum that progresses from perception to temporal analysis and finally to reasoning, improving data efficiency, convergence behavior, and robustness across the full surgical workflow.

4.3 Surg-R1: A Reasoning-Enhanced Model for Surgical Scene Understanding

Surg-R1 [33] is a reasoning-enhanced multimodal foundation model for surgical scene understanding with hierarchical chain-of-thought [82] reasoning capabilities. Built upon the reasoning-annotated data within SurgΣ-DB, Surg-R1 interprets complex surgical scenes through a structured three-level reasoning hierarchy: (1) perceptual grounding for instrument and tissue identification, (2) relational understanding for tool-tissue-action interactions, and (3) contextual reasoning for phase recognition and safety assessment. Initialized with Qwen2.5-VL-7B [9], Surg-R1 is trained through a comprehensive four-stage pipeline: First, supervised fine-tuning is performed to establish foundational vision–language alignment using question-answer pairs without reasoning. Secondly, structured

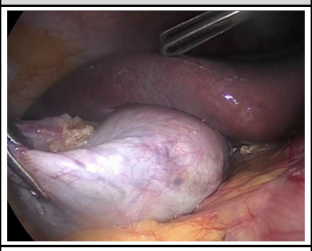
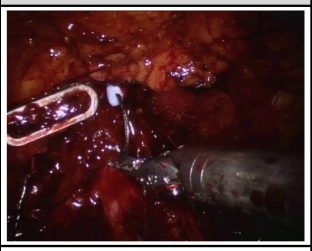
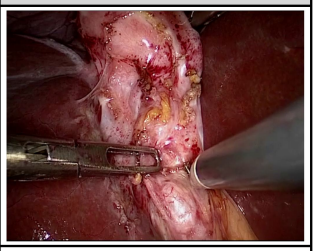
Task	Phase Recognition	Action Recognition	Triplet Recognition
Image			
Question	Given the surgical image, which **surgical phase** is being performed?	Given the radical prostatectomy image, what **surgical action** is being performed?	Given the laparoscopic cholecystectomy image, describe the complete surgical action in terms of **tool** , **action** , and **tissue** .
Thinking	*This is a Level 3 task...* **L1:** *gallbladder identified as the central anatomical structure... A grasper is visible gripping the infundibulum of the gallbladder, and the liver bed is partially exposed beneath.* **L2:** *grasper retracting gallbladder superiorly away from the liver bed, creating traction to expose the dissection plane...* **L3:** *...Gallbladder Dissection phase. The retraction direction and exposed dissection plane confirm the procedural context.*	*This is a Level 2 task...* **L1:** *two fenestrated graspers identified, suturing needle visible held between the jaws of the right grasper. The tissue surface appears consistent with a radical prostatectomy operative site.* **L2:** *right grasper engaged with needle, positioning it at a specific angle relative to the tissue surface... the left grasper provides counter-traction on adjacent tissue to stabilize the suturing target. The action is Positioning the needle tip.*	*This is a Level 2 task...* **L1:** *gallbladder identified, grasper and hook present. Calot's triangle region is partially exposed with connective tissue attachments visible.* **L2:** *grasper retracting gallbladder to expose Calot's triangle and the cystic duct junction. Hook dissecting connective tissue surrounding the cystic artery and duct. This dual-instrument coordination reveals two concurrent triplets: <grasper, retract, gallbladder> and <hook, dissect, gallbladder>.*
Answer	**Gallbladder Dissection**	**Positioning The Needle Tip**	**Grasper Retract Gallbladder, Hook Dissect Gallbladder**

Figure 8: Qualitative results of Surg-R1-7B across three representative surgical tasks. Each column shows the model's structured multi-level reasoning chain, progressing from visual identification (Level 1) through tool-tissue interaction analysis (Level 2) to procedural understanding (Level 3). **L**x denotes reasoning Level x. Reasoning traces are abbreviated with ellipses (. . .) for brevity.

reasoning priors are introduced through cold-start fine-tuning on reasoning trajectories synthesized under surgery-aware constraints that encode structured domain knowledge. Then, reasoning capability is further refined via reinforcement learning using Group Relative Policy Optimization [67]. Finally, an iterative refinement stage combines rejection sampling for correctly predicted samples and teacher-guided knowledge distillation for hard cases, progressively improving reasoning generalization capabilities beyond initial training data. Please refer to Surg-R1 [33] for more implementation details.

