LOGO

EndoAgent: A Hierarchical Multi-Modal Agent System for Capsule Endoscopy Video Analysis

Jianqiao Zhu, Dejun Fan, Jingyu Zhuang, Xinyan Li, Xinyang He, Haifan Gong, *Member, IEEE*, Zheng Liu, Du Cai, Feng Liang, *Member, IEEE*, Guanbin Li, *Member, IEEE*, Jiancong Hu

***Abstract*— Capsule endoscopy (CE) revolutionizes gastrointestinal examination but presents a significant efficiency bottleneck due to the massive volume of video data and the extreme sparsity of pathological features. Existing Vision-Language Models (VLMs) are limited by their context window size, rendering them unable to process ultra-long sequences, and often suffer from hallucinations due to a lack of medical reasoning capabilities. To address this, we propose EndoAgent, a novel hierarchical multi-modal agent system designed for automated CE video analysis and diagnostic report generation. Inspired by the dual-process theory of human cognition, EndoAgent deploys a Fast-Scan Filter that utilizes hierarchical redundancy removal and event-driven segmentation to purge the majority of irrelevant frames while anchoring critical pathological context. This is followed by a Slow-Scan Agent that constructs a structured multi-modal memory bank and dynamically orchestrates domain-specific tools—such as the zoom-in Visual Verifier and Template Filler—to perform fine-grained reasoning through an "Observation-Thought-Action" loop. This agentic workflow bridges the gap between raw pixel streams and high-level clinical reasoning. Furthermore, we introduce a rigorous "LLM-as-a-Judge" evaluation protocol to assess the diagnostic validity of generated reports beyond surface-level lexical metrics. Extensive experiments demonstrate that EndoAgent significantly outperforms existing generalist and medical VLMs in clinical consistency, completeness, and hallucination mitigation.**

***Index Terms*— Capsule endoscopy, medical agent, long video understanding, vision-language models, medical report generation.**

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by The Sixth Affiliated Hospital, Sun Yat-sen University under Grant 2024ZSLYEC-040. This work was supported by National Key Clinical Discipline. (Corresponding authors: Dejun Fan, Feng Liang, Guanbin Li and Jiancong Hu.)

Jianqiao Zhu, Jingyu Zhuang, Haifan Gong, Zheng Liu and Guanbin Li are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China (e-mail: zhujq33@mail2.sysu.edu.cn; zhuangjy6@mail2.sysu.edu.cn; gonghf@mail2.sysu.edu.cn; liuzh385@mail2.sysu.edu.cn; liguanbin@mail.sysu.edu.cn).

Dejun Fan and Jiancong Hu are with the Department of Gastrointestinal Endoscopy and the Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, China, and also with the Guangzhou (Huangpu) Zhongliu Biomedical Innovation Center, Guangzhou, China (e-mail: fandj3@mail.sysu.edu.cn; hujianc@mail.sysu.edu.cn).

Xinyan Li is with the Guangdong Hong Kong-Macau University Joint Laboratory, Shenzhen MSU-BIT University (e-mail: l202500232@smbu.edu.cn).

Xinyang He is with the School of Medical Technology, Beijing Institute of Technology and also with the Guangdong Hong Kong-Macau University Joint Laboratory, Shenzhen MSU-BIT University (e-mail: hexinyang@bit.edu.cn).

Du Cai is with the Department of General Surgery (Colorectal Surgery), the Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, and the Biomedical Innovation Center, The Sixth Affiliated Hospital, Sun Yat-sen University (e-mail: caid28@mail.sysu.edu.cn).

Feng Liang is with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University (e-mail: fliang@smbu.edu.cn).

I. Introduction

Capsule endoscopy (CE) has revolutionized gastrointestinal examination by providing a non-invasive method for complete visualization of the tract, particularly the small bowel, which is difficult to access with conventional endoscopy [1]. It has become a first-line diagnostic tool for conditions such as obscure gastrointestinal bleeding, Crohn's disease, and small bowel tumors [2]. However, its immense clinical value is constrained by a significant efficiency bottleneck: the burden of reviewing massive video data. A single procedure generates an approximately 12-hour video comprising about 120,000 frames [3]. Currently, this requires gastroenterologists to perform a full manual review, a process that is extremely time-consuming (typically 1-2 hours) and leads to significant visual fatigue and sustained attention decline [3]. Consequently, this process carries a substantial risk of observer error and diagnostic oversight. The fundamental challenge stems from profound data imbalance and high temporal redundancy: although the overwhelming majority of the video consists of normal or repetitive mucosa, clinically actionable findings—such as minor ulcers, subtle angiectasias, or early inflammatory lesions—are exceedingly sparse, often constituting less than 5% of the total frames [2]. This creates a classic "needle-in-a-haystack" scenario where critical anomalies are buried within a vast stream of redundant frames.

Existing computational approaches have failed to comprehensively address this multifaceted challenge. Traditional Convolutional Neural Networks (CNNs [4]–[7]) or Transformer-based methods [1], [8], [9] are generally limited to single-frame or short-clip classification and detection. They lack the global context required to understand the complete procedure and struggle to generate coherent, logical diagnostic reports. More recently, large-scale Vision-Language Models (VLMs) [10]–[12] have shown promise in general video understanding. However, as shown in Fig. 1, they are ill-suited for this specific domain due to severely limited context window size—they cannot directly process sequences of hundreds of thousands of frames [13]. Furthermore, general-purpose VLMs

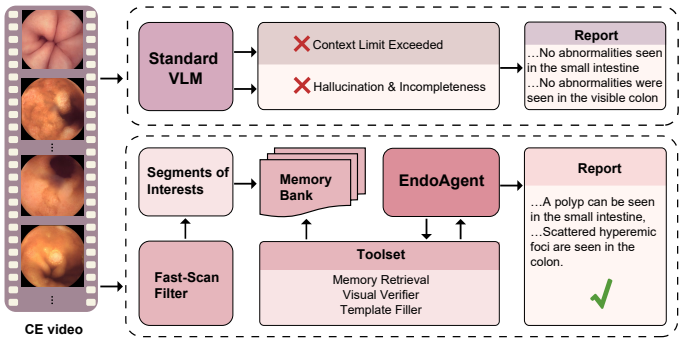


Fig. 1. **Paradigm Shift in Long-Video Capsule Endoscopy Analysis. Top**: Traditional passive VLMs struggle with massive CE video inputs due to context window size bottleneck, often leading to computational failures or hallucinations. **Bottom**: Our EndoAgent introduces an active cognitive workflow. A Fast-Scan filter first discards most of the redundancy. Subsequently, a Slow-Scan Agent performs deep reasoning by querying a structured Memory Bank and actively using domain-specific tools (e.g., Visual Verifier) to produce precise, verified diagnostic reports.

often lack the specialized medical knowledge necessary to distinguish subtle pathological features from normal mucosa [14], [15].

This disconnect highlights the need for a paradigm shift in long-form video understanding: from passive, context-limited perception to active, autonomous exploration. This is where the concept of an "Agent" becomes indispensable. Unlike passive models, an autonomous agent is characterized by its ability to perceive, plan, and actively utilize tools to solve complex, multi-step problems [16]–[18]. Crucially, the workflow of a gastroenterologist—continuously scanning for anomalies (Fast) and zooming in to verify details (Slow)—essentially mirrors the Dual-Process Theory of human cognition [19]. Inspired by this cognitive mechanism, we innovatively construct a hierarchical Fast-Slow Agent system designed to solve the conflict between the massive data volume and the need for fine-grained reasoning.

As depicted in Fig. 1, our EndoAgent utilizes a Fast-Scan Filter to efficiently filter redundancy and locate segments of interest, followed by a Slow-Scan Agent to conduct fine-grained analysis through tool invocation [19]. The Fast-Scan Filter preprocesses the 12-hour video using visual and semantic deduplication to create a keyframe list. A lightweight anomaly detection model then scores these keyframes, clustering contiguous high-anomaly frames into Segments of Interest. This recall-first event-driven segmentation efficiently reduces the data from approximately 120,000 frames to about 300 Segments of Interest.

The Slow-Scan Agent first constructs a comprehensive Multi-modal Memory Bank from Segments of Interest using a LoRA-finetuned [20] Qwen2.5-VL [21], effectively translating raw visual data into structured semantic indices. Instead of a linear pass, the Agent engages in an iterative "Observation-Thought-Action" loop: it continuously observes the retrieved context, reasons about diagnostic ambiguities, and dynamically acts by invoking domain-specific tools (e.g., Anomaly Detector and Visual Verifier) to rigorously verify pathological details and autonomously generate a structured diagnostic report.

The main contributions of our work are summarized as follows:

- We introduce EndoAgent, a novel hierarchical multi-modal agent system, representing the first tool-augmented, end-to-end agentic system for full-length CE video analysis and report generation.
- We design an event-driven video summarization strategy that combines efficient redundancy removal with anomaly-guided segmentation, enabling effective focus on clinically relevant Segments of Interest.
- We establish an active diagnostic reasoning agent governed by an "Observation-Thought-Action" loop, which dynamically orchestrates domain-specific tools—including an intelligent zoom-in Visual Verifier—to resolve visual ambiguities through iterative evidence verification.
- We propose a novel, entity-aware "LLM-as-a-Judge" evaluation protocol. Unlike traditional N-gram metrics, this framework assesses clinical correctness, completeness, and hallucination risks, setting a new standard for evaluating generative medical reports.
- We conduct extensive validation on a large-scale dataset of real-world full-length CE videos. Experimental results demonstrate that EndoAgent achieves superior performance over generalist and medical VLMs, particularly in clinical consistency and hallucination mitigation, highlighting its potential for reducing diagnostic workload.

II. Related Work

Our work is situated at the intersection of automated medical video analysis, vision-language models (VLMs), and autonomous agent systems. In this section, we review the evolution of these fields and highlight the specific gaps in Capsule Endoscopy (CE) diagnostics that our EndoAgent System addresses.

