SurgGoal: Rethinking Surgical Planning Evaluation via Goal-Satisfiability

Anonymous ACL submission

Abstract

Surgical planning integrates visual perception, long-horizon reasoning, and procedural knowledge, yet it remains unclear whether current evaluation protocols reliably assess vision-language models (VLMs) in safety-critical settings. Motivated by a goal-oriented view of surgical planning, we define planning correctness via phase-goal satisfiability, where plan validity is determined by expert-defined surgical rules. Based on this definition, we introduce a multicentric meta-evaluation benchmark with valid procedural variations and invalid plans containing order and content errors. Using this benchmark, we show that sequence similarity metrics systematically misjudge planning quality, penalizing valid plans while failing to identify invalid ones. We therefore adopt a rule-based goal-satisfiability metric as a high-precision meta-evaluation reference to assess Video-LLMs under progressively constrained settings, revealing failures due to perception errors and under-constrained reasoning. Structural knowledge consistently improves performance, whereas semantic guidance alone is unreliable and benefits larger models only when combined with structural constraints.

1 Introduction

Vision-language models (VLMs) have become powerful foundation models capable of reasoning over visual content through natural language (Zhang et al., 2025a). In the video domain, they have progressed beyond low-level perception toward long-horizon temporal reasoning (Grauman et al., 2022; Bai et al., 2025), enabling action recognition (Fan and Zheng, 2024), anticipation (Lin et al., 2024), and task-oriented planning (Li et al., 2025a; Zhao et al., 2023; Li et al., 2025b), and supporting online assistance in everyday scenarios. As these capabilities mature, VLMs are increasingly considered for deployment in high-stakes settings, where errors carry significant consequences. One such domain is surgery, where intelligent intraoperative surgical planning (Boels et al., 2025b,a; Xu et al., 2025b) predicts future actions conditioned on the current procedural state and the goal, which is highly desirable but subject to strict requirements on safety and reliability issues.

Despite recent progress in surgical planning models, it remains unclear whether existing evaluation protocols reliably measure clinically valid planning behavior. Most prior work evaluates planning outputs by comparing predicted step or phase sequences to a single reference trajectory using surface sequence similarity metrics (Ding et al., 2025; Xu et al., 2025a) such as edit distance or relative order accuracy. These metrics equate planning correctness with resemblance to one observed execution, potentially rewarding unsafe ordering errors while penalizing clinically plausible alternatives. This mismatch raises concerns about whether current evaluations meaningfully reflect planning capability in safety-critical surgical settings.

To address this, motivated by a goal-oriented, multi-step view of surgical planning, we define planning correctness via phase-goal satisfiability, with plan validity determined by expert-defined surgical rules encoding hierarchical phase–step relations and procedural constraints. Building on this definition, we introduce a multicentric meta-evaluation benchmark grounded in MultiBypass140 (Lavanchy et al., 2024), comprising clinically valid procedural variations and systematically constructed invalid plans with order and content errors. This benchmark enables meta-evaluation of planning metrics by assessing whether their validity judgments align with goal satisfiability rather than surface-level resemblance.

Using this benchmark, we show that widely used sequence similarity metrics are fundamentally misaligned with goal-satisfiability, and that LLM-as-a-judge baselines often capture semantic omissions but fail to enforce strict procedural dependencies.

