Training LLMs with Reinforcement Learning over Digital Twin Representations for Reasoning-Intensive Surgical VideoQA

Yiqing Shen[0000−0001−7866−3339], Han Zhang, and Mathias Unberath(✉)

Johns Hopkins University, Baltimore, MD, USA
{yshen92,unberath}@jhu.edu

Abstract. Surgical video question answering requires multi-step reasoning across semantic, spatial, and temporal dimensions. Existing methods architecturally compress videos into discrete token representations and couple visual perception with reasoning. This approach fragments continuous spatial-temporal relationships and has been shown to restrict multi-step reasoning capabilities. We introduce a reinforcement learning (RL) framework that trains large language models (LLMs) to decouple perception from reasoning by operating over digital twin representations constructed from surgical foundation models. Additionally, we introduce hierarchical representations across frame, temporal window, and procedure levels with probabilistic uncertainty estimates. Finally, we propose a novel reward that combines format validation with accuracy assessment through clinical plausibility evaluation and uncertainty-aware calibration for training. To demonstrate the capabilities of this approach, we introduce REAL-Colon-Reason, a colonoscopic benchmark with 2000 question-answer pairs across three complexity levels. We achieve state-of-the-art performance on REAL-Colon-Reason and two existing surgical VideoQA benchmarks REAL-Colon-VQA and EndoVis18-VQA.

Keywords: Surgical Video Question Answering · Digital Twin Representation · Reinforcement Learning · Large Language Model

1 Introduction

Surgical video question answering (VideoQA) interprets natural language questions about videos to support automated workflow analysis, trainee education, and many other applications in surgery [6, 7, 1, 5]. However, most existing formulations treat this as a direct visual matching task, where models learn to map explicit questions about visible elements to corresponding visual features [19, 13]. Recent datasets such as REAL-Colon-VQA [7] have begun incorporating temporal questions that span multiple frames, yet these still rely on single-step visual recognition rather than multi-step reasoning. In other words, existing surgical VideoQA neglects scenarios where answering requires multi-step semantic reasoning about instrument function, spatial reasoning to understand geometric/depth relations, or temporal reasoning to trace motion trajectories [7, 25, 29].

The absence of reasoning capacity prevents VideoQA from performing complex surgical tasks, such as anticipating adverse events during safety monitoring or providing guidance when surgeons face ambiguous intraoperative situations [8].

On the other hand, existing surgical VideoQA methods do not address these scenarios requiring reasoning [27]. General vision-language models (VLMs) such as Qwen-VL [2, 3] require costly domain-specific adaptation to capture surgical-specific patterns because their visual encoders are trained on natural images. Most surgery-specific approaches including SSG-VQA [33], SurgicalGPT [18] and PitVQA [9] operate on single frames, making them unable to track temporal dependencies. More recent surgical VideoQA methods like SurgViVQA [7] introduce temporal modeling, yet compress visual information into discrete token representations that fragment continuous spatial-temporal relationships, thereby restricting their ability to perform multi-step spatial and temporal inference. Furthermore, all these approaches architecturally bind visual perception to reasoning, forcing the same model to simultaneously learn fine-grained visual feature extraction and high-level reasoning during training [25, 26]. It can consequently introduce competing optimization objectives that require large-scale annotated surgical datasets [26, 37].

To overcome these limitations, we introduce a reinforcement learning (RL) framework that trains large language models (LLMs) to perform reasoning over structured intermediate representations rather than directly processing video. Formally, the LLM learns to plan the construction of intermediate representation and subsequently reasons over it via a structured rollout. We follow previous work [25, 26, 24, 23] by terming this representation as digital twin (DT) representation because it functions as a virtual replica of the surgical video [22], which is constructed by calling surgical foundation models to extract semantic entities together with their spatial and temporal relations. Because surgical procedures involve temporal hierarchies ranging from frame-level motions to workflow phases [34], our DT representation encodes observations at multiple time-scales through hierarchical structuring. Additionally, to handle perceptual ambiguity inherent in surgical videos [15], we formulate entities and relations in the DT representation as probabilistic instead of deterministic, carrying uncertainty estimates. Finally, we design rewards that enforce clinical validity and calibrate answer confidence to DT uncertainty, eliminating the need for manually annotated reasoning chains.