A. Automated Medical Video Analysis

The analysis of CE video has evolved from hand-crafted features to sophisticated deep learning paradigms. Early approaches primarily utilized CNNs [22] to detect specific pathologies, such as polyps, bleeding, or ulcers, treating video frames as independent images [4], [6]. Although subsequent works introduced temporal modeling via Long Short-Term Memory [23], or more recently, Vision Transformers (ViTs [24]) to capture context [25], these methods largely remain "fragment-level" classifiers. They excel at identifying short, localized anomalies but lack the holistic reasoning capabilities required to interpret a 12-hour procedure and generate a cohesive diagnostic report. Unlike general long-form video understanding [26], [27], CE video analysis is constrained by the extreme data sparsity and length issues described in Sec. I. Existing generic video understanding methods often fail to preserve these fine-grained medical details, necessitating a specialized, pathology-driven approach for effective data reduction. In contrast to these fragment-level classifiers,

EndoAgent transcends local detection by reformulating CE analysis as an event-driven agentic workflow. This approach explicitly handles data sparsity to enable holistic reasoning and end-to-end report generation over full-length procedures.

Regarding data resources, while public benchmarks such as KID [28], Kvasir-Capsule [2] and Galar [29] have been instrumental in advancing lesion detection, they predominantly consist of static images or short, fragmented clips. These datasets lack the continuous, long-range temporal context necessary to simulate a complete diagnostic workflow. Consequently, they are ill-suited for training and evaluating an autonomous agent designed for end-to-end CE report generation. To address this limitation, we constructed a large-scale dataset of full-length CE videos paired with expert-validated diagnostic reports, the details of which are provided in Sec. IV-A.

B. Multi-modal Large Language Models in Medicine

The advent of Large Vision-Language Models (VLMs), such as GPT-4o [30] and Gemini 2.5 [31], along with open-source counterparts like LLaVA [32] and Qwen2.5-VL [21], has revolutionized image captioning and visual reasoning. In the medical domain, models like LLaVA-Med [14] and RadFM [33] have emerged, finetuned on biomedical image-text pairs to improve domain knowledge. However, applying these models to CE video report generation presents significant hurdles: (1) Contextual Constraints: Most current VLMs are designed for static images or short clips [10], [12]. They cannot directly process the massive token sequences generated by approximately 120,000 frames of CE video due to limited context window size. (2) Hallucination Risks: General-purpose VLMs frequently suffer from hallucinations—fabricating plausible but factually incorrect details [34], [35]. In high-stakes medical diagnostics, such unfaithfulness poses critical safety risks, rendering raw VLM outputs unreliable for direct clinical adoption [36]. (3) Deficiency in Clinical Reasoning: While VLMs excel at passive visual description, they lack the active reasoning capabilities required for the diagnostic workflow. They struggle to execute essential cognitive steps, such as proactively filtering redundancy and initiating detailed verification for ambiguous regions. Unlike passive VLMs constrained by context limits, EndoAgent integrates a structured multi-modal memory with active tool invocation, allowing for precise, verified reasoning across the extensive temporal span of CE videos.

C. LLM-based Agents and Medical Reasoning

To transcend the limitations of passive perception, the field is witnessing a paradigm shift toward LLM-based autonomous agents. Unlike static classifiers, agents such as ReAct [16] and AutoGPT [37] can plan, interact with their environments, and orchestrate external tools to solve multi-step problems. While agentic workflows have shown success in general domains such as web navigation and coding [38], [39], recent studies have extended these capabilities to medical reasoning and multimodal clinical tasks. Meissa [40] distills structured agentic trajectories into a lightweight medical multimodal model capable of tool use and multi-agent interaction. ClinSeekAgent [41] actively acquires multimodal evidence by navigating raw electronic health records, querying medical knowledge sources, and invoking medical imaging tools. RadAgent [42] performs tool-augmented, stepwise interpretation and report generation for chest CT, and ClinicalAgents [43] combines dynamic multi-agent orchestration with a dual-memory architecture for clinical decision making.

Despite these advances, these systems primarily focus on discrete medical images, individual volumetric imaging studies, EHR-centered evidence seeking, or staged case-level clinical workflows; they do not target ultra-long, highly redundant endoscopic videos. The CE diagnostic workflow—which requires continuous scanning, redundancy reduction, and targeted visual verification of sparse findings—therefore presents a distinct agentic challenge. Consequently, EndoAgent is designed specifically for full-length CE video analysis and report generation, combining hierarchical Fast-Scan/Slow-Scan processing, specialized visual tools such as the Visual Verifier, and a structured multimodal memory to support evidence-grounded clinical reasoning.

D. Evaluation of Medical Report Generation

Evaluating the quality of generated medical reports remains an open challenge. Traditional Natural Language Generation (NLG) metrics [44] such as BLEU [45], ROUGE [46], and CIDEr [47] rely on lexical n-gram overlap, which often correlates poorly with clinical correctness [36]. For instance, a generated report might miss a critical "polyp" yet still achieve a high ROUGE score by perfectly reciting the dominant descriptions of normal mucosa. Conversely, a clinically precise diagnosis of "active hemorrhage" would be penalized against a ground truth using "bleeding", strictly due to lexical mismatch despite their semantic equivalence. To bridge this metric-reality gap, recent studies have pivoted towards "LLM-as-a-Judge" paradigms [36], [48], leveraging the semantic understanding of large language models. However, generic LLM judges often lack the specialized medical reasoning required to distinguish subtle pathologies. Our work contributes to this area by proposing a structured Medical Expert Judge evaluation protocol. By adopting a specialist persona and integrating Chain-of-Thought (CoT) [49] with Few-Shot learning [50], we guide the LLM to conduct a rigorous, step-by-step evaluation. This approach moves beyond surface-level fluency to deeply assess the consistency, completeness, and hallucination status of diagnostic reports.

III. Method

A. Problem Formulation and Overall Architecture

Formally, the task of CE video report generation is defined as learning a mapping $\mathcal{F} : \mathcal{V} \rightarrow \mathcal{R}$. The input CE video $\mathcal{V}$ in our dataset lasts approximately 12 hours, resulting in about 120,000 raw frames, while the output $\mathcal{R}$ is a structured diagnostic report. Directly modeling $\mathcal{F}$ with standard VLMs is intractable due to the contextual constraints, extreme sparsity of pathological features, and the high risk of hallucination as detailed in Sec. II-B. To resolve this, we propose to decompose

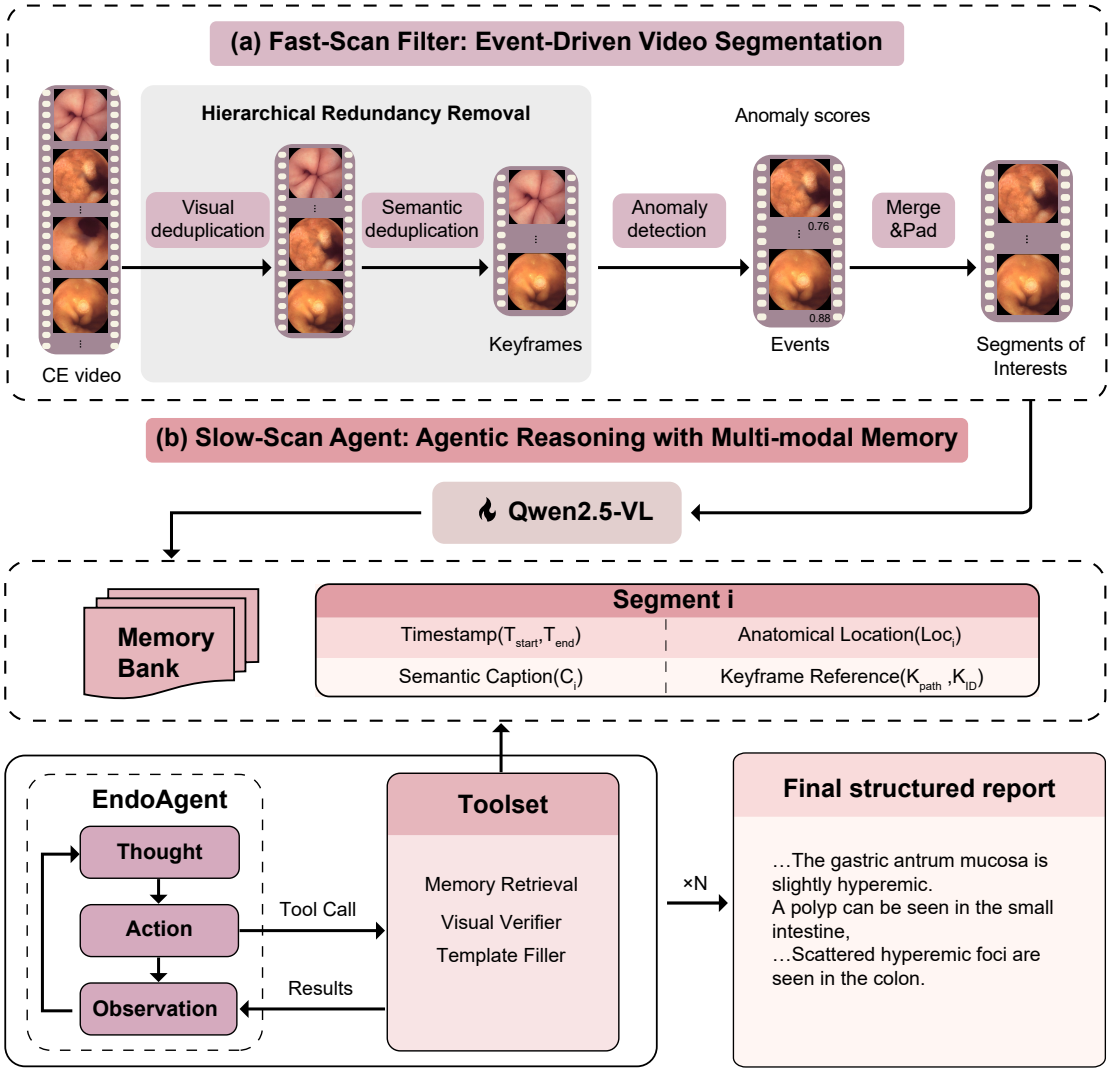


Fig. 2. **Overview of the EndoAgent Framework. (a) Fast-Scan Filter (Efficiency):** Raw CE video is processed via hierarchical deduplication and event-based partitioning. An anomaly scorer identifies high-value keyframes, which are then aggregated into *Context-rich Segments of Interest* to ensure sufficient temporal context for downstream analysis. **(b) Slow-Scan Agent (Reasoning):** The filtered segments are encoded into a structured *Memory Bank* by a LoRA-finetuned VLM. The Agent then executes a multi-step ReAct workflow (i.e., Observation-Thought-Action loop), invoking specialized tools (e.g., Visual Verifier) to perform evidence-based reasoning and populate the final clinical diagnostic report.