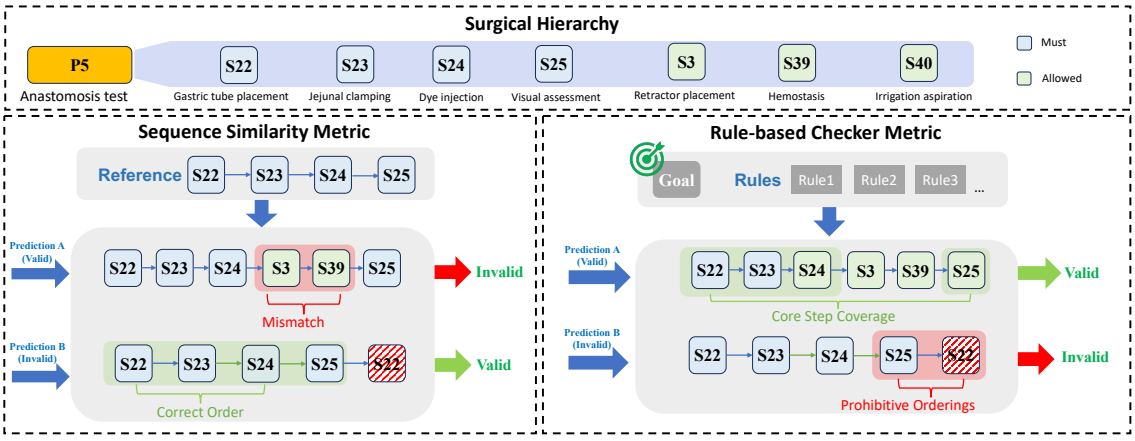

Figure 1: **Comparison of sequence similarity metrics and rule-based checker metric. Top**: Surgical procedures follow a hierarchical structure: each phase (e.g., P5) contains mandatory core steps (blue) and permissible generic steps (green). **Bottom Left**: Sequence similarity metrics compare predictions to a fixed reference, causing false negatives for valid clinical variations (Prediction A) and false positives for prohibited orderings (Prediction B). **Bottom Right**: The rule-based checker correctly distinguishes valid from invalid plans using surgical rules.(Section 3.1)

Lacking a reliable alternative for planning correctness, we then adopt a rule-based checker metric as a high-precision meta-evaluation reference to assess Video-LLMs under progressively informative planning settings, from end-to-end video planning to explicit state and injected knowledge. This reveals failure modes driven by visual misrecognition and under-constrained reasoning, and shows that explicit procedural structure yields consistent gains, whereas semantic knowledge benefits depend strongly on model capacity.

These insights highlight the need for more robust and clinically grounded evaluation protocols to support the development of reliable surgical planning systems in high-stakes settings. Contributions:

- We define planning correctness via goal-satisfiability under expert-defined surgical rules.
- We introduce a meta-evaluation benchmark with valid procedural variations and invalid plans.
- We benchmark traditional metrics and LLM-based judges under a meta-evaluation framework, revealing systematic evaluation misalignment.
- We introduce a rule-based checker metric and use it to evaluate Video-LLMs across progressively constrained planning tasks.

2 Rethinking Surgical Planning: Task Formulation and Prior Work

2.1 Previous Work: Flat Prediction under Strong Structural Constraints

Task formulation mismatch (flat vs. hierarchical). Prior work in surgical planning predominantly formulates the problem as flat prediction, including next-action classification (Xu et al., 2025b,a), phase transition forecasting (Boels et al., 2025a; Chen et al., 2025). These formulations assume a single temporal granularity and reduce planning to local transitions between adjacent actions or phases. However, real surgical procedures are inherently hierarchical (Biagini et al., 2025; Lavanchy et al., 2024; Lalys and Jannin, 2014). Phases correspond to strategic sub-goals, steps instantiate tactical plans, and actions serve as execution primitives. Planning cognition operates not at isolated action transitions, but in organizing multi-step structures toward phase-level objectives. Flat formulations obscure this hierarchy and fail to capture operative procedural logic.

Objective mismatch (transition-based vs. goal-oriented). Existing approaches largely focus on predicting what comes next, treating planning as a sequence of independent transition decisions. In contrast, surgical decision-making is fundamentally goal-oriented. Surgeons do not decide on isolated actions; instead, they engage in top-down reasoning. A surgeon first establishes a strategic phase level objective (e.g., jejunal separation) , which then constrains and motivates a sequence of coordinated steps (e.g., mesenteric opening, jejunal transection) planned over a long temporal horizon. As a result, transition-centric models lack the semantic grounding needed to explain why a step is performed and whether a sequence meaningfully contributes to procedural completion. For instance, a dissection action may be appropriate during the

dissection of Calot's triangle, but would be questionable if it occurs during the abdominal closure phase (Lavanchy et al., 2025).

Path multiplicity ignored (single trajectory vs. multiple valid plans). Clinical reality admits multiple valid step-level paths to achieve the same phase-level goal, driven by surgeon preference, patient anatomy, intraoperative findings, and institutional style. In practice, even the same surgical team may legitimately reorder phases; for instance, in Roux-en-Y gastric bypass (Lavanchy et al., 2024), performing omentum division before gastric pouch creation can improve operative exposure. Such variations are clinically equivalent as they preserve essential dependencies (e.g., performing anastomosis before its integrity test). Prior work, however, typically treats the observed sequence as a unique ground truth, implicitly equating deviations with errors. This one-to-one assumption contradicts surgical practice, where procedural correctness is defined by dependency preservation rather than exact sequence identity. Modeling surgical planning as deterministic single-path prediction therefore systematically penalizes valid variations and conflates alternative plans with incorrect ones, despite strong structural constraints.