Surg-R1 is evaluated on thirteen datasets spanning six core surgical AI tasks, with seven public benchmarks and six multi-center external validation sets from five institutions, against proprietary reasoning models (GPT-5.1, Gemini 3.0 Pro), open-source generalist VLMs, and surgical-domain baselines. As shown in Figure 8, Surg-R1 produces structured multi-level reasoning chains that ground predictions in visual evidence. Surg-R1 achieves state-of-the-art performance across both settings, with the largest gains on compositional tasks. On CholecT50 triplet recognition, for example, Surg-R1 attains 51.69% accuracy versus 6.77% for GPT-5.1 and 8.01% for Qwen2.5-VL-7B-Surg. On multi-center external data it achieves an average arena score of 60.0%, compared with 44.9% for the leading surgical baseline. From a data-centric perspective, the consistent failure of general-purpose chain-of-thought reasoning on surgical compositional tasks, even in frontier models such as GPT-5.1, highlights the necessity of domain-specific structural priors. SurgΣ-DB's multi-granular annotation taxonomy provides the hierarchical scaffolding that makes effective surgical reasoning possible, and its structured instrument, tissue, and action vocabularies anchor the CoT synthesis pipeline in

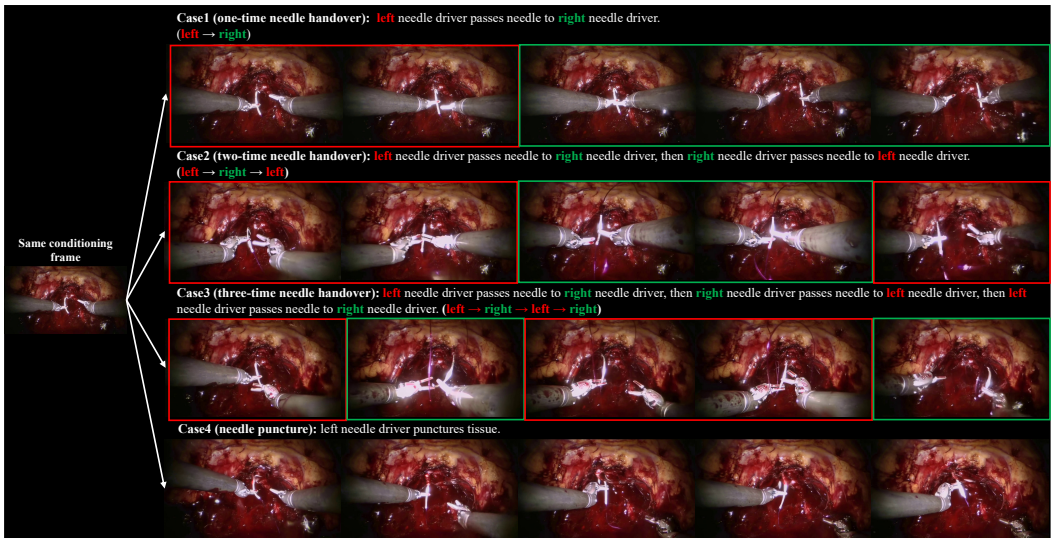


Figure 9: Cosmos-H-Surgical results: New behavior generalization via strong text–video alignment. Given the same conditioning frame, our surgical world model generates distinct video rollouts corresponding to four task prompts: (1) one-time needle handover, (2) two-time needle handover, (3) three-time needle handover, and (4) needle puncture.

visual observations, suppressing the hallucination artifacts that arise when models reverse-engineer explanations from labels. The three-level reasoning structure also brings practical advantages beyond training. Heterogeneous downstream systems can selectively consume only the granularity level relevant to their task. Skill assessment modules analyze Level 2 interaction patterns, workflow management systems leverage Level 3 phase predictions, and safety monitors extract CVS criteria, all without post-processing unstructured natural-language output. The structured hierarchy further enables automatic cross-level consistency checking and precise fault localization when predictions are wrong. Finally, every inference produces a hierarchically labeled reasoning trace that, after clinical review, can be ingested as new supervision, forming a data flywheel that continuously lowers annotation cost and grows the structured reasoning corpus in SurgΣ-DB.