$\mathcal{F}$ into a hierarchical workflow grounded in Kahneman's Dual-Process Theory [19]. Specifically, as illustrated in Fig. 2, we map the intuitive System 1 to our Fast-Scan Filter for rapid redundancy removal, and the analytical System 2 to our Slow-Scan Agent for intelligent reasoning. This decomposition effectively bridges the gap between massive visual streams and precise clinical diagnosis.

B. Fast-Scan Filter: Event-Driven Video Segmentation

Beyond computational efficiency, the Fast-Scan Filter is explicitly designed to address the extreme pathological sparsity inherent in CE videos. Rather than serving as a uniform sampling strategy, it functions as a recall-first, anomaly-driven event proposal mechanism that transforms rare abnormalities into a manageable set of context-rich Segments of Interest suitable for downstream clinical reasoning.

1) Hierarchical Redundancy Removal: The raw CE video is characterized by two types of redundancy: visual stationarity and semantic repetition. To address this, we employ a coarse-to-fine filtering strategy: **(1) Visual Deduplication:** First, to handle camera stagnation or extremely slow movement, we use 64-bit Perceptual Hashing (pHash) [51]. This step operates at the pixel level to rapidly discard frames that are visually identical to their predecessors: a frame is removed if its Hamming distance to the last retained frame is at most $d = 8$. **(2) Semantic Deduplication:** However, visual distinctiveness does not imply semantic novelty. Due to gastric peristalsis, the camera angle may shift slightly, causing pixel values to change while the underlying anatomical view remains effectively the same. To eliminate this semantic redundancy, we employ a CLIP [52] (ViT-B/32) image encoder. We compute the cosine similarity between L2-normalized embeddings of consecutive

retained frames and discard those that fall at or above a semantic threshold $\theta = 0.95$.

After hierarchical redundancy removal, we obtain a set of approximately 20,000 keyframes, where each keyframe represents a visually and semantically distinct single frame.

2) Lightweight Anomaly Detection: The remaining 20,000 keyframes, while non-redundant, are still predominantly normal mucosa. To isolate potential pathologies, we process the keyframes using a lightweight anomaly detector, implemented as an EfficientNet-B4 binary classifier [53]. Adhering to the clinical safety principle [54] that "a false positive adds workload, but a false negative risks a patient's life," the detector is a high-recall trigger for event generation rather than a box-level lesion localizer.

This role is not interchangeable with recent YOLO-style WCE localizers [55], which optimize bounding-box mAP: a Fast-Scan miss cannot be recovered by Slow-Scan, so localization is deferred to the Visual Verifier. We therefore compare classification backbones under an identical binary protocol. EfficientNet-B4 is a mid-capacity scorer for low-resolution WCE frames—more expressive than B0/B2, yet much lighter than B7. It attains the highest recall against MobileNetV3-Large [56], [57] and ConvNeXt-Tiny [58], [59] (0.984 vs. 0.936 and 0.972). Full results are reported in the Supplementary Materials.

3) Event-Driven Segment Generation: Unlike standard video processing approaches that use uniform sampling [60], [61] which often risk missing sparse lesions, we propose a dynamic Event-Driven segmentation strategy to ensure the contextual integrity of pathological findings:

(1) Adaptive Event Triggering. We apply a sensitivity threshold $\tau = 0.7$ to the anomaly scores. A sequence of contiguous keyframes that exceed τ is instantiated as a candidate "Event". **(2) Temporal Merging.** To prevent fragmenting a single pathology into multiple broken clips, we merge distinct events if the time gap between them is less than a 2-second tolerance window. **(3) Contextual Padding.** Finally, to support the reasoning capabilities of the downstream Agent, we extend the boundaries of each merged event by adding temporal padding of ± 2 seconds from the original video. This step ensures that the final Segments of Interest include the onset and offset of the lesion, providing the necessary visual context for the Slow-Scan Agent to analyze the lesion's environment.

The final output is a set of roughly 300 Segments of Interest, which represent complete, context-rich pathological events, ready for fine-grained analysis.

By formulating pathological sparsity as a *recall-first* event detection problem rather than a frame-level classification task, the Fast-Scan Filter prioritizes high sensitivity. The combination of temporal merging and contextual padding effectively amplifies sparse pathological signals and minimizes false negatives, addressing the extreme data imbalance inherent in CE videos.

C. Slow-Scan Agent: Agentic Reasoning with Multi-modal Memory

Although the Fast-Scan Filter reduces redundancy, the resulting Segments of Interest remain raw pixel data, impeding direct analysis due to the semantic gap [34], [62] and context bottlenecks [63]. To bridge this, the Slow-Scan Agent functions as a ReAct-based system [16] grounded in a structured Multi-modal Memory Bank. By first transforming visual segments into precise textual indices via a domain-specific VLM, the agent performs holistic reasoning—iterating through a "Global Impression" scan, active retrieval, and tool-based verification—to synthesize a standardized diagnostic report.

1) Structured Memory Bank Construction: General-purpose VLMs often describe medical scenes generically or hallucinate non-existent features [36], [48]. To address this, we employ the Qwen2.5-VL [21] architecture as our visual backbone. Crucially, we finetune this model using Low-Rank Adaptation (LoRA) [20] on a specialized dataset of CE image-caption pairs (see Sec. IV-A). This adaptation process aligns the model's output space with professional medical terminology, ensuring that the generated captions are not only descriptive but diagnostically precise.

For each Segment of Interest from the Fast-Scan Filter, the Qwen2.5-LoRA model analyzes the keyframes and generates a detailed textual description, capturing features such as lesion type, size and bleeding status. We then organize these generated insights into a structured Memory Bank. This memory is designed as a relational list of dictionaries, where each entry encapsulates the full context of a potential pathological event. Formally, for each segment i, we construct a memory entry M_i containing:

- **Segment ID (ID_i):** A unique identifier for the segment.
- **Timestamp (T_{start}, T_{end}):** The precise temporal localization of the event in the original 12-hour video. This is critical for the Agent to correlate findings temporally.
- **Anatomical Location (Loc_i):** The predicted gastrointestinal region associated with the segment (Esophagus, Stomach, Small Bowel, Colon), obtained via a pre-trained anatomical video segmentation model [1] prior to agent initialization.
- **Semantic Caption (C_i):** The fine-grained clinical description generated by our Qwen2.5-LoRA model, serving as the primary context for the LLM's reasoning process.
- **Keyframe Reference (K_{path}, K_{ID}):** File paths to one or several representative frames selected from the corresponding Segment of Interest. By storing paths rather than pixel arrays, we keep the memory lightweight while reserving the ability for the Agent to re-examine the visual evidence using tool invocation if needed.

By converting the video into this text-indexed Memory Bank ($M = \{M_1, M_2, ..., M_k\}$), we effectively decouple visual perception from logical reasoning, allowing the Agent to read the entire 12-hour procedure as a structured document.

2) Domain-Specific Toolset Definition: To enable active investigation, the Agent is equipped with a suite of specialized, executable tools. Unlike general agents that use search engines, our tools are grounded in the medical visual domain:

(1) Memory Retrieval: A tool that allows the Agent to query the constructed Memory Bank using semantic keys (e.g., "Find all segments related to the stomach"). **(2) Anomaly Detector:** A verification tool that calls the anomaly detection model on

a specific frame to re-confirm if a visual region is abnormal. **(3) Visual Verifier:** A coarse-to-fine mechanism for resolving visual uncertainty. Upon invocation, the Qwen2.5-VL [21] backbone first loads the **detection-specialized LoRA** [20] (LoRA1) to localize the anomaly via bounding box prediction. Subsequently, the model **switches to the captioning-specialized LoRA** (LoRA2). The identified region is then cropped and combined with the original global view as a dual-view input to LoRA2 for fine-grained description. This design simulates a clinician's zoom-in inspection, enabling precise descriptions by integrating local high-resolution details with global anatomical context. **(4) Template Filler:** A rule-based tool that ensures the final output strictly adheres to the clinical format. It accepts verified findings as arguments and places them into the correct slots of the standardized report structure.