2.2 Evaluation Mismatch: From Sequence Similarity to Goal Satisfiability

Given that surgical planning is goal-oriented, and admits multiple valid paths, a key issue lies in how it is evaluated. Prior work predominantly relies on sequence match metrics and their variants, including edit distance, Jaccard similarity, and relative order accuracy. Despite different formulations, these metrics share a common design: they measure surface-level similarity between a predicted sequence and a single reference trajectory.

This evaluation paradigm implicitly assumes a unique correct execution path, equates deviation with error, and treats order similarity as a proxy for procedural validity. As a result, planning is assessed by agreement with an observed rollout rather than by whether a sequence can meaningfully accomplish the intended surgical objective.

These assumptions contradict surgical reality. Because multiple step-level paths may validly satisfy the same phase-level goal, sequence similarity metrics produce false negatives by penalizing legitimate alternatives, and false positives by rewarding superficially similar but dependency-violating plans (Figure 1). Therefore, procedural correctness should be defined not by sequence identity, but by whether a plan plausibly satisfies the surgical goal under essential procedural dependencies. This task-evaluation mismatch motivates the need for a goal-satisfiability-based perspective and leads to our meta-evaluation framework.

3 Meta-Evaluation: A Goal-Satisfiability Benchmark for Surgical Planning

This section introduces a meta-evaluation benchmark to test whether existing planning metrics correctly assess goal-satisfiability in surgical planning.

3.1 Formalizing Goal-Satisfiability with Surgical Rules

Our benchmark is grounded in the MultiBypass140 dataset (Lavanchy et al., 2024), a multicentric collection of 140 laparoscopic Roux-en-Y gastric bypass procedures with untrimmed videos and surgeon-annotated hierarchical labels spanning 11 phases and 45 fine-grained steps.

We formalize goal-satisfiability via a set of expert-defined surgical rules that determine whether a step sequence can plausibly complete a target phase. The rules encode clinically essential constraints derived from phase semantics and procedural dependencies and are intended as a high-precision reference for meta-evaluation. Rather than enumerating all valid surgical variations, the rules conservatively distinguish sequences that are *definitively invalid* from those that are *plausibly goal-satisfiable* within a well-defined scope.

Rule Components. An example of phase P5 (anastomosis test).

- **Required Steps Set** defines mandatory steps to achieve the phase goal: S22 (gastric tube placement), S23 (jejunal clamping), S24 (dye injection), and S25 (visual assessment).
- **Allowed Steps Set** enumerates additional steps that may plausibly occur without violating the phase objective, such as S3 (retractor placement), S39 (hemostasis), or S40 (irrigation aspiration).
- **Procedural Dependencies** enforce critical ordering constraints among core steps, such as, if S24 occurs, the first occurrence of S23 must strictly precede it, and S25 must occur after S24.
- **Prohibitive Orderings** further restricts when ancillary steps may appear. Actions such as S39 (hemostasis) or S40 (irrigation aspiration), if present, are permitted only after clamping has begun. Crucially, once the last S25 is completed,

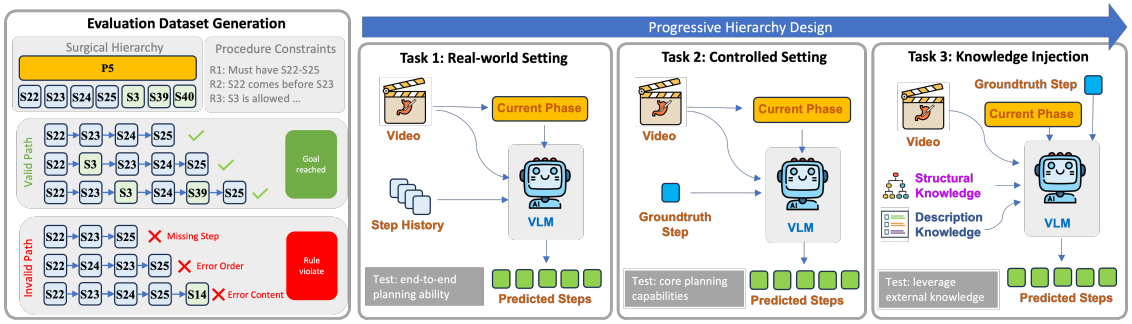

Figure 2: **Meta-evaluation and Evaluation Pipelines for Surgical Planning.** Left: Rule-based benchmark defining goal-satisfiability via hierarchical phase-step relations and procedural constraints (dependencies and prohibitive orderings), separating valid and invalid step sequences. Right: Progressive evaluation of Video-LLMs, from end-to-end planning to planning with ground-truth steps and injected knowledge.

core test steps are prohibited from reappearing, explicitly marking completion of the phase.