Our major contributions are four-fold. First, we introduce an RL framework that trains LLM to decouple perception from reasoning by planning and operating over DT representations constructed by surgical foundation models. Second, we design hierarchical and probabilistic DT representations that encode multi-scale temporal structures and uncertainty estimates to address the temporal hierarchies and perceptual ambiguities in surgical video analysis. Third, we develop a reward that enforces clinical validity and calibrates answer confidence to uncertainty. Fourth, we introduce REAL-Colon-Reason, which is a colonoscopic benchmark for reasoning-intensive surgical VideoQA. It includes questions that

require multi-step semantic, spatial, and temporal reasoning across varying complexity levels.

2 Methods

Structure of Rollout Sequence. Given a surgical video $\mathcal{V} = \{I^{(1)}, \ldots, I^{(T)}\}$ with T frames and an implicit query Q, we train LLM to generate the corresponding answer A via structured rollout sequence, as shown in Fig. 1. The rollout sequence begins with initial query reasoning, where the LLM produces chain-of-thoughts [30] R_0 between ⟨`think`⟩ and ⟨`/think`⟩ tokens to decompose the query and identify required information. Based on this analysis, LLM then outputs a DT representation construction plan G enclosed by ⟨`plan`⟩ and ⟨`/plan`⟩. The G specifies which surgical foundation models to invoke and their execution dependencies as a directed acyclic graph (DAG) [25, 26], where the LLM receives descriptions of available foundation models within its input prompt. Once the ⟨`/plan`⟩ token is detected, LLM generation pause unitl the execution of the designated foundation models. DT representation $\mathcal{D}$ is then appended between ⟨`dt`⟩ and ⟨`/dt`⟩ tokens. Afterwards, the generation resumes with another reasoning R_1 on DT representation within ⟨`think`⟩ and ⟨`/think`⟩ tokens. Finally, the LLM generates answer A enclosed by ⟨`ans`⟩ and ⟨`/ans`⟩ tokens. The complete rollout sequence follows the structure $Y = \langle\langle\text{think}\rangle R_0 \langle/\text{think}\rangle, \langle\text{plan}\rangle G \langle/\text{plan}\rangle, \langle\text{dt}\rangle \mathcal{D} \langle/\text{dt}\rangle, \langle\text{think}\rangle R_1 \langle/\text{think}\rangle, \langle\text{ans}\rangle A \langle/\text{ans}\rangle\rangle$.

Hierarchical and Probabilistic Digital Twin Representations. The construction of the DT representation is guided by the DAG plan $G = (V, E)$, where the vertices V correspond to surgical foundation models and the directed edges E specify the execution dependencies between them, following the previous work [26]. Formally, for each frame $I^{(t)}$ at time t in video $\mathcal{V}$, we first construct frame-level DT representations $D^{(t)}$ by applying the selected models according to the topological order in G. In this work, we consider the following foundation models: SurgSAM-2 [14] provides instance segmentation masks $M^{(t)} = \{m_i^{(t)}\}_{i=1}^{N^{(t)}}$ where $m_i^{(t)}$ denotes the binary mask for object i and $N^{(t)}$ indicates the number of detected instances; DepthAnything2 [32] generates dense depth maps $Z^{(t)}$ capturing spatial relationships; RASO [12] produces semantic tags $S^{(t)} = \{s_i^{(t)}\}_{i=1}^{N^{(t)}}$ identifying surgical objects; OWLv2 [16] outputs bounding boxes $B^{(t)} = \{b_i^{(t)}\}_{i=1}^{N^{(t)}}$ for efficient spatial localization. To avoid redundant storage of high-dimensional mask arrays within the text-based DT representation, we encode file paths $P^{(t)} = \{p_i^{(t)}\}_{i=1}^{N^{(t)}}$ where each path $p_i^{(t)}$ refers to the location of the corresponding mask $m_i^{(t)}$ on disk. Importantly, rather than invoking all foundation models uniformly, the LLM selects only those required by the specific query in the plan G, reducing computational overhead.