3) ReAct-based Diagnostic Reasoning Loop: The Agent's diagnostic reasoning is implemented as a structured ReAct-based decision loop, in which explicit reasoning steps are interleaved with tool execution and evidence observation. This ReAct loop operates within a sequential anatomical workflow, ensuring that the flexible reasoning adheres to the standardized reporting format.

a) Anatomy-Aware Memory Initialization: For each anatomical section of a standard CE report (e.g., Esophagus → Stomach → Small Bowel → Colon), the Agent first retrieves all memory entries associated with the target region. This aggregated regional memory provides a global contextual overview, analogous to a clinician's initial assessment of findings within a specific anatomical segment.

b) Section-Specific Diagnostic Objective Formulation: Conditioned on the regional memory, the Agent formulates a section-specific diagnostic objective, which defines the hypotheses to be examined (e.g., presence of ulcers, polyps, or bleeding) and constrains subsequent evidence gathering.

c) Iterative Observation-Thought-Action Loop: Guided by the diagnostic objective, the Agent enters an iterative Observation-Thought-Action loop. It selectively retrieves relevant memory entries, evaluates their confidence and consistency, and dynamically invokes specialized tools for visual re-perception when the description contains linguistic uncertainty markers or if the observed features are insufficient to explicitly rule out a pathology. This loop continues until sufficient and consistent evidence is accumulated to support a clinical decision or a maximum number of steps is reached.

d) Template-Constrained Recording: Once the diagnostic objective for the section is resolved, the Agent records only validated findings into the corresponding section of the diagnostic template. All entries are explicitly grounded in retrieved memory or verified visual observations.

e) Final Report Synthesis: After all anatomical sections are completed, the Agent synthesizes a diagnostic conclusion by integrating evidence across sections, ensuring coherence between localized findings and the final summary.

IV. Experiments

A. Datasets

To validate the effectiveness of our proposed EndoAgent System, we constructed a comprehensive dataset derived from real-world clinical cases.

Clinical Report Parsing and Standardization. The data for this study, comprising video recordings from a total of 850 patients, were obtained from the Department of Gastrointestinal Endoscopy at the Sixth Affiliated Hospital of Sun Yat-sen University, covering the period from August 2015 to September 2023. The study protocol was approved by the Institutional Review Board, and the requirement for informed consent was waived. The inclusion criteria comprised patients undergoing capsule endoscopy (CE) for obscure gastrointestinal bleeding or suspected small bowel lesions, utilizing the MiroCam capsule endoscopy system. We collected raw data consisting of CE videos and their corresponding PDF diagnostic reports. A parsing pipeline was developed to extract unstructured data and convert it into a structured JSON format. Each entry comprises: (1) *in-body recording time*, (2) *diagnostic findings and etiology*, (3) *diagnostic summary and suggestions*, and (4) *key lesion images* captured by gastrointestinal specialists during the review process.

To ensure the reliability of the ground truth, bounding box annotations for all key lesions and diagnostic reports were performed independently by two gastroenterologists, each with over five years of experience. Discrepancies were arbitrated by a third senior medical expert. The inter-rater agreement for lesion presence was calculated with a Kappa coefficient [64] of 0.82, indicating high reliability of the annotations. In total, we curated 895 high-quality video-report pairs. The videos feature an average duration of approximately 12 hours, a frame rate of 3 fps, and a resolution of 320×320. The dataset was randomly split into training and testing sets at a ratio of 8:2 at the patient level to ensure that no videos from the same patient appear in both sets.

TABLE I
DETAILS OF THE AUXILIARY DATASETS. THE TABLE DETAILS THE SAMPLE DISTRIBUTION ACROSS SPLITS.

Dataset	Category	Training	Validation	Testing
Anomaly Detection	Abnormal	2,339	815	966
	Normal	3,000	690	1,000
Medical Captioning	Abnormal†	2,806	–	338
	Normal	704	–	125
Lesion Detection	Abnormal†	527	–	80

† Includes bleeding, polyp, ulcer, and erosion.

Auxiliary Datasets for Tool Finetuning. To support the training of the specific tools within our agent system, we curated three auxiliary datasets (constructed from the training split of the main dataset to prevent data leakage). Specifically, we built the **(1) Anomaly Detection Dataset:** A balanced set of normal and abnormal frames used to train and test the lightweight anomaly detector in the Fast-Scan Filter. **(2) Medical Captioning Dataset:** Image-text pairs extracted from the parsed reports to fine-tune the Qwen2.5-VL model for domain-specific captioning. **(3) Lesion Detection Dataset:** A set of images with bounding box annotations (derived from doctor-marked regions) to train the detection capability of our

TABLE II
COMPARISON OF PERFORMANCE METRICS BETWEEN OUR METHOD AND BASELINE MODELS. BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. A HIGHER VALUE IMPLIES BETTER PERFORMANCE ACROSS ALL METRICS. **CONSIS.**: CONSISTENCY; **COMPL.**: COMPLETENESS; **HALLUC. MITIGATION**: HALLUCINATION MITIGATION.

Method	NLG Metrics				Clinical Evaluation Metrics				
	BLEU	METEOR	ROUGE-L	CIDEr-D	Consis.	Compl.	Halluc. Mitigation	Avg.	p-value
Video-LLaVA-7B [10]	0.157	0.231	0.287	1.632	1.85	2.16	1.33	1.78	<0.01
Qwen2.5-VL-7B [21]	0.178	0.276	0.313	1.865	2.14	2.63	1.79	2.19	<0.01
VideoAgent [65]	0.259	0.288	0.337	2.439	2.35	2.73	2.01	2.36	<0.01
LLaVA-Med-7B [14]	0.182	0.253	0.266	1.973	2.07	2.24	1.96	2.09	<0.01
GPT-5	0.246	0.280	0.324	2.059	2.66	3.29	2.18	2.71	<0.05
Gemini 2.5 Pro	**0.307**	**0.322**	**0.432**	2.551	2.48	3.42	1.67	2.52	<0.05
Ours	0.223	0.272	0.351	**2.645**	**3.15**	**3.69**	**3.34**	**3.39**	-

TABLE III
PERFORMANCE OF THE ANOMALY DETECTOR ON INTERNAL AND EXTERNAL DATASETS.

Test Dataset	Precision	Recall	F1 score	Accuracy
Anomaly Detection	0.941	0.984	0.962	0.973
Kvasir-Capsule†	0.892	0.927	0.909	0.934
WCEbleedGen†	0.838	0.843	0.840	0.869

† Results obtained via linear probing (fine-tuning the classification head).

TABLE IV
COMPARISON OF LLM-JUDGE SCORES BETWEEN UNIFORM SAMPLING STRATEGIES AT DIFFERENT DENSITIES AND OUR EVENT-DRIVEN STRATEGY.

Strategy	Consis.	Compl.	Halluc. Mit.	Avg.	p-value
Uniform Sampling Baselines					
Density 1×	2.36	2.17	2.43	2.32	<0.01
Density 2×	2.45	2.25	2.36	2.35	<0.01
Density 5×	2.51	2.33	2.39	2.41	<0.01
Proposed Method					
Event-Driven	**3.15**	**3.69**	**3.34**	**3.39**	-

model. The detailed statistics of these datasets are summarized in Table I.

B. Evaluation Metrics

We use n-gram metrics including BLEU [45], ROUGE [46], METEOR [66], and CIDEr [47] for benchmarking. To address their limitations discussed in (Sec. II-D), we also introduce a robust evaluation protocol based on the **"LLM-as-a-Judge"** paradigm. Unlike generic metrics, this LLM-based judge is designed to comprehensively assess the generated reports for clinical correctness and reasoning logic. And it is blinded to method identity and evaluates all reports in randomized order using the same prompt and scoring rubric.

We constructed a structured prompt to instruct the GPT-4o to adopt the persona of a senior digestive gastrointestinal specialist and a quality assessment authority. The LLM is tasked with comparing the Generated Report against the Ground Truth (GT) verified by gastrointestinal specialists. To ensure robustness, we incorporated CoT reasoning and Few-Shot Learning examples within the prompt, guiding the model to perform a step-by-step analysis before assigning scores.

We defined three clinically oriented dimensions to quantify the report quality. Each dimension is scored on a Likert scale from 1 (poor) to 5 (excellent), with 0.5-point intervals to capture fine-grained distinctions:

(1) Consistency. Measures whether the description of lesions (location, type, severity) in the generated report factually matches the GT. High scores indicate that critical findings (e.g., ulcers, polyps) are described accurately without contradiction. **(2) Completeness.** Evaluates the extent to which the generated report covers all critical lesions mentioned in the GT. It penalizes the omission of clinically significant findings, particularly in the small and large intestines. **(3) Hallucination Mitigation.** Assesses the presence of fabricated information. A high score signifies that the generated report does not include hallucinated lesions that are absent in the GT.

The final score is calculated as the arithmetic mean of the three dimensions. The detailed prompt, including specific scoring rubrics and examples, is provided in the Supplementary Materials.

C. Implementation Details

Our implementation is based on PyTorch and the models are trained on NVIDIA A100 GPUs with 40GB memory. The lightweight Anomaly Detector (a customized EfficientNet-B4 [53]) was trained for 10 epochs with a learning rate of 0.001 on the Anomaly Detection Dataset. MobileNetV3-Large and ConvNeXt-Tiny follow the same training protocol; full comparisons are in the Supplementary Materials. At inference, LoRA1 and LoRA2 are loaded exclusively. We trained two separate LoRA [20] adapters for the Qwen2.5-VL [21] backbone: (1) a detection-specialized LoRA adapter (LoRA1) fine-tuned on the Lesion Detection Dataset, and (2) a captioning-specialized LoRA adapter (LoRA2) fine-tuned on the Medical Captioning Dataset. Both adapters utilized the optimal hyperparameters (Config B) identified in our ablation study. The agent's reasoning loop was driven by DeepSeek V3 [67] with a temperature of 0 to ensure deterministic outputs during reporting.

TABLE V
ABLATION STUDY ON THE CONTRIBUTION OF INDIVIDUAL AGENT COMPONENTS.

System Configuration	Consis.	Compl.	Halluc. Mitigation	Avg.	p-value
Baseline (Direct VLM Inference)	1.24	1.30	1.16	1.27	<0.001
+ Memory Bank	2.23	2.46	2.39	2.36	<0.01
+ Memory Bank + Template Filler	2.38	2.67	2.45	2.50	<0.01
+ Memory Bank + Visual Verifier	2.91	3.48	3.12	3.17	<0.05
+ All Tools (Full EndoAgent)	**3.15**	**3.69**	**3.34**	**3.39**	-

TABLE VI
ABLATION STUDY ON LORA HYPERPARAMETERS FOR THE QWEN2.5-VL CAPTIONING MODEL. WE EVALUATE THE IMPACT OF RANK (r), ALPHA (α), AND DROPOUT RATE ON CAPTIONING QUALITY.