We construct a comprehensive rule specification that encodes all constraints described above, including 50 expert-defined procedural dependencies and prohibitions, hierarchical phase-step relations, and phase-specific allowed and required steps from the MultiBypass140 annotation protocol. A step sequence is labeled as valid if it satisfies all rules defined for the target phase; any violation of rules renders it invalid. These rules provide a high-precision reference labeling of goal-satisfiability within the scope of the benchmark, serving as the basis for evaluating whether different metrics capture the intended notion of procedural validity. They are not intended to represent an exhaustive clinical ground truth, but rather a conservative reference that reliably identifies definitively invalid plans.

3.2 Meta-Evaluation Dataset Construction

Building on these surgical rules, we construct a meta-evaluation dataset from MultiBypass140 designed to probe metric behavior under a multi-path formulation. We curate a test set of step-phase pairs, categorized by expert surgeons (Figure 2):

- **Correct Sequences** ($N = 191$): Clinically valid paths that satisfy all rules and achieve the phase goal, allowing legitimate variations in step order and ancillary maneuvers (e.g., hemostasis).
- **Incorrect Sequences** ($N = 199$): Paths that violate one or more rules and fail to achieve the phase goal, further sub-classified into *Order Errors* (OE; violations of procedural dependencies), *Content Errors* (CE; missing core steps or inclusion of any step outside the allowed set of the corresponding phase), and *Both* (BE).

3.3 Meta-Evaluation Protocol

We define a unified protocol to evaluate planning metrics under a goal-satisfiability formulation.

Input. The input to a metric is a candidate step sequence and its associated target phase. For metrics that require a reference sequence, we construct a canonical reference using the phase-specific core steps defined by the surgical rules, ordered according to standard procedural dependencies.

Output. Each metric produces a binary judgment indicating whether the candidate sequence is considered valid (goal-satisfiable) or invalid for completing the target phase. For continuous-valued metrics, scores are thresholded at 0.7 to obtain a binary decision.

Comparison and Reporting. This protocol evaluates whether a metric's decision boundary aligns with the rule-based definition of goal-satisfiability, rather than surface similarity to a single reference sequence. Performance is reported as binary classification accuracy, optionally stratified by sequence category (Valid, OE, CE, BE) to analyze metric sensitivity to different failure modes.

4 Benchmarking Planning Metrics under Goal-Satisfiability Meta-Evaluation

Under the proposed meta-evaluation protocol, planning metrics are evaluated by their ability to classify step sequences as goal-satisfiable or invalid.

4.1 Metrics under Comparison.

Sequence Similarity Metrics We evaluate a representative set of surface-level metrics, including Normalized Edit Distance (NED) (Marzal and Vidal, 2002), Jaccard Index on Sequences (JIS) (Broder, 1997), and Relative Order Accuracy (ROA) (Kendall, 1938), using their standard

	Traditional Metrics			⋆	LLM-based Judges			
Subset	**NED**	**JIS**	**ROA**	**Rule**	**Gemini3**	**GPT 5.2**	**Claude 4.5**	**HuluMed**
Valid	18.8	40.3	93.2	100.0	99.5	97.4	61.8	99.0
OE	87.3	46.5	11.3	100.0	11.3	18.3	80.3	8.5
CE	86.8	85.3	17.6	100.0	85.3	97.1	97.1	58.8
BE	96.7	85.0	20.0	100.0	98.3	100.0	100.0	75.0
Invalid	89.9	71.4	17.1	100.0	63.8	69.8	92.0	45.7

Table 1: Accuracy (%) comparison of traditional similarity metrics and LLM-based judges across valid and erroneous subsets. NED: Normalized Edit Distance. JIS: Jaccard Index on Sequences. ROA: Relative Order Accuracy. OE: order error. CE: content error. BE: both error.

formulations. These metrics, together with their commonly used variants, have been widely adopted to measure sequence similarity and relative order agreement in prior surgical workflow analysis.

Rule-based Checker Metric. We include an expert-defined rule-based checker derived from the surgical rules in Section 3.1, which labels sequences as valid or invalid; since the meta-evaluation dataset strictly follows these rules, the checker serves as an upper bound on performance.

LLM Judge. We evaluate several LLM-based judges that assess plan plausibility using injected phase-step relationships and descriptions, rather than the explicit rules in Section 3.1. These judges output both a binary validity decision and a textual explanation, providing a more flexible and potentially scalable alternative to the rule-based checker.