4.4 Cosmos-H-Surgical: A Surgical World Model for Scalable Robot Policy Learning

Cosmos-H-Surgical [31] is a surgical world model and data-augmentation pipeline designed to bridge the gap between abundant unlabeled surgical video and the scarce paired video–kinematics data required for surgical robot vision–language–action (VLA) policy training. Based on SurgΣ-DB, a surgery-focused captioned dataset is curated with expert-authored action descriptions aligned to short surgical clips. The generative backbone of Cosmos-H-Surgical is based on a state-of-the-art video world model [4]. During training, Cosmos-H-Surgical learns to model surgical scene appearance and spatiotemporal dynamics under typical surgical imaging nuisances (specular highlights, occlusions, constrained tool motion), and condition generation on fine-grained text descriptions so that synthesized videos preserve actionable affordances and tool–tissue interactions required for downstream policy learning. An inverse-dynamics model (IDM) is trained on limited real paired demonstrations (when available) and then applied to synthetic video to recover pseudo kinematics (approximate action/robot-state sequences), thus producing large-scale synthetic (video, pseudo-kinematics, text) triples suitable for supervised VLA or imitation-style policy learning. Cosmos-H-Surgical therefore turns unlabeled surgical video corpora into paired training data at scale, enabling standard VLA optimization without requiring dense manual robot-state annotation. Please refer to Cosmos-H-Surgical [31] for more implementation details.

Experimental results show that Cosmos-H-Surgical–augmented policies significantly outperform those trained solely on limited real demonstrations, achieving higher task success rates and improved sample efficiency. Although IDM-generated pseudo-kinematics are imperfect, they provide sufficiently informative supervision when paired with realistic world-model synthesis. The combination of generative diversity and structured inverse inference enables effective scaling of embodied training data. From a data-centric perspective, Cosmos-H-Surgical highlights three key insights.

First, domain-specific data curation is essential. Generic video generation is insufficient for surgical scenarios; the SATA dataset, curated from BSA with structured action annotations, provides the semantic grounding required to align generation with physically meaningful surgical actions. Second, forward physical consistency matters. World models must capture constrained tool kinematics and tissue deformation; otherwise, synthetic demonstrations can introduce policy bias. Third, hybrid supervision is most effective. While synthetic augmentation reduces reliance on real demonstrations, the best performance arises from mixed training (synthetic + limited real data), where real data anchors policies within true robot dynamics. Together with SurgVLM (large-scale perception and reasoning alignment) and Surg-R1 (hierarchical structured reasoning with reinforcement refinement), Cosmos-H-Surgical completes the pipeline toward embodied surgical intelligence: perception $\rightarrow$ reasoning $\rightarrow$ action. These results suggest that scalable surgical AI requires not only multimodal foundation models, but also structured world models that translate visual understanding into physical control.

5 Limitations and Future Work

We aim to progressively achieve comprehensive multi-task annotation coverage across all surgical scenes in SurgΣ-DB. Although its current v0.1 release spans diverse understanding, reasoning, planning, and generation tasks, full-spectrum supervision has not yet been uniformly established for every sample. While certain subsets already provide comprehensive multi-task and reasoning-level annotations, other samples remain limited to task-specific supervision (*e.g.*, perception-level labels), and structured reasoning annotations are not consistently available across all conversations.

This imbalance largely stems from the intrinsic complexity and high cost of surgical data collection and annotation. Surgical scenes are dynamic, anatomically intricate and safety-critical, requiring domain expertise to precisely characterize fine-grained details and procedural context. In particular, reasoning annotations demand multi-level clinical interpretation and careful validation, making large-scale consistent annotation substantially more challenging than standard perception labeling.

In future iterations beyond SurgΣ-DB v0.1, we will continue expanding cross-task coverage and enriching structured reasoning annotation under a unified label space and annotation framework, moving toward more holistic and fully aligned multimodal surgical foundation training.