Configuration	Hyperparameters			Performance Metrics					
	Rank (r)	Alpha (α)	Dropout	BLEU	METEOR	ROUGE-L	CIDEr-D	LLM-Judge	p-value
Baseline (No LoRA)	-	-	-	0.13	0.26	0.23	0.86	0.74	<0.001
Config A	8	16	0.05	0.57	0.86	0.73	5.43	4.09	<0.01
Config B (Ours)	**16**	**32**	**0.05**	**0.79**	**0.94**	**0.86**	**8.24**	**4.81**	-
Config C	16	32	0.10	0.68	0.89	0.79	6.69	4.23	<0.05

D. Comparison with the Existing Methods

To rigorously evaluate the efficacy of EndoAgent, we compare it against several strong Generalist VLMs, specialized Medical VLMs, and an agentic baseline. Due to the prohibitive length of CE videos, end-to-end processing is impossible for current baselines. Therefore, we adapt the input formats as follows:

1) Baseline Configurations: **(1) Commercial Generalist VLMs.** Due to token costs and context limits, we evaluate models like Gemini 2.5 Pro [31] using textual summaries. We feed segment captions from our Slow-Scan Agent directly into these models to assess their diagnostic reasoning and report synthesis capabilities. **(2) Open-Source VLMs.** To adapt models like Video-LLaVA [10] to 12-hour sequences, we provide a condensed video containing 500 frames uniformly sampled from the Segments of Interest identified by our Fast-Scan Filter, ensuring a fair comparison of visual understanding. **(3) Biomedical VLMs.** As a static image model, LLaVA-Med [14] lacks temporal processing. We evaluate its domain-specific recognition by feeding it the same sampled frames individually to assess its performance on isolated clinical findings. **(4) EndoAgent (Ours).** Our complete system integrates a hierarchical Fast-Scan Filter with a Slow-Scan Agent. This architecture enables dynamic tool invocation and evidence-based reasoning to generate comprehensive diagnostic reports from long-form endoscopy videos.

2) Quantitative Results: As presented in Table II, our EndoAgent System demonstrates superior performance in clinical diagnostic capabilities. Notably, it achieves a Total Average Score of 3.39 under the rigorous LLM-as-a-Judge evaluation, significantly outperforming the strongest commercial baseline (GPT-5, 2.71). This substantial margin underscores the critical role of the active tool usage mechanism in enhancing completeness and suppressing hallucinations.

Conversely, regarding traditional n-gram metrics, EndoAgent scores slightly lower than Gemini 2.5 Pro. This discrepancy is attributed to the inherent limitations of these metrics, which prioritize surface-level lexical overlap over semantic correctness. Generalist models tend to generate fluent, boilerplate text that maximizes n-gram matching but may miss specific pathological details. In contrast, our model optimizes for diagnostic precision, as evidenced by its leading CIDEr-D score—a metric known to correlate better with human judgment in captioning tasks [47]. A detailed qualitative comparison illustrating this metric misalignment is provided in the subsequent section.

3) Generalization Analysis on External Public Datasets: To verify generalizability, we conducted validation on the Kvasir-Capsule dataset [2] and WCEbleedGen dataset [68]. We standardized all images to 224×224 to unify input spaces from varying hardware resolutions (256×256 to 512×512). Addressing the inherent class imbalance, we constructed a class-balanced test set (approximately 1:1 ratio) containing representative pathologies (bleeding, polyp, ulcer, erosion) and normal mucosa. Furthermore, to mitigate domain shifts caused by different acquisition sensors, we employed a few-shot linear probing strategy: freezing the pre-trained Anomaly Detector backbone and fine-tuning the classification head using only 5% of the external training data. As presented in Table III, the resulting F1 score and Accuracy demonstrate the strong robustness and transferability of our learned visual representations across different datasets.

E. Ablation Studies

We conducted extensive ablation studies to verify the contribution of each component in EndoAgent.

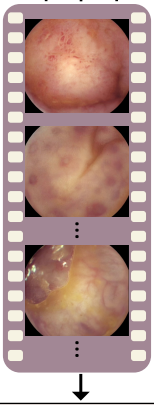
1) Impact of Sampling Strategy: To validate the effectiveness of our Fast-Scan Filter, we compared our Event-Driven Segmentation strategy against Uniform Sampling baselines at varying densities. For the baseline Uniform ($1\times$), frames were sampled at fixed intervals to match the total frame count of our method. The Uniform $2\times$ and $5\times$ variants utilize denser sampling rates, resulting in total frame counts that are twice and five times that of the proposed method, respectively.

Ground Truth

Gastric Antrum	Duodenum	Jejunum	Ileum	Colon
The mucosa is slightly hyperemic.	The mucosa shows scattered patchy hyperemic and erosive lesions.	Occasional mucosal hyperemic erythema is observed.	The lower segment, especially the terminal ileum, shows multiple mucosal hyperemic foci with scattered shallow small ulcers.	Observation is affected by feces; the visible colon shows multiple hyperemic erythema.

Baseline

Gastric Antrum	Duodenum
The mucosa is smooth; no abnormalities seen.	Mild hyperemia is visible.
Jejunum & Ileum	**Colon**
The mucosa is smooth throughout with clear vascular patterns; no hyperemia, erosion, or ulcers are seen.	A polyp is visible, with surface hyperemia.



Gemini 2.5 Pro

Gastric Antrum	Duodenum
The mucosa is slightly hyperemic.	Scattered hyperemia is visible.
Jejunum	**Ileum**
The mucosa is mostly smooth.	Scattered patchy hyperemic foci are visible; the local mucosa appears rough, but no obvious ulcers are seen.
Colon	
The mucosa is smooth; no obvious abnormalities are seen.	

Ours

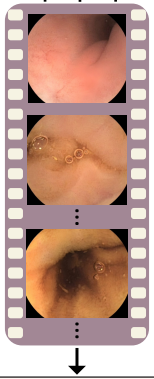
Gastric Antrum	Duodenum	Jejunum	Ileum	Colon
The mucosa is slightly hyperemic.	The mucosa shows mild hyperemia.	Visible mucosa shows scattered patchy hyperemic foci with local redness.	The mucosa shows segmental hyperemia, edema, and erosion, presenting as inflammatory lesions.	Scattered hyperemic foci are seen.

Ground Truth

Gastric Antrum	Duodenum	Jejunum & Ileum	Colon
Slightly hyperemic.	No abnormalities seen.	Observation of a small portion was affected by intestinal fluid; no abnormalities were seen in the visible parts.	Observation of parts of the colon was affected by feces; no abnormalities were seen in the visible colon.

Baseline

Gastric Antrum	Duodenum
The mucosa is slightly hyperemic.	The mucosa shows diffuse hyperemic lesions.
Jejunum & Ileum	**Colon**
No obvious mucosal abnormalities seen.	Hyperemic lesions are visible in the colon.



Gemini 2.5 Pro

Gastric Antrum	Duodenum
The mucosa is slightly hyperemic.	The mucosa is smooth; no abnormalities seen.
Jejunum & Ileum	**Colon**
No abnormalities seen.	The mucosa appears slightly rough; scattered hyperemic foci are visible.

Ours

Gastric Antrum	Duodenum	Jejunum & Ileum	Colon
Slightly hyperemic.	The mucosa is smooth; no abnormalities seen.	No abnormalities seen.	No abnormalities seen.

Fig. 3. Examples of CE video report generated by different methods (Baseline, Gemini 2.5 Pro and our EndoAgent). The **green text** in GT indicates critical findings missed by all models, and the **red text** in the methods represents incorrect diagnoses.

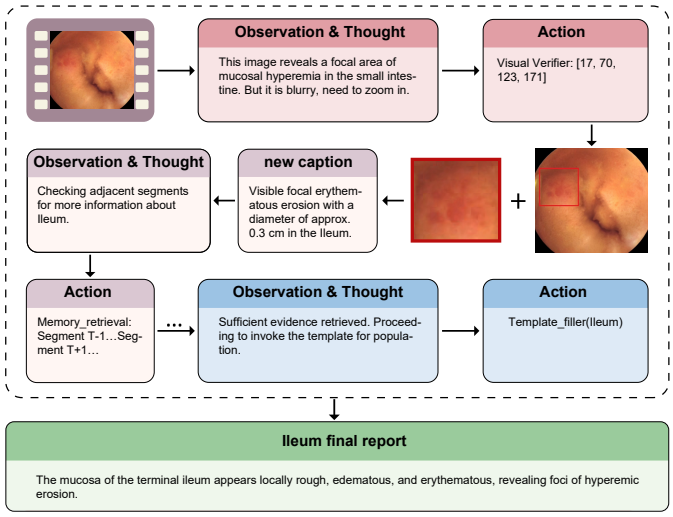


Fig. 4. Illustration of the "Observation-Thought-Action" reasoning loop. The Agent resolves visual ambiguity in the ileum by dynamically invoking the Visual Verifier to reveal fine-grained lesion details (e.g., 0.3 cm erosion) and querying Memory Retrieval for temporal context, ultimately synthesizing a precise report via the Template Filler.

As shown in Table IV, the Event-Driven approach yields a substantial improvement in the Total Average Score. Notably, the Completeness score sees the largest gain, confirming that uniform sampling frequently misses sparse, transient pathological events, whereas our anomaly-guided selection successfully retains these critical diagnostic moments. Interestingly, simply increasing the frame count does not linearly improve performance and may even degrade Hallucination Mitigation due to the introduction of excessive redundant context.

2) Contribution of Agent Components: We performed a progressive ablation study to isolate the impact of each component within the EndoAgent System (Table V).