Then, we extend the DT representation with probabilistic attributes to quantify the reliability. This uncertainty information originates from the confidence

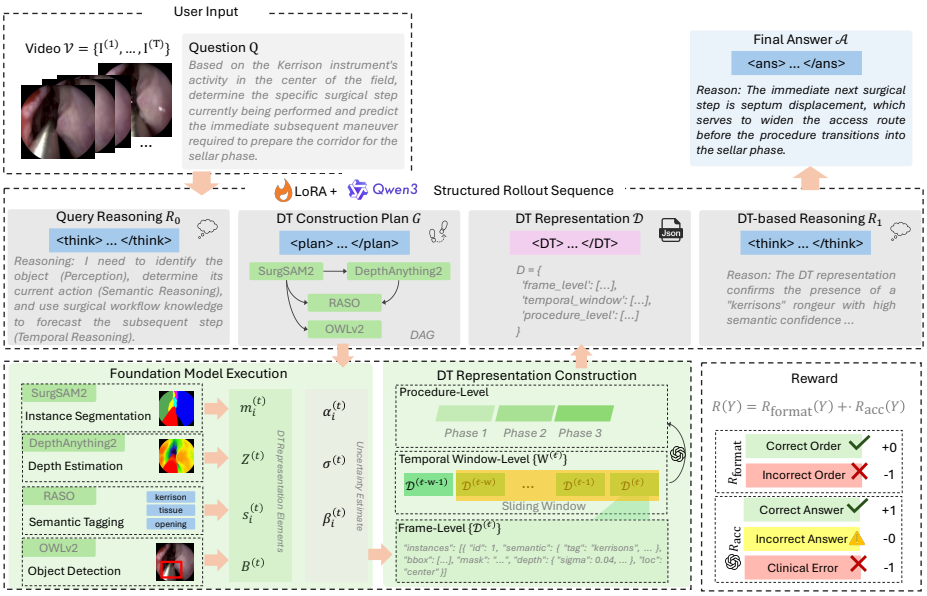

Fig. 1. Overview of the proposed method. LLM generates structured rollout sequences to plan, construct, and reason over digital twin representations of surgical videos.

estimates inherently produced by the foundation models. Specifically, SurgSAM-2 [14] produces not only binary masks $m_i^{(t)}$ but also predicted IoU scores $\alpha_i^{(t)} \in [0, 1]$ that measure segmentation certainty for each instance. Correspondingly, RASO [12] associates each semantic tag $s_i^{(t)}$ with a confidence score $\beta_i^{(t)} \in [0, 1]$ derived from the sigmoid probability output of its tag decoder prior to thresholding, reflecting the uncertainty for rare instruments or ambiguous anatomical structures. When DepthAnything2 [32] is invoked, we also compute the standard deviation to quantify spatial position uncertainty across the masked region.

Finally, surgical procedures can unfold across multiple temporal scales, such as frame-level instrument motions occurring within seconds, tissue interactions span multiple seconds, while workflow phases extend over minutes. To accommodate this multi-scale temporal structure, we organize the complete DT representation $\mathcal{D}$ hierarchically across three levels. At the frame level, each $D^{(t)}$ captures instantaneous observations at time t. At the temporal window level, we maintain a sliding context $\mathcal{W}^{(t)} = \{D^{(k)} \mid t - w \leq k \leq t\}$ with window size w that aggregates recent frames to preserve short-term dynamics, including instrument trajectories, tissue deformations, and transient occlusions. Within this window, object correspondence across frames is established through tracking identifiers $\text{id}_i^{(t)}$ assigned by SurgSAM-2's temporal consistency mechanism [14], which links the same entity across consecutive frames based on mask overlap and visual feature similarity. At the procedure level, we maintain aggregated statistics that summarize long-term context beyond the sliding window by an external LLM

(*e.g.*, GPT), compressing extended temporal sequences into workflow-relevant summaries such as phase transitions and cumulative instrument usage patterns.

Reward Design. To train the LLM with RL, we design a rule-based reward that evaluates both the correctness of rollout structure and the answers. The total reward $R(Y)$ combines a format reward $R_{\text{format}}(Y)$ and an accuracy reward $R_{\text{acc}}(Y)$, represented as $R(Y) = R_{\text{format}}(Y) + \cdot R_{\text{acc}}(Y)$. The format reward verifies that the rollout sequence contains all required token pairs in the correct order, which assigns 0 when all requirements are satisfied and -1 otherwise. The accuracy reward evaluates whether the final answer A within $\langle \texttt{ans} \rangle$ tokens matches the ground truth reference A^*. Since surgical answers can be expressed in different languages, we employ an LLM-as-a-judge [36] approach where a separate evaluation LLM (such as GPT) determines the semantic equivalence between A and A^*. Beyond semantic matching, we instruct the judge LLM to identify clinically implausible claims that violate surgical knowledge, such as contradictory instrument functions, anatomically impossible spatial relationships, or medically unsafe procedural sequences. Therefore, the accuracy reward is computed as $R_{\text{acc}}(Y) = \mathbb{I}[\text{match}(A, A^*)] - \mathbb{I}[\text{implausible}(A)]$, where the first indicator returns $+1$ when the answers match and 0 otherwise, while the second indicator imposes a penalty of -1 when implausible claims are detected. This formulation results in final values of $+1$ for correct answers, 0 for incorrect but plausible answers, and -1 for answers containing clinical errors.