6 Conclusion

In this work, we introduced SurgΣ, a unified spectrum of large-scale multimodal data and foundation models for surgical intelligence. At its core, SurgΣ-DB provides a systematically curated and scalable multimodal data foundation, consolidating heterogeneous surgical resources into a unified schema with consistent semantics and standardized formats across diverse procedures. SurgΣ-DB supports comprehensive supervision spanning understanding, reasoning, planning, and generation tasks, and is designed around three key principles: large-scale multimodal data, unified data representations, and structured reasoning annotations. Empirical evidence from four foundation models built upon SurgΣ-DB demonstrates that these data-centric design choices substantially enhance cross-task generalization and interpretable reasoning in complex surgical environments. We envision SurgΣ-DB as a scalable infrastructure for surgical foundation modeling, and will continue expanding its scale, diversity, and annotation completeness toward dense cross-task coverage within unified surgical scenes, advancing clinically reliable and generalizable multimodal surgical intelligence.

Acknowledgments

We sincerely thank Dillan Imans for his invaluable assistance with dataset annotation during his visit as a researcher at CUHK. We also thank Erli Zhang from NUS for his invaluable assistance with dataset summarizing.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[2] Oluwatosin Alabi, Ko Ko Zayar Toe, Zijian Zhou, Charlie Budd, Nicholas Raison, Miaojing Shi, and Tom Vercauteren. Cholecinstanceseg: A tool instance segmentation dataset for laparoscopic surgery. *Scientific Data*, 12(1):825, 2025.

[3] Deepak Alapatt, Jennifer Eckhoff, Zhiliang Lyu, Yutong Ban, Jean-Paul Mazellier, Sarah Choksi, Kunyi Yang, Po-Hsing Chiang, Noemi Zorzetti, Samuele Cannas, et al. The SAGES critical view of safety challenge: A global benchmark for AI-assisted surgical quality assessment. *arXiv preprint arXiv:2509.17100*, 2025.

[4] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

[5] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

[6] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

[7] Nicolás Ayobi, Santiago Rodríguez, Alejandra Pérez, Isabela Hernández, Nicolás Aparicio, Eugénie Dessevres, Sebastián Peña, Jessica Santander, Juan Ignacio Caicedo, Nicolás Fernández, et al. Pixel-wise recognition for holistic surgical scene understanding. *Medical Image Analysis*, page 103726, 2025.

[8] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[9] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[10] Yutong Ban, Jennifer A. Eckhoff, Thomas M. Ward, Daniel A. Hashimoto, Ozanan R. Meireles, Daniela Rus, and Guy Rosman. Concept graph neural networks for surgical video understanding. *IEEE Transactions on Medical Imaging*, 43(1):264–274, 2024.

[11] Vivek Singh Bawa, Gurkirt Singh, Francis KapingA, Inna Skarga-Bandurova, Elettra Oleari, Alice Leporini, Carmela Landolfo, Pengfei Zhao, Xi Xiang, Gongning Luo, et al. The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods. *arXiv preprint arXiv:2104.03178*, 2021.

[12] Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, Rogier Van Der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay S Chaudhari. A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering*, 9(4):494–506, 2025.

[13] Matthias Carstens, Franziska M Rinner, Sebastian Bodenstedt, Alexander C Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R Kolbinger. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):3, 2023.

[14] João Cartucho, Alistair Weld, Samyakh Tukra, Haozheng Xu, Hiroki Matsuzaki, Taiyo Ishikawa, Minjun Kwon, Yong Eun Jang, Kwang-Ju Kim, Gwang Lee, et al. Surgt challenge: Benchmark of soft-tissue trackers for robotic surgery. *Medical image analysis*, 91:102985, 2024.

[15] Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C Garcia-Peraza-Herrera. Lemon: A large endoscopic monocular dataset and foundation model for perception in surgical settings. *arXiv preprint arXiv:2503.19740*, 2025.

[16] Kexin Chen, Yuyang Du, Tao You, Mobarakol Islam, Ziyu Guo, Yueming Jin, Guangyong Chen, and Pheng-Ann Heng. Llm-assisted multi-teacher continual learning for visual question answering in robotic surgery. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10772–10778, 2024.

[17] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

[18] Yiliang Chen, Zhixi Li, Cheng Xu, Alex Qinyang Liu, Ruize Cui, Xuemiao Xu, Jeremy Yuen-Chun Teoh, Shengfeng He, and Jing Qin. Prostatd: Bridging surgical triplet from classification to fully supervised detection. *arXiv preprint arXiv:2506.01130*, 2025.