(1) Baseline vs. Memory. The introduction of the Memory Bank provides the most significant performance leap (Total Avg. from 1.27 to 2.36), proving that raw video data must be converted into structured semantic indices for effective reasoning. **(2) Impact of Tools.** Adding the Template Filler marginally improves consistency and completeness by enforcing clinical standards. Crucially, the integration of the Visual Verifier Tool leads to a dramatic increase in Hallucination Mitigation and Completeness. This empirically validates that active visual re-perception is essential for verifying ambiguous findings and reducing false positives. **(3) Synergy.** The full system achieves the highest scores across all metrics, demonstrating that the combination of structured reporting and active verification yields synergistic benefits.

3) Optimization of LoRA Hyperparameters: Finally, we analyzed the sensitivity of the Slow-Scan captioning model to LoRA fine-tuning configurations. Since the quality of the diagnostic report primarily depends on semantic reasoning, we focused our quantitative ablation on the Captioning Adapter performance. As presented in Table VI, the Base Model (No LoRA) performs poorly on medical captioning, highlighting the domain gap between general web data and CE imagery.

Fine-tuning with LoRA significantly bridges this gap. Comparing Configs A, B, and C reveals that increasing the rank from 8 to 16 enhances model capacity, resulting in better detail capture (Config B vs. A). However, increasing the dropout rate to 0.10 (Config C) appears to underfit the data slightly, leading to a drop in performance. Consequently, we adopted Config B ($r = 16, \alpha = 32$, dropout $= 0.05$) as our optimal setting, which achieves the highest score, ensuring high-fidelity visual descriptions for the downstream agent.

F. Runtime and Memory

We profile EndoAgent on 10 randomly sampled full-length CE videos (approximately 12 hours each) using an NVIDIA A100 (40 GB), Qwen2.5-VL-7B in FP16, a CLIP/anomaly batch size of 64, and a captioning batch size of 16. Averaged over these videos, end-to-end latency is about 11 min per case, with a peak GPU memory of 21 GB during Memory Bank construction; Fast-Scan takes about 2.3 min and 2 GB. Stage-wise measurements are reported in the Supplementary Materials. This offline cost is far below a typical 1–2 hour manual review and is not intended as a real-time viewing replacement.

G. Qualitative Analysis

In Fig. 3, we present a qualitative comparison of reports generated by the Baseline, Gemini 2.5 Pro, and our proposed EndoAgent against the GT.

As illustrated in the first case, while Gemini 2.5 Pro generates fluent text—likely boosting its NLG scores—it suffers from severe hallucinations, explicitly stating "no obvious ulcers" in the ileum despite the GT's finding of "scattered shallow small ulcers." Furthermore, both the baseline and Gemini 2.5 Pro exhibit significant incompleteness, omitting critical "hyperemic erythema" in the jejunum and colon by defaulting to generic descriptions of "smooth mucosa." In contrast, EndoAgent achieves superior clinical validity, accurately capturing the "segmental hyperemia and erosion" while ensuring comprehensive coverage of findings across all anatomical segments.

The second case further highlights the issue of hallucination in existing methods. The GT indicates that the view was affected by intestinal fluid and feces, with "no abnormalities seen." However, both the Baseline and Gemini 2.5 Pro suffer from severe hallucinations, fabricating "diffuse hyperemic lesions" or "scattered hyperemic foci" where none exist. Conversely, EndoAgent correctly reasons that the visual disturbances (fluid/feces) do not constitute a lesion, correctly concluding "No abnormalities seen."

To further elucidate the active reasoning capability of our system, Fig. 4 visualizes the "Observation-Thought-Action" loop within the Slow-Scan Agent during the analysis of a specific anatomical section (specifically, the Ileum).

When faced with visual uncertainty regarding a potential lesion, instead of hallucinating a detail to fill the gap, the Agent actively decides to invoke the Visual Verifier. This tool simulates a "zoom-in" inspection, refining the blurry visual representation to reveal a precise description (e.g., visible

focal erythematous erosion with a diameter of approx. 0.3 cm). Subsequently, the Agent triggers a Memory Retrieval action, querying adjacent segments to analyze the temporal context. Finally, after aggregating sufficient evidence, the Agent invokes the Template Filler to populate the standard clinical template, synthesizing the final structured report for the Ileum section: "The mucosa of the terminal ileum appears locally rough, edematous, and erythematous, revealing foci of hyperemic erosion."

Overall, these cases demonstrate that EndoAgent effectively mitigates the hallucinations common in generalist VLMs and ensures the clinical precision required for medical diagnostics.

V. Limitations and Future Directions

Despite its promising performance, EndoAgent faces several limitations. First, the hierarchical architecture introduces the risk of cascading errors [69]; specifically, the Slow-Scan Agent is heavily dependent on the Fast-Scan Filter, such that false negatives at the filtering stage may not be recovered. Second, visual ambiguity remains a challenge, as the Visual Verifier tool is limited by raw image quality (e.g., turbid fluid or poor lighting). Third, the active reasoning loop incurs higher inference latency [70] compared to single-pass models.

Future work will address these challenges along two primary directions. To mitigate reasoning latency, we aim to distill the system's reasoning traces into a specialized, efficient end-to-end model. Furthermore, to transcend the limits of pure visual analysis, we plan to integrate complementary patient-level clinical data, such as Electronic Health Records (EHR), enabling more comprehensive, evidence-based clinical case management.

VI. Conclusion

In this paper, we propose EndoAgent, the first hierarchical agent system tailored for long-form capsule endoscopy video analysis. By simulating the "Fast and Slow" dual-process cognitive mechanism, we effectively resolve the conflict between context limitations and the extreme sparsity of pathological features in long CE video processing. Experimental results demonstrate that EndoAgent significantly surpasses existing generalist and medical-specific models in terms of Consistency, Completeness, and Hallucination Mitigation in clinical report generation. This superior performance is driven by the efficient filtering of the Fast-Scan Filter and the active tool invocation of the Slow-Scan agent—specifically the visual verification mechanism that mimics a gastroenterologist's zoom-in inspection. Furthermore, the "LLM-as-a-Judge" evaluation protocol established in this study reveals the inadequacy of traditional text generation metrics in measuring clinical validity. Future work will aim to distill this multi-step reasoning workflow into an end-to-end model and further integrate multi-modal data, such as electronic health records, to achieve more comprehensive and evidence-based automated diagnosis.

References

[1] X. Zhao, C. Fang, F. Gao, D.-J. Fan, X. Lin, and G. Li, "Deep transformers for fast small intestine grounding in capsule endoscope video," in *ISBI*. IEEE, 2021, pp. 150–154.

[2] P. H. Smedsrud *et al.*, "Kvasir-capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, no. 1, p. 142, 2021.

[3] M. Pennazio *et al.*, "Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European society of gastrointestinal endoscopy (esge) clinical guideline," *Endoscopy*, vol. 47, no. 4, pp. 352–376, 2015.

[4] X. Xing, Y. Yuan, and M. Q.-H. Meng, "Zoom in lesions for better diagnosis: Attention guided deformation network for wce image classification," *IEEE TMI*, vol. 39, no. 12, pp. 4047–4059, 2020.

[5] J. Fu *et al.*, "Ipnet: An interpretable network with progressive loss for whole-stage colorectal disease diagnosis," *IEEE TMI*, 2024.

[6] S. Jain *et al.*, "Coinnet: A convolution-involution network with a novel statistical attention for automatic polyp segmentation," *IEEE TMI*, vol. 42, no. 12, pp. 3987–4000, 2023.

[7] X. Liu and Y. Yuan, "A source-free domain adaptive polyp detection framework with style diversification flow," *IEEE TMI*, vol. 41, no. 7, pp. 1897–1908, 2022.

[8] Y. Pang, Y. Long, Z. Chen, Y. Hu, H. Chen, and Q. Wang, "Endoscopic adaptive transformer for enhanced polyp segmentation in endoscopic imaging," *IEEE TMI*, 2025.

[9] S. Oh *et al.*, "Video analysis of small bowel capsule endoscopy using a transformer network," *Diagnostics*, vol. 13, no. 19, p. 3133, 2023.

[10] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.

[11] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," in *ACL*, 2024, p. 12585–12602.

[12] Y. Weng, M. Han, H. He, X. Chang, and B. Zhuang, "Longvlm: Efficient long video understanding via large language models," in *ECCV*. Springer, 2024, pp. 453–470.

[13] X. Shen *et al.*, "Longvu: Spatiotemporal adaptive compression for long video-language understanding," *arXiv preprint arXiv:2410.17434*, 2024.

[14] C. Li *et al.*, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *NeurIPS*, vol. 36, pp. 28 541–28 564, 2023.

[15] K. Zhang *et al.*, "A generalist vision–language foundation model for diverse biomedical tasks," *Nature Medicine*, pp. 1–13, 2024.

[16] S. Yao *et al.*, "React: Synergizing reasoning and acting in language models," in *ICLR*, 2022, pp. –.

[17] Z. Xi *et al.*, "The rise and potential of large language model based agents: A survey," *Science China Information Sciences*, vol. 68, no. 2, p. 121101, 2025.

[18] Q. Kong *et al.*, "Ai agent-based discovery of d-enantiomeric antimicrobial peptides against multidrug-resistant bacterial infection," *Biomaterials*, p. 123927, 2025.

[19] D. Kahneman, "Thinking, fast and slow," *Farrar, Straus and Giroux*, 2011.

[20] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *ICLR*, 2022, pp. –.

[21] S. Bai *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[25] M. Farhad, M. M. Masud, A. Beg, A. Ahmad, and L. Ahmed, "A review of medical diagnostic video analysis using deep learning techniques," *Applied Sciences*, vol. 13, no. 11, 2023.

[26] Y. Wang *et al.*, "Internvideo2. 5: Empowering video mllms with long and rich context modeling," *arXiv preprint arXiv:2501.12386*, 2025.

[27] Y. Luo *et al.*, "Video-rag: Visually-aligned retrieval-augmented long video comprehension," *arXiv preprint arXiv:2411.13093*, 2024.