To calibrate the confidence of the answer with the uncertainty encoded in the DT representation, we modulate the accuracy reward by aggregating the uncertainty from queries relevant instances. Let $\mathcal{R} \subset \{1, \ldots, N^{(t)}\}$ denote the set of instance indices to which the LLM refers when generating the answer A. For each instance $i \in \mathcal{R}$, we combine all uncertainty estimates if available: segmentation confidence $\alpha_i^{(t)}$ from SurgSAM-2 [14], semantic recognition confidence $\beta_i^{(t)}$ from RASO [12], and normalized depth uncertainty $\hat{\sigma}_i^{(t)} = 1 - \sigma_i^{(t)}/\sigma_{\max}$ where higher values indicate greater reliability. The overall confidence factor is therefore computed as the average reliability across all relevant instances and uncertainty types, denoted as

$$\gamma = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \left(w_\alpha \cdot \alpha_i^{(t)} + w_\beta \cdot \beta_i^{(t)} + w_\sigma \cdot \hat{\sigma}_i^{(t)} \right), \tag{1}$$

where w_α, w_β, and w_σ are balancing coefficients, satisfying $w_\alpha + w_\beta + w_\sigma = 1$. The final modulated reward becomes $\tilde{R}_{\text{acc}}(Y) = \gamma \cdot R_{\text{acc}}(Y)$, which scales the magnitude of the reward according to the reliability of visual observation. Finally, the LLM is trained using Group Relative Policy Optimization (GRPO) [21] with reward function $R(Y)$.

Dataset Construction. To train and evaluate the proposed method for reasoning-intensive surgical VideoQA, we introduce REAL-Colon-Reason based on the colonoscopic REAL-Colon dataset [4]. REAL-Colon comprises 60 full-resolution

colonoscopy videos, which are annotated at the frame level for endoscope motion trajectories, surgical tool interactions, tissue visibility conditions, and lesion characteristics. We first segment these videos into three to five minute clips by sampling frames at stride ten from the original 30 FPS footage to balance extended temporal coverage with computational tractability. To construct reasoning-intensive question-answer pairs, we employ a VLM (*i.e.*, GPT-5.2) to generate candidate questions along with explicit intermediate reasoning steps required to arrive at each answer by taking both the video clip and the frame-level meta information as input. All generated pairs undergo manual refinement by clinical experts to verify reasoning chain correctness, eliminate ambiguous formulations, and ensure accuracy. Questions are organized into three complexity levels based on reasoning chain depth, where level 1 involves basic single-step inference, level 2 requires two intermediate reasoning operations, and level 3 demands chains exceeding three steps of logical deduction. The final dataset contains 2000 question-answer pairs.

3 Experiments

Implementation Details. We employ Qwen3-8B [31] as the backbone LLM and train it using GRPO with LoRA [10] of rank 8 with TRL (v0.26). The training uses AdamW optimizer with a learning rate of 2×10^{-4}, a batch size of 8, and a sampling of 4 rollouts per query. We employ GPT-5-nano to judge the semantic equivalence and clinical plausibility in reward. For DT construction, we include SurgSAM-2 [14], DepthAnything2 [32], RASO [12], and OWLv2 [16], following their official implementations. We conduct training on 8 NVIDIA NVIDIA 4090 GPUs of 24Gb memory using DeepSpeed for distributed optimization.