4.2 Results and Analysis

Surface-level similarity metrics exhibit a systematic bias. NED and JIS achieve high accuracy on invalid sequences but perform poorly on valid plans (Table 1), misclassifying the majority of clinically correct variations. This confirms a similarity trap: deviations from a single reference trajectory are penalized regardless of whether the phase goal is satisfied. As a result, these metrics conflate procedural diversity and flexibility with error.

ROA is permissive but unsafe. ROA achieves high accuracy on valid sequences but fails catastrophically on order errors (Table 1). By measuring only relative pairwise order, it overlooks repetitions and critical misplacements that render a procedure infeasible, thereby rewarding sequences that violate essential temporal and causal constraints.

Rule-based evaluation provides a high-precision reference. The expert-defined rule checker metric achieves perfect accuracy across all subsets within its defined scope, as it is directly derived from the same rules used to construct the dataset; we therefore treat it as a high-precision upper bound. However, this approach requires substantial expert effort and is highly task-specific, limiting its practicality and scalability to broader procedures or settings.

LLM-based Judges: Semantics over Structure. Table 1 indicates that LLM-based evaluators perform well on content errors, indicating strong semantic understanding of phase goals and missing steps. They utilize internal medical knowledge to recognize that *Gastric Pouch Creation* cannot be completed if the *stapling* step is missing. However, most models struggle with order errors, frequently approving sequences that violate critical procedural dependencies. For phase P5, S24 is placed before S23, which would allow the dye to escape downstream before the test segment is sealed, thereby invalidating the anastomotic leak test; nevertheless, GPT judged the sequence as correct, reasoning that all required instruments for the test were present. This failure highlights a systematic tendency of LLMs to prioritize semantic completeness over strict procedural ordering, leading to errors when correct execution depends on temporal or causal constraints rather than the mere presence of actions. Moreover, different models exhibit distinct biases: HuluMed tends to over-accept plausibly complete plans, while Claude over-reject them; yet none consistently enforce dependency-level correctness.

Overall, existing planning metrics fail to reliably assess goal-satisfiability in realistic multi-path settings: rule-based evaluation is precise but unscalable, while LLM-based evaluators are flexible yet unable to infer strict procedural constraints.

5 Evaluation of Video-LLMs for Goal-Oriented Surgical Planning

5.1 Experimental Setup

Dataset. Experiments are conducted on MultiBypass140. To focus on logical planning rather than

temporal boundary detection, we segment each video into 5,032 discrete step-level clips based on expert annotations. Each clip preserves its full temporal context (up to ten minutes), ensuring that the model has access to the complete visual evidence of the ongoing maneuver.

Models. We evaluate VideoLLaMA3-7B (Zhang et al., 2025b), LLaVA-NeXT-Video-7B (Li et al., 2024), Qwen2.5-VL (7B/32B) (Bai et al., 2025), and medical models Hulu-Med (7B/32B) (Jiang et al., 2025) and Lingshu (7B/32B) (Xu et al., 2025c). All models are evaluated zero-shot (temperature = 0, max output = 2048 tokens).

5.2 Progressive Task Formulation

To disentangle visual perception from procedural reasoning, we define progressively constrained planning tasks. (Figure 2)

Real-World Setting: End-to-End Planning. *Task 1*: Models are given raw surgical video clips, the current phase label, the history of completed steps, and a set of candidate step labels. They must infer the current procedural state from visual evidence and generate a plausible future step sequence. This setting mimics a real-world intraoperative assistant that must simultaneously ground its understanding in visual evidence (What is happening now?) and extrapolate future actions.

Controlled Setting: Planning with Explicit State. *Task 2*: To isolate planning from perception, models are additionally provided with the current step identity, while retaining video input. This removes ambiguity about the procedural state and allows us to directly assess the model's ability to plan future steps from a fixed starting point.

Knowledge Injection for Surgical Planning. *Task 3*: We further investigate how external medical expertise modulates planning quality by injecting three forms of knowledge into the Task 2 setup:

3.1 Structural knowledge: phase-step hierarchy.
3.2 Semantic knowledge: expert-written natural language descriptions of the phases and steps.
3.3 Combined knowledge: both knowledge.

5.3 Planning Output

Each model produces a unified structured output following a fixed JSON schema:

- Remaining steps to complete the current phase
- Next phase (as a phase name)
- Reasonable step sequence for the next phase
- Explanation (brief justification)

This output supports two planning tasks under a unified evaluation protocol:

current-phase completion planning, evaluated on the combined plan of completed steps, the current step, and predicted remaining steps;

next-phase planning, evaluated on the generated step sequence for the predicted next phase.