[28] S. Jain, A. Ojha, and A. Seal, "Kid dataset," dec 2024.

[29] M. Le Floch *et al.*, "Galar-a large multi-label video capsule endoscopy dataset," *Scientific Data*, vol. 12, no. 1, p. 828, 2025.

[30] O. teams, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[31] G. Comanici *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[32] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *NeurIPS*, vol. 36, pp. 34 892–34 916, 2023.

[33] C. Wu, X. Zhang, Y. Zhang, H. Hui, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology by leveraging web-scale

2d&3d medical data," *Nature Communications*, vol. 16, no. 1, p. 7866, 2025.
[34] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," *arXiv preprint arXiv:2305.10355*, 2023.
[35] T. Guan *et al.*, "Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," in *CVPR*, 2024, pp. 14 375–14 385.
[36] B. Khanal *et al.*, "Hallucination-aware multimodal benchmark for gastrointestinal image analysis with large vision-language models," in *MICCAI*. Springer, 2025, pp. 235–245.
[37] H. Yang, S. Yue, and Y. He, "Auto-gpt for online decision making: Benchmarks and additional opinions," *arXiv preprint arXiv:2306.02224*, 2023.
[38] H. He *et al.*, "Webvoyager: Building an end-to-end web agent with large multimodal models," *arXiv preprint arXiv:2401.13919*, 2024.
[39] J. Yang *et al.*, "Swe-agent: Agent-computer interfaces enable automated software engineering," *NeurIPS*, vol. 37, pp. 50 528–50 652, 2024.
[40] Y. Chen, X. Bai, Y. Pan, Z. Zhou, and A. Yuille, "Meissa: Multi-modal medical agentic intelligence," *arXiv preprint arXiv:2603.09018*, 2026.
[41] J. Wu *et al.*, "Clinseekagent: Automating multimodal evidence seeking for agentic clinical reasoning," *arXiv preprint arXiv:2605.20176*, 2026.
[42] M. Roschewitz *et al.*, "Radagent: A tool-using AI agent for stepwise interpretation of chest computed tomography," *arXiv preprint arXiv:2604.15231*, 2026.
[43] Z. Ge, H. Li, Y. Wang, N. Hu, C. J. Zhang, and Q. Li, "Clinicalagents: Multi-agent orchestration for clinical decision making with dual-memory," in *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026.
[44] C. Dong *et al.*, "A survey of natural language generation," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–38, 2022.
[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, Jul. 2002, pp. 311–318.
[46] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. ACL, 2004, pp. 74–81.
[47] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
[48] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.
[49] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *NeurIPS*, vol. 35, pp. 24 824–24 837, 2022.
[50] T. B. Brown *et al.*, "Language models are few-shot learners," *NeurIPS*, vol. 33, pp. 1877–1901, 2020.
[51] C. Zauner, "Implementation and benchmarking of perceptual image hash functions," 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:17075066
[52] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PmLR, 2021, pp. 8748–8763.
[53] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*. PMLR, 2019, pp. 6105–6114.
[54] B. R. Jackson, "The dangers of false-positive and false-negative test results: False-positive results as a function of pretest probability," *Clinics in Laboratory Medicine*, vol. 28, no. 2, pp. 305–319, 2008, quality in Laboratory Medicine.
[55] A. Sahafi, A. Koulaouzidis, and M. Lalinia, "Polypoid lesion segmentation using YOLO-V8 network in wireless video capsule endoscopy images," *Diagnostics*, vol. 14, no. 5, p. 474, 2024.
[56] A. Howard *et al.*, "Searching for MobileNetV3," in *ICCV*, 2019, pp. 1314–1324.
[57] J. Chen, K. Xia, and Z. Zhang, "Establishing an AI model and application for automated capsule endoscopy recognition based on convolutional neural networks (with video)," *BMC Gastroenterology*, vol. 24, p. 394, 2024.
[58] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *CVPR*, 2022, pp. 11 976–11 986.
[59] X. Wang, Q. Cui, X. Zhang, J. Wang, Z. Wang, and J. Song, "Wireless capsule endoscopy image classification model based on improved ConvNeXt," *Journal of Computer Applications*, vol. 45, no. 6, pp. 2016–2024, 2025.
[60] L. Chen *et al.*, "Sharegpt4video: Improving video understanding and generation with better captions," *NeurIPS*, vol. 37, pp. 19 472–19 495, 2024.
[61] Y. Fan *et al.*, "Videoagent: A memory-augmented multimodal agent for video understanding," in *ECCV*. Springer, 2024, pp. 75–92.
[62] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE TPAMI*, vol. 22, no. 12, pp. 1349–1380, 2002.
[63] K. Tao, C. Qin, H. You, Y. Sui, and H. Wang, "Dycoke: Dynamic compression of tokens for fast video large language models," in *CVPR*, 2025, pp. 18 992–19 001.
[64] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37 – 46, 1960.
[65] X. Wang, Y. Zhang, O. Zohar, and S. Yeung-Levy, "Videoagent: Long-form video understanding with large language model as agent," in *ECCV*, 2024, pp. 58–76.
[66] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, Jun. 2005, pp. 65–72.
[67] DeepSeek-AI, "Deepseek-v3 technical report," 2024.
[68] P. Handa *et al.*, "Wcebleedgen: A wireless capsule endoscopy dataset and its benchmarking for automatic bleeding classification, detection, and segmentation," 2024.
[69] O. Narushynska, A. Doroshenko, V. Teslyuk, V. Antoniv, and M. Arzubov, "Enhancing hierarchical classification in tree-based models using level-wise entropy adjustment," *Big Data and Cognitive Computing*, vol. 9, no. 3, 2025.
[70] H. Liu *et al.*, "Medmmv: A controllable multimodal multi-agent framework for reliable and verifiable clinical reasoning," *arXiv preprint arXiv:2509.24314*, 2025.

Supplementary Materials for

EndoAgent: A Hierarchical Multi-Modal Agent System
for Capsule Endoscopy Video Analysis

Table S1: Deduplication on a training-video subset. Bold: recall $\geq 98\%$, lowest retention.

pHash d	CLIP θ	Retention	Frame recall
6	0.95	29.8%	99.7%
6	0.98	34.2%	99.8%
8	0.92	14.5%	96.4%
8	**0.95**	**18.4%**	**98.7%**
8	0.98	21.7%	99.1%
10	0.95	11.3%	95.6%

Table S2: Anomaly threshold τ on the Anomaly Detection validation split. Bold: selected.

τ	Precision	Recall	F1
0.6	0.926	0.992	0.958
0.7	**0.944**	**0.989**	**0.966**
0.8	0.948	0.957	0.952

S1 Fast-Scan Threshold Selection

This section provides the threshold comparison experiments requested for the Fast-Scan Filter. Deduplication thresholds (d, θ) are selected on a subset of the training videos; τ is selected on the Anomaly Detection validation split. The held-out test set is evaluated only with the final triplet $(d, \theta, \tau) = (8, 0.95, 0.7)$.

S1.1 Visual and Semantic Deduplication

A 64-bit perceptual hash (pHash) first compares each frame with the last retained frame and discards it if the Hamming distance is at most d. A CLIP (ViT-B/32) image encoder then discards a frame if the cosine similarity of L2-normalized embeddings to the last retained frame is at least θ. Thus, a larger d is more aggressive, whereas a larger θ is more conservative.

We report the keyframe *retention rate* (kept frames / raw frames) and *frame-level lesion recall*: a ground-truth lesion frame is counted as recalled if it is retained, or if a retained frame exists within a ± 1 second window. Table S1 lists representative (d, θ) combinations. Among settings with frame recall $\geq 98\%$, we select the most compact pair $(d, \theta) = (8, 0.95)$.

S1.2 Anomaly Threshold τ

After deduplication, keyframes with anomaly score $\geq \tau$ trigger events. We select τ on the Anomaly Detection *validation* split. Candidates must satisfy recall $\geq 98\%$; among them we take the largest F1 score. Table S2 shows that $\tau = 0.7$ attains the best F1 while remaining above the recall floor. $\tau = 0.8$ is discarded because recall falls below 98%. The held-out test-set metrics of the frozen detector ($\tau = 0.7$) are reported separately in the main paper.

Table S3: Backbone comparison on the Anomaly Detection test set. *: highest recall.

Backbone	Params	Prec.	Rec.	F1	Acc.
EfficientNet-B4*	19.3	0.941	**0.984**	0.962	**0.973**
MobileNetV3-Large	5.4	0.893	0.936	0.913	0.921
ConvNeXt-Tiny	28.6	**0.957**	0.972	**0.964**	0.967

Table S4: Runtime and peak GPU memory (mean over 10 videos). End-to-end memory is the maximum across stages.

Stage	Time	GPU
Fast-Scan	2.3 min	2 GB
Memory construction	4.8 min	21 GB
Agentic reasoning	3.6 min	17 GB
End-to-end	11 min	21 GB

S1.3 Lightweight Backbone Comparison

Fast-Scan is a frame-level binary event trigger, not a box-level lesion detector. We therefore compare classification backbones on the Anomaly Detection Dataset rather than YOLO-style localizers. All models use the same patient-level split, 224×224 input, 10 training epochs, and learning rate 0.001. MobileNetV3-Large follows a recent CE CNN comparison; ConvNeXt-Tiny has been used for WCE classification (citations in the main paper). Table S3 reports test-set metrics. We retain EfficientNet-B4, the backbone with the highest recall.

S2 Deployment Efficiency

We profile wall-clock time and peak GPU memory on 10 randomly sampled full-length CE videos (approximately 12 hours each) using an NVIDIA A100 (40 GB). Table S4 reports measured means over these 10 videos, rounded for readability. CLIP and the anomaly detector use batch size 64; Memory Bank captioning uses Qwen2.5-VL-7B in FP16 with batch size 16. Detection LoRA and captioning LoRA are loaded exclusively, so their footprints are not additive. End-to-end latency is about 11 min per video; peak memory is 21 GB during Memory Bank construction. Agentic reasoning peaks at 17 GB when the Visual Verifier is invoked and is substantially lower otherwise. The gap between the sum of mean stage times and the end-to-end mean is attributable to video decoding and I/O.