Dataset and Evaluation Metrics. We first evaluate our method on the proposed REAL-Colon-Reason dataset, which is partitioned at the video level with an 80%:20% split. During training, we augment our dataset with 4,450 additional question-answer pairs from REAL-Colon-VQA [7] to improve the model's generalization across diverse query formulations. For REAL-Colon-Reason, we adopt exact match (EM) as the primary metric, which evaluates whether the generated answer semantically matches the ground truth reference through GPT-5-nano. We also use SMILE [11], which combines sentence-level semantic understanding with keyword-level semantics and lexical matching, as additional metric. Moreover, we conduct evaluation on REAL-Colon-VQA [7] and EndoVis18-VQA [17, 19] datasets following previous work. For these benchmarks, we report BLEU-4 for lexical precision, ROUGE-L and METEOR for semantic coherence, and keyword accuracy for clinical term correctness.

Performance on REAL-Colon-Reason. Tab. 1 presents the performance comparison across three levels of reasoning complexity on REAL-Colon-Reason. Our method achieves the best results with an overall EM of 0.584 and a SMILE

Table 1. Performance comparison on the REAL-Colon-Reason benchmark. We report EM and SMILE scores as decimal values [0, 1] (mean ± std). Standard deviations are computed over 3 independent runs with different seeds. Best results are in **bold**; second-best results are underlined.

	Level 1		Level 2		Level 3		Overall	
Model	EM	SMILE	EM	SMILE	EM	SMILE	EM	SMILE
SurgicalGPT [18]	$0.224_{\pm 0.005}$	$0.282_{\pm 0.006}$	$0.145_{\pm 0.004}$	$0.193_{\pm 0.005}$	$0.082_{\pm 0.003}$	$0.125_{\pm 0.004}$	$0.150_{\pm 0.004}$	$0.200_{\pm 0.005}$
PitVQA [9]	$0.451_{\pm 0.012}$	$0.508_{\pm 0.010}$	$0.286_{\pm 0.009}$	$0.352_{\pm 0.011}$	$0.153_{\pm 0.006}$	$0.221_{\pm 0.008}$	$0.297_{\pm 0.009}$	$0.360_{\pm 0.010}$
Qwen2.5-VL-3B [2]	$0.302_{\pm 0.008}$	$0.356_{\pm 0.009}$	$0.224_{\pm 0.007}$	$0.281_{\pm 0.008}$	$0.148_{\pm 0.005}$	$0.205_{\pm 0.007}$	$0.225_{\pm 0.006}$	$0.281_{\pm 0.008}$
Qwen3-VL-4B [3]	$0.355_{\pm 0.010}$	$0.412_{\pm 0.012}$	$0.289_{\pm 0.009}$	$0.345_{\pm 0.011}$	$0.194_{\pm 0.008}$	$0.258_{\pm 0.009}$	$0.279_{\pm 0.009}$	$0.338_{\pm 0.011}$
Qwen3-VL-8B [3]	$0.428_{\pm 0.011}$	$0.485_{\pm 0.013}$	$\underline{0.452}_{\pm 0.012}$	$\underline{0.511}_{\pm 0.014}$	$\underline{0.356}_{\pm 0.010}$	$\underline{0.423}_{\pm 0.012}$	$\underline{0.412}_{\pm 0.011}$	$\underline{0.473}_{\pm 0.013}$
MedGemma-4B [20]	$0.256_{\pm 0.007}$	$0.314_{\pm 0.008}$	$0.182_{\pm 0.006}$	$0.245_{\pm 0.007}$	$0.105_{\pm 0.004}$	$0.168_{\pm 0.006}$	$0.181_{\pm 0.006}$	$0.242_{\pm 0.007}$
InternVL3-1B [38]	$\underline{0.482}_{\pm 0.012}$	$\underline{0.535}_{\pm 0.014}$	$0.354_{\pm 0.010}$	$0.412_{\pm 0.011}$	$0.221_{\pm 0.008}$	$0.295_{\pm 0.010}$	$0.352_{\pm 0.010}$	$0.414_{\pm 0.012}$
Surgical-LVLM [28]	$0.385_{\pm 0.010}$	$0.441_{\pm 0.011}$	$0.312_{\pm 0.009}$	$0.375_{\pm 0.010}$	$0.208_{\pm 0.007}$	$0.272_{\pm 0.009}$	$0.302_{\pm 0.009}$	$0.363_{\pm 0.010}$
VideoLLaMA3-2B [35]	$0.185_{\pm 0.004}$	$0.241_{\pm 0.006}$	$0.123_{\pm 0.003}$	$0.178_{\pm 0.005}$	$0.064_{\pm 0.002}$	$0.112_{\pm 0.004}$	$0.124_{\pm 0.003}$	$0.177_{\pm 0.005}$
SurgViVQA [7]	$0.405_{\pm 0.009}$	$0.462_{\pm 0.011}$	$0.264_{\pm 0.008}$	$0.328_{\pm 0.010}$	$0.142_{\pm 0.005}$	$0.205_{\pm 0.007}$	$0.270_{\pm 0.007}$	$0.332_{\pm 0.009}$
Ours	$\mathbf{0.653}_{\pm 0.013}$	$\mathbf{0.708}_{\pm 0.015}$	$\mathbf{0.584}_{\pm 0.012}$	$\mathbf{0.642}_{\pm 0.014}$	$\mathbf{0.515}_{\pm 0.011}$	$\mathbf{0.589}_{\pm 0.013}$	$\mathbf{0.584}_{\pm 0.012}$	$\mathbf{0.646}_{\pm 0.014}$