5.4 Evaluation

We evaluate VLM surgical planning using *goal-satisfiability accuracy*. In the absence of a reliable alternative metric (Section 4.2), we use an expert-defined rule-based checker metric (Section 4.1) to determine whether each generated step sequence constitutes a plausible path for completing the target phase. Accuracy is computed as the fraction of sequences judged valid.

For Task 1, we additionally report *current step recognition*, defined as exact matching between the predicted and ground-truth current step, to separate step recognition from planning quality. We do not evaluate exact next-phase prediction. Surgical planning admits multiple valid phase transitions and does not assume a single canonical next phase.

5.5 Results and Analysis

We analyze the results of Section 5.2 by progressively isolating the roles of visual perception, planning logic, and medical knowledge in goal-oriented surgical planning.

5.5.1 The Perception-Reasoning Gap in End-to-End Planning

Task 1 reflects the real-world setting where models must infer procedural state directly from video. As shown in Table 2, even the strongest model in our comparison (Lingshu-32B) achieves only 39.4% step recognition accuracy, resulting in substantial downstream planning errors. A clinician-led analysis reveals two dominant perception failure modes.

Confusing exploration with procedural steps. Models often fail to distinguish long, repetitive exploratory video segments (e.g., tissue retraction to locate target organs or vessels) from well-defined surgical steps. For example, prolonged tissue exploration before definitive vessel exposure is often misclassified as a subsequent surgical step. As a result, models may prematurely conclude that a phase objective has been met, producing implausible plans that omit essential preparatory actions.

Failing to recognize repeated steps. Models also struggle when the same step appears multiple

Task	Metric	Hulu-7B	Hulu-32B	Qwen-7B	Qwen-32B	DAMO	Lingshu-7B	Lingshu-32B	LLaVA
Real-world	StepAcc	23.5%	34.0%	21.7%	31.4%	14.4%	26.1%	**39.4%**	2.8%
	Current	10.3%	23.8%	2.9%	20.9%	8.8%	7.4%	**31.1%**	1.3%
	Next	2.9%	45.6%	2.4%	31.8%	1.4%	1.0%	**46.5%**	0.1%
Controlled	Current	27.9%	40.2%	13.7%	48.6%	22.5%	22.3%	**51.1%**	20.5%
	Next	2.7%	31.0%	2.9%	29.4%	1.5%	4.5%	**40.7%**	0.1%

Table 2: Goal-satisfiability accuracy (%) across models. Task 1 (real-world) evaluates current step recognition (StepAcc) and downstream planning, while Task 2 (controlled) focuses on phase-aware planning. Current and Next denote current-phase completion and next-phase planning accuracy. Best results are shown in bold.

times within a phase. During gastric pouch creation phase, steps such as horizontal stapling, retrogastric dissection, vertical stapling, and hemostasis may repeat multiple times. Models often misinterpret such repetition as phase progression or completion, leading to incorrect phase status estimation and subsequently flawed planning.

These perception errors propagate directly to planning. For most small models, current-phase completion accuracy remains below 25%, while for larger models it is consistently lower than next-phase planning accuracy. This cascading failure indicates that step recognition errors directly undermine current-phase completion judgments. Importantly, this should not be interpreted as limited reasoning ability, as larger models perform well on next-phase planning. Rather, the results expose a dominant perception bottleneck: without reliable video-procedure alignment, even strong language backbones fail to support coherent surgical planning, indicating that end-to-end planning from surgical video is constrained primarily by video understanding rather than higher-level reasoning.

5.5.2 The Reasoning Bottleneck: Under-Constrained Planning Space

Task 2 removes perceptual ambiguity by providing the current step explicitly. Although performance improves relative to Task 1, planning quickly saturates: current-phase completion remains below 52%, and next-phase planning accuracy falls below 5% for small models.

Semantically plausible but procedurally invalid plans. Without explicit procedural guidance, models default to semantic proximity rather than procedural logic. As a result, generated plans often omit critical steps or assemble loosely related actions that fail to collectively achieve the phase goal. For example, sequences may mix steps from the omentum division phase (e.g., omentum exposure and omental transection) with unrelated steps from gastrojejunal anastomosis, including biliary limb measurement or jejunal opening.

This indicates that, although Video-LLMs encode general surgical knowledge, particularly in medical VLMs such as Lingshu and HuluMed, their planning space remains under-constrained. Without explicitly encoded procedural constraints, models fail to produce stable, executable step-level plans, leading to performance saturation even when perceptual uncertainty is removed.