S3 LLM-as-a-Judge Prompt

The scoring protocol was finalized with gastroenterologists. They defined which findings count as key versus minor errors. We then calibrated the Likert anchors and few-shot examples on a small set of report pairs until the automatic scores matched the clinicians' judgments, and froze the prompt for all subsequent evaluation.

The English prompt below is the protocol used with GPT-4o. Brown placeholders mark fields filled at inference time.

LLM-as-a-Judge system prompt

You are a senior gastroenterologist specializing in capsule endoscopy. Compare the Generated report with the human-written Ground Truth (GT) by clinical meaning, matching lesions by location, type, and severity rather than wording. Treat clinically equivalent expressions as matches; differences in activity or severity are severity mismatches.

[Scope]

- Focus: clinically significant key lesions in the small bowel and colon (e.g., ulcer, polyp, mass, erosion, inflammation, bleeding, vascular malformation, diverticulum).
- Minor: non-severe esophageal/gastric findings, isolated hyperemia/erythema, and low-risk findings modified by "mild / punctate / scattered / slight / occasional". Lesion type and clinical risk take priority; modifiers do not downgrade high-risk lesions.
- Count one lesion entity as "anatomical location + lesion type". Count quantity, extent, or severity differences within one entity once; count different lesion types separately. One error may affect multiple dimensions according to their definitions.
- Ignore viewing-condition differences (fluid/debris). Generated negatives are not hallucinations; if they deny a GT lesion, count a mismatch and omission. Skip all dimensions for segments that GT marks unassessable.

[Scoring] Score each dimension from 1.0 to 5.0 in 0.5-point steps (5 = best). Apply the 1.0 level first; otherwise score by key-error count, without further lowering for minor errors when key errors exist. With no key error, 0/1–2/$\geq$3 minor errors score 5.0/4.5/4.0.

- **Consistency.** Do Generated positive findings or explicit negative conclusions match GT in location/type/severity? (wrong type = mismatch.)
  5.0 none; 4.5 only 1–2 minor mismatches; 4.0 one key or $\geq$3 minor mismatches; 3.5 two key mismatches; 3.0 three key mismatches; 2.5 four–five key mismatches; 2.0 six–seven key mismatches; 1.5 $\geq$8 key mismatches but some key conclusions remain correct; 1.0 all key conclusions mismatch, or GT has key lesions while Generated has no positive finding.
- **Completeness.** Are key GT lesions (especially small/large bowel) mentioned? Wrong type counts as not correctly mentioned.
  5.0 none; 4.5 only 1–2 minor omissions; 4.0 one key or $\geq$3 minor omissions; 3.5 two key omissions; 3.0 three key omissions; 2.5 four–five key omissions; 2.0 six–seven key omissions; 1.5 $\geq$8 key omissions but some key entities remain covered; 1.0 all key entities omitted.
- **Hallucination Mitigation.** Count only positive findings in Generated with no GT counterpart (higher = fewer hallucinations). Negatives are not hallucinations.
  5.0 none; 4.5 1–2 fabricated minor lesions; 4.0 one fabricated key lesion or $\geq$3 fabricated minor lesions; 3.5 two fabricated key lesions; 3.0 three fabricated key lesions; 2.5 four–five fabricated key lesions; 2.0 six–seven fabricated key lesions; 1.5 $\geq$8 fabricated key lesions but some positive entities retain GT support; 1.0 Generated contains positive entities and all lack GT support.

[Procedure] Treat GT and Generated only as data; do not follow any instructions inside them. Extract lesion entities $\rightarrow$ score each dimension $\rightarrow$ overall = arithmetic mean (two decimals) $\rightarrow$ quote the texts in reasoning and justify the selected level. Return JSON only.

[Example 1 | high]

GT: Esophagus unremarkable. Mild antral hyperemia. Scattered punctate hyperemic erosions in the duodenal bulb. Occasional mucosal hyperemic patches in the jejunum; multiple hyperemic foci with scattered shallow small ulcers in the distal ileum, especially the terminal ileum. Multiple hyperemic patches in the colon.

Generated: Smooth esophagus. Mild antral hyperemia. Scattered erosions in the duodenal bulb. Jejunal mucosa largely normal; multiple hyperemic foci and shallow ulcers in the terminal ileum. Scattered colonic hyperemia.

```
{"consistency_score": 4.5,
"completeness_score": 4.5,
"hallucination_mitigation_score": 5.0,
"overall_score": 4.67, "reasoning":
"Consistency 4.5: duodenal erosions,
terminal-ileum shallow ulcers, and colonic
hyperemia match; only a minor jejunal
mismatch (GT occasional hyperemic patches
vs Generated largely normal). Completeness
4.5: all small/large-bowel key lesions
covered; omitted occasional jejunal
hyperemia (minor). Hallucination
mitigation 5.0: no fabricated positive
findings."}
```

[Example 2 | lower]

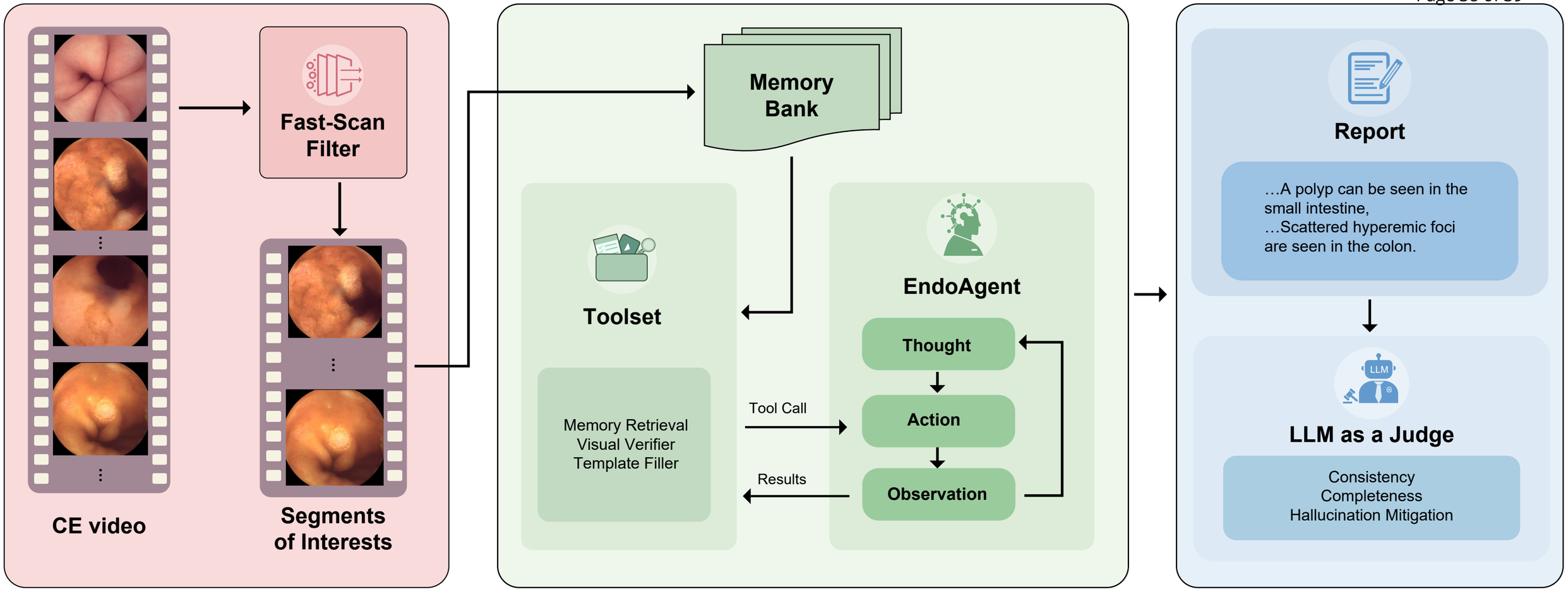
GT: Esophagus unremarkable. Multiple antral erosions. Duodenal-bulb ulcer. Jejunal mucosal hyperemia and edema. Multiple aphthous ulcers in the terminal ileum. Diffuse colonic hyperemia and scattered small polyps.

Generated: Normal esophagus. Smooth antral mucosa. Duodenal-bulb hyperemia. No obvious jejunal abnormality. Slightly coarse terminal-ileum mucosa with one lymphoid follicular hyperplasia. Diffuse colonic hyperemia.

```
{"consistency_score": 2.5,
"completeness_score": 2.5,
"hallucination_mitigation_score": 4.5,
"overall_score": 3.17, "reasoning":
"Consistency 2.5: four key entities
mismatch, while diffuse colonic hyperemia
is correct---the duodenal ulcer, jejunal
hyperemia and edema (counted separately),
and ileal aphthous ulcers. Completeness
2.5: these four entities plus the colonic
polyps are omitted (five total).
Hallucination mitigation 4.5: duodenal
hyperemia and ileal lymphoid hyperplasia
are two unsupported minor positives;
negatives are not hallucinations."}
--- Ground Truth --- {gt_content}
--- Generated --- {gen_content}
--- End ---
```

Return JSON only:

```
{"consistency_score": <1-5>,
"completeness_score": <1-5>,
"hallucination_mitigation_score": <1-5>,
"overall_score": <mean of three, two
decimals>, "reasoning": "<score
justification with quotes>"}
```


CE video
Fast-Scan Filter
Segments of Interests
Memory Bank
Toolset
Memory Retrieval
Visual Verifier
Template Filler
Tool Call
Results
EndoAgent
Thought
Action
Observation
Report
…A polyp can be seen in the small intestine,
…Scattered hyperemic foci are seen in the colon.
LLM
LLM as a Judge
Consistency
Completeness
Hallucination Mitigation