score of 0.646, outperforming the second-best method Qwen3-VL-8B [3] by 17.2% and 17.3% respectively. Specifically, our method shows improvements of 15.9% at Level 1, 13.2% at Level 2, and 15.9% at Level 3 in EM over the strongest baseline, namely Qwen3-VL-8B [3]. This demonstrates that DT representations effectively preserve the fine-grained spatial-temporal information required for multi-step reasoning. Image-based methods such as SurgicalGPT [18] and PitVQA [9] show performance degradation across all levels, illustrating that single-frame approaches fail to capture temporal dependencies needed for temporal reasoning. SurgViVQA [7] achieves an overall EM of 0.270 compared to our 0.584, despite incorporating temporal modeling through masked video encoding. This supports our hypothesis that token-based compression fragments continuous spatial-temporal relationships and limits multi-step reasoning capability. Zero-shot general-purpose VLMs including Qwen3-VL-8B [3] and InternVL3-1B [38] demonstrate competitive performance at level 1 but show steeper degradation level 2 and 3. This indicates that without an explicit reasoning structure, even powerful general-purpose VLMs cannot fully handle reasoning-intensive surgical scenarios.

Performance on Existing VideoQA Benchmarks. Tab. 2 presents performance on REAL-Colon-VQA and EndoVis18-VQA to assess generalization to existing surgical VideoQA benchmarks. Our method achieves the best results with BLEU-4 scores of 75.42% and 87.20% for in-template evaluations on the two datasets respectively. On keyword accuracy, our method outperforms SurgViVQA [7] by 7.85% and 11.23%, demonstrating that explicit reasoning over DT representations improves factual grounding. The improvements extend to out-of-template evaluations, where our method achieves keyword accuracy of 57.12% and 53.40%, outperforming SurgViVQA [7]. Zero-shot VLMs show stronger out-of-template performance than supervised methods, yet our method surpasses them by maintaining both linguistic fluency and clinical accuracy. These results

Table 2. Performance comparison on REAL-Colon-VQA and EndoVis18-VQA. We report BLEU-4 (B-4), ROUGE-L (R-L), METEOR (MET), and Keyword Accuracy (K-ACC) in percentages (%). Best results are in **bold**; second-best results are underlined.

	REAL-Colon-VQA [7]								**EndoVis18-VQA** [19]							
	In-Template				Out-of-Template				In-Template				Out-of-Template [17]			
Methods	B-4	R-L	MET	K-ACC	B-4	R-L	MET	K-ACC	B-4	R-L	MET	K-ACC	B-4	R-L	MET	K-ACC
SurgicalGPT [18]	14.93	47.85	52.36	33.33	12.35	42.87	50.49	46.67	29.55	58.60	57.99	4.00	2.52	43.97	44.99	11.85
PitVQA [9]	64.55	78.48	79.99	54.13	23.63	50.03	53.22	42.93	81.73	86.18	83.28	40.35	17.57	46.87	45.49	9.73
Qwen2.5-VL-3B [2]	4.53	45.70	46.83	49.33	0.50	38.80	40.77	50.27	19.04	57.60	68.60	51.22	11.92	52.77	63.72	46.90
Qwen3-VL-4B [3]	6.80	48.20	51.50	54.00	2.10	40.50	43.80	52.40	22.50	59.30	70.20	52.80	13.20	53.50	64.80	48.10
Qwen3-VL-8B [3]	8.90	50.10	54.70	58.20	3.50	42.80	46.50	54.10	26.30	61.40	72.50	54.60	15.80	55.20	66.90	50.30
MedGemma-4B [20]	18.09	36.92	31.86	49.47	5.31	30.96	27.80	34.13	8.23	44.32	64.19	37.90	7.74	40.74	61.59	37.86
InternVL3-1B [38]	7.13	43.90	54.31	68.67	3.87	32.10	44.34	53.73	5.62	51.50	55.33	32.21	0.57	38.73	38.73	21.99
Surgical-LVLM [28]	7.90	49.80	53.80	57.20	3.20	42.00	45.80	53.80	24.60	60.80	72.00	53.50	15.00	54.80	66.20	49.80
VideoLLaMA3-2B [35]	2.74	33.01	42.22	17.60	1.78	26.17	35.15	21.47	3.28	52.01	54.66	40.90	0.57	38.73	39.51	35.89
SurgViVQA [7]	71.98	82.85	84.11	63.20	31.19	53.62	54.89	47.73	84.94	89.65	86.08	48.27	24.84	50.86	50.13	9.23
Ours	**75.42**	**85.10**	**86.35**	**71.05**	**35.80**	**57.45**	**58.90**	**57.12**	**87.20**	**91.50**	**88.40**	**59.50**	**28.15**	**55.90**	**68.45**	**53.40**