[19] Zhen Chen, Qing Xu, Jinlin Wu, Biao Yang, Yuhao Zhai, Geng Guo, Jing Zhang, Yinlu Ding, Nassir Navab, and Jiebo Luo. How far are surgeons from surgical world models? a pilot study on zero-shot surgical video generation with expert assessment. *arXiv preprint arXiv:2511.01775*, 2025.

[20] Jiajun Cheng, Xianwu Zhao, Sainan Liu, Xiaofan Yu, Ravi Prakash, Patrick J. Codd, Jonathan Elliott Katz, and Shan Lin. Surgxbench: Explainable vision-language model benchmark for surgery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8188–8198, March 2026.

[21] Tae-Min Choi, Tae Kyeong Jeong, Garam Kim, Jaemin Lee, Yeongyoon Koh, In Cheul Choi, Jae-Ho Chung, Jong Woong Park, and Juyoun Park. Surgmllmbench: A multimodal large language model benchmark dataset for surgical scene understanding. *arXiv preprint arXiv:2511.21339*, 2025.

[22] Ronald de Jong, HJ Carolus, HA Franciscus, Romy C van Jaarsveld, Richard van Hillegersberg, PW Josien, Peter HN de With, Yasmina al Khalil, Fons van Der Sommen, et al. Scaling up self-supervised learning for improved surgical foundation models. *Medical Image Analysis*, page 103873, 2025.

[23] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, pages 1–13, 2025.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[25] Pierre E Dupont, Bradley J Nelson, Michael Goldfarb, Blake Hannaford, Arianna Menciassi, Marcia K O'Malley, Nabil Simaan, Pietro Valdastri, and Guang-Zhong Yang. A decade retrospective of medical robotics research from 2010 to 2020. *Science robotics*, 6(60):eabi8017, 2021.

[26] Davide Ferrari, Tommaso Violante, Marco Novelli, Patrick P Starlinger, Rory L Smoot, Janani S Reisenauer, and David W Larson. The death of laparoscopy. *Surgical endoscopy*, 38(5):2677–2688, 2024.

[27] Negin Ghamsarian, Yosuf El-Shabrawi, Sahar Nasirihaghighi, Doris Putzgruber-Adamitsch, Martin Zinkernagel, Sebastian Wolf, Klaus Schoeffmann, and Raphael Sznitman. Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos. *Scientific data*, 11(1):373, 2024.

[28] Fengyue Guo, Chengkun Li, Bin Peng, Yonghao Long, Jialun Pei, Mengya Xu, Ziling He, Guangsuo Wang, and Qi Dou. Surgical key step recognition with global-local modeling mamba in laparoscopic pulmonary lobectomy. In *International Workshop on Collaborative Intelligence and Autonomy in Image-Guided Surgery*, pages 11–20. Springer, 2025.

[29] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[30] Runlong He, Mengya Xu, Adrito Das, Danyal Z Khan, Sophia Bano, Hani J Marcus, Danail Stoyanov, Matthew J Clarkson, and Mobarakol Islam. Pitvqa: Image-grounded text embedding llm for visual question answering in pituitary surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024.

[31] Yufan He, Pengfei Guo, Mengya Xu, Zhaoshuo Li, Andriy Myronenko, Dillan Imans, Bingjie Liu, Dongren Yang, Mingxue Gu, Yongnan Ji, et al. Cosmos-h-surgica: Learning surgical robot policies from videos via world modeling. *arXiv preprint arXiv:2512.23162*, 2025.

[32] Aimin Jiang, Zhao Tang, Hanzhong Zhang, Jinxin Li, Jialin Meng, Ying Liu, Yu Fang, Juan Lu, Xu Zhang, Le Qu, et al. Current application status and innovative development of surgical robot. *Med Research*, 1(3):378–396, 2025.

[33] Jian Jiang, Chenxi Lin, Yiming Gu, Zengyi Qin, Zhitao Zeng, Kun Yuan, Yonghao Long, Xiang Xia, Cheng Yuan, Yuqi Wang, Zijie Yue, Kunyi Yang, Yuting Zhang, Zhu Zhuo, Dian Qin, Xin Wang, NG Chi Fai, Brian Anthony, Daguang Xu, Guy Rosman, Ozanan Meireles, Zizhen Zhang, Nicolas Padoy, Hesheng Wang, Qi Dou, Yueming Jin, and Yutong Ban. Surg-r1: A hierarchical reasoning foundation model for scalable and interpretable surgical decision support with multi-center clinical validation. *arXiv preprint arXiv:2603.12430*, 2026.