5.5.3 Knowledge Injection: Medical Guidance for Planning

Task 3 introduces explicit medical knowledge to constrain planning, revealing how different forms of knowledge affect model behavior.

Structural Knowledge constrains planning effectively. For most models, structural knowledge (Task 3.1) is the most effective intervention. By explicitly specifying the phase-step hierarchy, procedural structure sharply narrows the space of admissible plans. This leads to large and consistent gains over Task 2 in both current-phase completion and next-phase planning (Table 3), particularly for 7B-scale models. Generated plans exhibit fewer cross-phase intrusions and more reliably satisfy phase-level requirements under goal-satisfiability evaluation. For example, when planning the gastrojejunal anastomosis phase, unrelated steps such as Petersen space exposure or biliary limb opening are more consistently excluded, which were frequently misincorporated in earlier tasks.

Similarly, during jejunojejunal anastomosis planning, extraneous steps such as mesenteric defect exposure or mesenteric defect closure are largely eliminated. These results show that concise structural constraints directly align model generation with phase-level goals.

Semantic descriptions alone cannot enforce procedural correctness. In contrast, semantic descriptions without explicit structure (Task 3.2), as

Task	Metric	Hulu-7B	Hulu-32B	Qwen-7B	Qwen-32B	DAMO	Lingshu-7B	Lingshu-32B	LLaVA
Structural	Current	42.8%	66.0%	63.0%	**73.0%**	43.0%	46.1%	71.1%	23.4%
	Next	51.2%	49.2%	64.3%	68.7%	38.0%	62.6%	**70.5%**	22.9%
Description	Current	26.8%	46.5%	14.0%	44.8%	26.9%	18.5%	**49.7%**	14.1%
	Next	7.7%	26.7%	11.5%	**56.7%**	5.5%	0.9%	52.8%	0.1%
Combined	Current	36.2%	67.6%	55.0%	69.1%	38.6%	34.6%	**70.3%**	19.7%
	Next	41.7%	76.5%	60.8%	**89.2%**	53.8%	56.5%	82.1%	21.3%

Table 3: Goal-satisfiability accuracy (%) of Task 3 under different knowledge injection settings. Current and Next denote current-phase completion and next-phase planning accuracy. Best results are shown in bold.

well as the combined setting (Task 3.3), underperform structural guidance alone for 7B-scale models, especially for next-phase planning(Table 3). Although models often express correct surgical intent, they frequently omit critical execution steps needed to complete the phase goal. Rich semantic prompts tend to promote narrative plausibility rather than enforce discrete procedural requirements.

For instance, after Petersen space closure, models may correctly identify jejunojejunal anastomosis as the next phase, but still fail to include essential steps such as biliary limb opening or alimentary limb measurement. Under long-context prompts, semantic information is often diluted, leading to shallow or generic plans. Models may also lose global procedural coherence, treating jejunojejunal anastomosis as the final phase and appending steps from later phases (e.g., mesenteric defect closure or cleaning and coagulation) in an unstructured manner. These failures show that semantic guidance alone insufficiently constrains the planning space.

5.5.4 Model Capacity Determines Knowledge Integration.

Across all settings, larger models outperform their 7B-scale counterparts, reflecting stronger reasoning capacity and more reliable use of long-context inputs. Structural knowledge (Task 3.1) provides a stable benefit across model sizes, consistently improving planning performance relative to Task 2.

Small models struggle to combine multiple guidance. For 7B-scale models, combining semantic and structural knowledge (Task 3.3) consistently performs worse than structural constraints alone, particularly for next-phase planning (Table 3). This suggests that limited-capacity models have difficulty integrating heterogeneous information sources, leading to weaker adherence to procedural constraints.

Larger models better exploit combined knowledge. For 32B-scale models, Task 3.3 achieves the best next-phase planning performance (Table 3). This indicates that sufficient capacity enables models to leverage semantic descriptions to refine intent-level reasoning while still relying on structural constraints to maintain procedural validity. Notably, structural knowledge alone remains highly competitive even at 32B-scale, suggesting that explicit procedural structure benefits goal-oriented surgical planning across model capacities.

Overall, these results suggest that model capacity governs the ability to integrate multiple forms of knowledge. Semantic enrichment becomes beneficial primarily when sufficient capacity is available to reconcile heterogeneous guidance without compromising procedural validity.