confirm that our framework generalizes to standard VideoQA tasks while maintaining the advantages of explicit multi-step reasoning.

Table 3. Ablation study on REAL-Colon-Reason benchmark.

Architecture			**Reward**			**Level 1**		**Level 2**		**Level 3**		**Overall**	
DT	Hier.	Prob.	Format	Acc.	Uncert.	EM	SMILE	EM	SMILE	EM	SMILE	EM	SMILE
×	×	×	✓	✓	×	0.312	0.368	0.245	0.305	0.168	0.232	0.242	0.302
✓	×	×	✓	✓	×	0.468	0.524	0.382	0.445	0.298	0.365	0.383	0.445
✓	✓	×	✓	✓	×	0.521	0.578	0.445	0.508	0.356	0.424	0.441	0.503
✓	✓	✓	✓	✓	×	0.547	0.604	0.478	0.542	0.389	0.458	0.471	0.535
✓	✓	✓	×	✓	✓	0.423	0.482	0.348	0.412	0.267	0.335	0.346	0.410
✓	✓	✓	✓	×	✓	0.489	0.548	0.412	0.478	0.321	0.391	0.407	0.472
✓	✓	✓	✓	✓	×	0.547	0.604	0.478	0.542	0.389	0.458	0.471	0.535
✓	✓	✓	×	×	×	0.401	0.458	0.325	0.389	0.248	0.318	0.325	0.388
✓	✓	✓	✓	✓	✓	**0.653**	**0.708**	**0.584**	**0.642**	**0.515**	**0.589**	**0.584**	**0.646**

Ablation Study. Tab. 3 evaluates the contribution of each component in our framework on REAL-Colon-Reason. Introducing DT representation improves EM from baseline 0.242 to 0.383, while hierarchical temporal structure and probabilistic attributes further increase performance to 0.441 and 0.471 respectively. Among reward components, format reward proves most important with its removal causing degradation to 0.346 EM, indicating that structured rollout supervision is necessary for valid DT construction plans. The full RL-trained model achieves 0.584 EM compared to supervised fine-tuning baseline at 0.325 EM, confirming that reinforcement learning with our designed reward effectively trains reasoning capabilities.

4 Conclusion

In this paper, we present a RL framework that trains LLM to perform reasoning-intensive surgical video question answering through DT representations constructed from foundation models. By decoupling visual perception from reasoning and organizing representations hierarchically with probabilistic uncertainty estimates, our framework preserves spatial-temporal information that token-based methods fragment. Evaluation on three benchmarks demonstrates improvements over existing methods. Future work can explore real-time deployment through distilled foundation models for DT construction and LLMs. Another direction is the incorporation of multi-modal information such as audio cues and physiological signals to enrich DT representations.

Acknowledgments. Y. Shen was supported in part by the JHU Amazon Initiative for Artificial Intelligence (AI2AI) fellowship program.

Disclosure of Interests. The authors have no competing interests in the paper.