[34] Juseong Jin and Chang Wook Jeong. Surgical-llava: Toward surgical scenario understanding via large language and vision models. *arXiv preprint arXiv:2410.09750*, 2024.

[35] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging*, 37(5):1114–1126, 2017.

[36] Yueming Jin, Yonghao Long, Xiaojie Gao, Danail Stoyanov, Qi Dou, and Pheng-Ann Heng. Trans-svnet: hybrid embedding aggregation transformer for surgical workflow analysis. *International Journal of Computer Assisted Radiology and Surgery*, 17(12):2193–2202, 2022.

[37] Xinwei Ju, Rema Daher, Razvan Caramalau, Baoru Huang, Danail Stoyanov, and Francisco Vasconcelos. Segcol challenge: Semantic segmentation for tools and fold edges in colonoscopy data. *arXiv preprint arXiv:2412.16078*, 2024.

[38] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[39] Sreeram Kamabattula, Kai Chen, and Kiran Bhattacharyya. Weakly supervised pre-training for surgical step recognition using unannotated and heterogeneously labeled videos. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–11, 2025.

[40] Joël L Lavanchy, Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Paolo Fiorini, Beat P Müller-Stich, Philipp C Nett, Jacques Marescaux, Didier Mutter, and Nicolas Padoy. Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. *International journal of computer assisted radiology and surgery*, 19(11):2249–2257, 2024.

[41] Jiajie Li, Garrett Skinner, Gene Yang, Brian R Quaranto, Steven D Schwaitzberg, Peter CW Kim, and Jinjun Xiong. Llava-surg: towards multimodal surgical assistant via structured surgical video learning. *arXiv preprint arXiv:2408.07981*, 2024.

[42] Yaoqian Li, Xikai Yang, Dunyuan Xu, Yang Yu, Litao Zhao, Xiaowei Hu, Jinpeng Li, and Pheng-Ann Heng. Surgpub-video: A comprehensive surgical video dataset for enhanced surgical intelligence in vision-language model. *arXiv preprint arXiv:2508.10054*, 2025.

[43] Haofeng Liu, Ziyue Wang, Sudhanshu Mishra, Mingqi Gao, Guanyi Qin, Chang Han Low, Alex YW Kong, and Yueming Jin. Sam2s: Segment anything in surgical videos via semantic long-term tracking. *arXiv preprint arXiv:2511.16618*, 2025.

[44] Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[45] Chang Han Low, Ziyue Wang, Tianyi Zhang, Zhu Zhuo, Zhitao Zeng, Evangelos B Mazomenos, and Yueming Jin. Surgraw: Multi-agent workflow with chain of thought reasoning for robotic surgical video analysis. *IEEE Robotics and Automation Letters*, 2026.

[46] Imanol Luengo, Maria Grammatikopoulou, Rahim Mohammadi, Chris Walsh, Chinedu Innocent Nwoye, Deepak Alapatt, Nicolas Padoy, Zhen-Liang Ni, Chen-Chen Fan, Gui-Bin Bian, et al. 2020 cataracts semantic segmentation challenge. *arXiv preprint arXiv:2110.10965*, 2021.

[47] Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, et al. Heidelberg colorectal data set for surgical data science in the sensor operating room. *Scientific data*, 8(1):101, 2021.

[48] Pietro Mascagni, Deepak Alapatt, Aditya Murali, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, et al. Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. *Scientific Data*, 12(1):331, 2025.

[49] John G Meara, Andrew JM Leather, Lars Hagander, Blake C Alkire, Nivaldo Alonso, Emmanuel A Ameh, Stephen W Bickler, Lesong Conteh, Anna J Dare, Justine Davies, et al. Global surgery 2030: evidence and solutions for achieving health, welfare, and economic development. *The lancet*, 386(9993):569–624, 2015.

[50] Baudolino Mussa, Barbara Defrancisco, Ludovico Campi, and Mario Morino. Single-port laparoscopy compared with conventional laparoscopic surgery: a systematic review and meta-analysis. *Journal of Clinical Medicine*, 14(14):4915, 2025.