References

1. Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. *arXiv preprint arXiv:2305.11692*, 2023.
2. Shuai Bai et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
3. Shuai Bai et al. Qwen3-vl technical report, 2025.
4. Carlo Biffi et al. Real-colon: A dataset for developing real-world ai applications in colonoscopy. *Scientific Data*, 11(1):539, 2024.
5. Kexin Chen et al. 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. IEEE, 2024.
6. Di Ding, Tianliang Yao, Rong Luo, and Xusen Sun. Visual question answering in robotic surgery: A comprehensive review. *IEEE Access*, 2025.
7. Mauro Orazio Drago et al. Surgvivqa: Temporally-grounded video question answering for surgical scene understanding. *arXiv preprint arXiv:2511.03325*, 2025.
8. Michael B Eppler et al. Automated capture of intraoperative adverse events using artificial intelligence: a systematic review and meta-analysis. *Journal of Clinical Medicine*, 12(4):1687, 2023.
9. Runlong He et al. 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.
10. Edward J Hu et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
11. Shrikant Kendre et al. Smile: A composite lexical-semantic metric for question-answering evaluation. *arXiv preprint arXiv:2511.17432*, 2025.
12. Jiajie Li et al. Recognize any surgical object: Unleashing the power of weakly-supervised data. *arXiv preprint arXiv:2501.15326*, 2025.

13. Lili Liang, Guanglu Sun, Jin Qiu, and Lizhong Zhang. Neural-symbolic videoqa: Learning compositional spatio-temporal reasoning for real-world video question answering. *arXiv preprint arXiv:2404.04007*, 2024.
14. Haofeng Liu et al. Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. *arXiv preprint arXiv:2408.07931*, 2024.
15. Haofeng Liu et al. Sam2s: Segment anything in surgical videos via semantic long-term tracking. *arXiv preprint arXiv:2511.16618*, 2025.
16. Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.
17. Dennis Pierantozzi, Luca Carlini, Mauro Orazio Drago, Chiara Lena, Cesare Hassan, Elena De Momi, Danail Stoyanov, Sophia Bano, and Mobarak I Hoque. When to trust the answer: Question-aligned semantic nearest neighbor entropy for safer surgical vqa. *arXiv preprint arXiv:2511.01458*, 2025.
18. Lalithkumar Seenivasan et al. 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. Springer, 2023.
19. 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. Springer, 2022.
20. Andrew Sellergren et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.
21. Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
22. Yiqing Shen, Hao Ding, Lalithkumar Seenivasan, Tianmin Shu, and Mathias Unberath. Position: Foundation models need digital twin representations. *arXiv preprint arXiv:2505.03798*, 2025.
23. Yiqing Shen, Chenxiao Fan, Chenjia Li, and Mathias Unberath. Reasoning text-to-video retrieval via digital twin video representations and large language models. *arXiv preprint arXiv:2511.12371*, 2025.
24. Yiqing Shen, Chenjia Li, and Mathias Unberath. Text-driven reasoning video editing via reinforcement learning on digital twin representations. *arXiv preprint arXiv:2511.14100*, 2025.
25. Yiqing Shen, Bohan Liu, Chenjia Li, Lalithkumar Seenivasan, and Mathias Unberath. Online reasoning video segmentation with just-in-time digital twins. *arXiv preprint arXiv:2503.21056*, 2025.
26. Yiqing Shen and Mathias Unberath. Constructing and interpreting digital twin representations for visual reasoning via reinforcement learning. *arXiv preprint arXiv:2511.12365*, 2025.
27. Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. Explore multi-step reasoning in video question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 239–247, 2018.
28. Guankun Wang et al. Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. *arXiv preprint arXiv:2405.10948*, 2024.
29. Haibo Wang et al. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024.
30. Jason Wei et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

31. An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
32. Lihe Yang, Bingyi Kang, Zilong Huang, et al. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.
33. Kun Yuan et al. Advancing surgical vqa with scene graph knowledge. *International journal of computer assisted radiology and surgery*, 19(7):1409–1417, 2024.
34. Wenxi Yue et al. Cascade multi-level transformer network for surgical workflow analysis. *IEEE transactions on medical imaging*, 42(10):2817–2831, 2023.
35. Boqiang Zhang et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.
36. Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
37. Jingqi Zhou, Sheng Wang, Jingwei Dong, Kai Liu, Lei Li, Jiahui Gao, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. Proreason: Multi-modal proactive reasoning with decoupled eyesight and wisdom. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31650–31679, 2025.
38. Jinguo Zhu et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.