[51] Hirenkumar Nakawala. Nephrec9. *(No Title)*, 2017.

[52] Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *International conference on medical image computing and computer-assisted intervention*, pages 364–374. Springer, 2020.

[53] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

[54] Jialun Pei, Jiaan Zhang, Guanyi Qin, Kai Wang, Yueming Jin, and Pheng-Ann Heng. Instrument-tissue-guided surgical action triplet detection via textual-temporal trail exploration. *IEEE transactions on medical imaging*, 2025.

[55] Alejandra Perez, Chinedu Nwoye, Ramtin Raji Kermani, Omid Mohareri, and Muhammad Abdullah Jamal. Surglavi: Large-scale hierarchical dataset for surgical vision-language representation learning. *arXiv preprint arXiv:2509.10555*, 2025.

[56] Alejandra Perez, Anita Rau, Lee White, Busisiwe Mlambo, Chinedu Nwoye, Muhammad Abdullah Jamal, and Omid Mohareri. Sureon: A benchmark and vision-language-model for surgical reasoning. *arXiv preprint arXiv:2603.0657*, 2026.

[57] Dimitrios Psychogyios, Emanuele Colleoni, Beatrice Van Amsterdam, Chih-Yang Li, Shu-Yu Huang, Yuchong Li, Fucang Jia, Baosheng Zou, Guotai Wang, Yang Liu, et al. Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*, 2023.

[58] Guanyi Qin, Xiaozhen Wang, Zhu Zhuo, Chang Han Low, Yuancan Xiao, Yibing Fu, Haofeng Liu, Kai Wang, Chunjiang Li, and Yueming Jin. Surgo-r1: Benchmarking and modeling contextual reasoning for operative zone in surgical video. *arXiv preprint arXiv:2602.21706*, 2026.

[59] Anita Rau, Mark Endo, Josiah Aklilu, Jaewoo Heo, Khaled Saab, Alberto Paderno, Jeffrey Jopling, F Christopher Holsinger, and Serena Yeung-Levy. Systematic evaluation of large vision-language models for surgical artificial intelligence. *arXiv preprint arXiv:2504.02799*, 2025.

[60] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2024.

[61] Manuel Sebastián Ríos, María Alejandra Molina-Rodriguez, Daniella Londoño, Camilo Andrés Guillén, Sebastián Sierra, Felipe Zapata, and Luis Felipe Giraldo. Cholec80-cvs: An open dataset with an evaluation of strasberg's critical view of safety for ai. *Scientific Data*, 10(1):194, 2023.

[62] Samuel Schmidgall, Ji Woong Kim, Jeffrey Jopling, and Axel Krieger. General surgery vision transformer: A video pre-trained foundation model for general surgery. *arXiv preprint arXiv:2403.05949*, 2024.

[63] Klaus Schoeffmann, Heinrich Husslein, Sabrina Kletz, Stefan Petscharnig, Bernd Muenzer, and Christian Beecks. Video retrieval in laparoscopic video recordings with dynamic content descriptors. *Multimedia Tools and Applications*, 77(13):16813–16832, 2018.

[64] Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. Surgicalgpt: end-to-end language-vision gpt for visual question answering in surgery. In *International conference on medical image computing and computer-assisted intervention*, pages 281–290, 2023.

[65] Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 33–43, 2022.

[66] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[67] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[68] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[69] Yue Sun, Limei Wang, Gang Li, Weili Lin, and Li Wang. A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. *Nature Biomedical Engineering*, 9(4):521–538, 2025.

[70] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[71] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[72] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017.

[73] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 30(10):2924–2935, 2024.

[74] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *Medical image analysis*, 86:102770, 2023.

[75] Guankun Wang, Long Bai, Wan Jun Nah, Jie Wang, Zhaoxi Zhang, Zhen Chen, Jinlin Wu, Mobarakol Islam, Hongbin Liu, and Hongliang Ren. Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. *arXiv preprint arXiv:2405.10948*, 2024.

[76] Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, et al. Endochat: Grounded multimodal large language model for endoscopic surgery. *arXiv preprint arXiv:2501.11347*, 2025.

[77] Guankun Wang, Junyi Wang, Wenjin Mo, Long Bai, Kun Yuan, Ming Hu, Jinlin Wu, Junjun He, Yiming Huang, Nicolas Padoy, et al. Surgvidlm: Towards multi-grained surgical video understanding with large language model. *arXiv preprint arXiv:2506.17873*, 2025.

[78] Guankun Wang, Han Xiao, Renrui Zhang, Huxin Gao, Long Bai, Xiaoxiao Yang, Zhen Li, Hongsheng Li, and Hongliang Ren. 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*, pages 12636–12643, 2025.

[79] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International conference on medical image computing and computer-assisted intervention*, pages 101–111. Springer, 2023.

[80] Zipei Wang, Sitian Pan, Mengjie Fang, Ruofan Zhang, Jie Tian, and Di Dong. CholecMamba: A Mamba-based Multimodal Reasoning Model for Cholecystectomy Surgery . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15968, pages 107–116. Springer Nature Switzerland, September 2025.

[81] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 486–496, 2022.

[82] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[83] Jianhui Wei, Zikai Xiao, Danyu Sun, Luqi Gong, Zongxin Yang, Zuozhu Liu, and Jian Wu. Surgbench: A unified large-scale benchmark for surgical video analysis. *arXiv preprint arXiv:2506.07603*, 2025.

[84] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

[85] Jinlin Wu, Felix Holm, Chuxi Chen, An Wang, Yaxin Hu, Xiaofan Ye, Zelin Zang, Miao Xu, Lihua Zhou, Huai Liao, et al. Unisurg: A video-native foundation model for universal understanding of surgical videos. *arXiv preprint arXiv:2602.05638*, 2026.

[86] Yingcheng Charles Wu, Ming Yin, Baiyu Shi, Zaixi Zhang, Di Yin, Xiaotong Wang, Youjuan Wang, Jigang Fan, Ruofan Jin, Hanchen Wang, et al. Medos: Ai-xr-cobot world model for clinical perception and action. *medRxiv*, pages 2026–02, 2026.

[87] Mengya Xu, Daiyun Shen, Jie Zhang, Hon Chi Yip, Yujia Gao, Cheng Chen, Dillan Imans, Yonghao Long, Yiru Ye, Yixiao Liu, Rongyun Mai, Kai Chen, Hongliang Ren, Yutong Ban, Guangsuo Wang, Francis Wong, Chi-Fai Ng, Kee Yuan Ngiam, Russell H. Taylor, Daguang Xu, Yueming Jin, and Qi Dou. Generalized recognition of basic surgical actions enables skill assessment and vision-language-model-based surgical planning. *arXiv preprint arXiv:2603.12787*, 2026.

[88] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[89] Shu Yang, Fengtao Zhou, Leon Mayer, Fuxiang Huang, Yiliang Chen, Yihui Wang, Sunan He, Yuxiang Nie, Xi Wang, Yueming Jin, et al. Large-scale self-supervised video foundation model for intelligent surgery. *npj Digital Medicine*, 2026.

[90] Cheng Yuan, Jian Jiang, Kunyi Yang, Lv Wu, Rui Wang, Zi Meng, Haonan Ping, Ziyu Xu, Yifan Zhou, Wanli Song, Hesheng Wang, Yueming Jin, Qi Dou, and Yutong Ban. Systematic evaluation and guidelines for segment anything model in surgical video analysis. *arXiv preprint arXiv:2501.00525*, 2025.

[91] Kun Yuan, Manasi Kattel, Joël L Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. Advancing surgical vqa with scene graph knowledge. *International journal of computer assisted radiology and surgery*, 19(7):1409–1417, 2024.

[92] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. HecVL: Hierarchical Video-Language Pretraining for Zero-shot Surgical Phase Recognition . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15006, pages 306–316. Springer Nature Switzerland, October 2024.

[93] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *Advances in Neural Information Processing Systems*, 37:122952–122983, 2024.

[94] Kun Yuan, Vinkle Srivastav, Tong Yu, Joel L Lavanchy, Jacques Marescaux, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. *Medical Image Analysis*, 105:103644, 2025.

[95] Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, Erli Zhang, Junde Wu, Jiaan Zhang, Yuxuan Wang, Chang Han Low, Jian Jiang, Zilong Zheng, et al. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *arXiv preprint arXiv:2506.02555*, 2025.