6 Conclusion

This work shows that prevailing formulations and evaluations of surgical planning are misaligned with clinical reality, where planning is hierarchical, goal-oriented, and admits multiple valid execution paths. We introduce a goal-satisfiability-based meta-evaluation benchmark grounded in expert-defined procedural rules to test whether planning metrics align with this setting. Under this benchmark, widely used sequence similarity metrics reject most valid plans, while LLM-based judges, despite strong semantic understanding, frequently fail to enforce critical procedural dependencies. Through progressive evaluation of Video-LLMs, we further show that end-to-end planning is limited by perception bottlenecks and that reasoning remains under-constrained without explicit procedural structure. Injecting structural knowledge provides the most consistent gains, whereas semantic descriptions alone are insufficient and their combination is effective only at larger model scales. Together, these findings motivate a shift from single-trajectory similarity toward goal-satisfiability evaluation as a foundation for developing and interpreting clinically aligned surgical planning models.

Limitations

This study has several limitations that suggest directions for future work.

First, the expert-defined rule-based evaluator relies on manually constructed procedural rules derived from surgical principles and dataset-specific annotation protocols. While this approach provides high precision and interpretability within scope, it does not readily scale to new procedures, institutions, or surgical domains. Automated or semi-automated construction of procedural rules, potentially with LLMs assisting clinicians in formalizing surgical knowledge, remains a challenge.

Second, our evaluation treats goal-satisfiability as a binary criterion. Although appropriate for determining whether a plan can plausibly complete a surgical phase, this formulation does not capture finer-grained aspects of planning quality, such as efficiency, redundancy, or preferences among multiple valid procedural paths. Future work could explore more nuanced, goal-aware metrics without reverting to single-path assumptions.

Third, our experiments are conducted on a limited set of surgical datasets with rich hierarchical annotations. At present, few publicly available datasets provide phase–step hierarchies with sufficient granularity and consistency to support goal-oriented planning analysis. Extending this framework to additional procedures will require broader annotation efforts or alternative forms of weak or implicit supervision.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Diego Biagini, Nassir Navab, and Azade Farshad. 2025. Hierasurg: Hierarchy-aware diffusion model for surgical video generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 310–319. Springer.

Maxence Boels, Yang Liu, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. 2025a. Swag: long-term surgical workflow prediction with generative-based anticipation. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–11.

Maxence Boels, Harry Robertshaw, Thomas C Booth, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. 2025b. Daril: When imitation learning outperforms reinforcement learning in surgical action planning. *arXiv preprint arXiv:2507.05011*.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Zhen Chen, Xingjian Luo, Jinlin Wu, Long Bai, Zhen Lei, Hongliang Ren, Sebastien Ourselin, and Hongbin Liu. 2025. Surgplan++: Universal surgical phase localization network for online and offline inference. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12782–12788. IEEE.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, and 1 others. 2025. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38.

Junming Fan and Pai Zheng. 2024. A vision-language-guided robotic action planning approach for ambiguity mitigation in human–robot collaborative manufacturing. *Journal of Manufacturing Systems*, 74:1009–1018.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, and 1 others. 2025. Hulumed: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Florent Lalys and Pierre Jannin. 2014. Surgical process modelling: a review. *International journal of computer assisted radiology and surgery*, 9(3):495–511.

Joël L Lavanchy, Deepak Alapatt, Luca Sestini, Marko Kraljević, Philipp C Nett, Didier Mutter, Beat P Müller-Stich, and Nicolas Padoy. 2025. Analyzing the impact of surgical technique on intraoperative adverse events in laparoscopic roux-en-y gastric bypass surgery by video-based assessment. *Surgical Endoscopy*, 39(3):2026–2036.

Joël L. Lavanchy, Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Paolo Fiorini, Beat P. Müller-Stich, Philipp C. Nett, Jacques Marescaux, Didier Mutter, and Nicolas Padoy. 2024. Challenges in multicentric generalization: Phase and step recognition in Roux-en-Y gastric bypass surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19(11):2249–2257.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Weili Guan, Dongmei Jiang, and Liqiang Nie. 2025a. Optimus-3: Towards generalist multimodal minecraft agents with scalable task experts. *arXiv preprint arXiv:2506.10357*.

Zhiwei Li, Yong Hu, and Wenqing Wang. 2025b. Encouraging good processes without the need for good answers: Reinforcement learning for llm agent planning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1654–1666.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.

Andres Marzal and Enrique Vidal. 2002. Computation of normalized edit distance and applications. *IEEE transactions on pattern analysis and machine intelligence*, 15(9):926–932.

Mengya Xu, Zhongzhen Huang, Dillan Imans, Yiru Ye, Xiaofan Zhang, and Qi Dou. 2025a. Sap-bench: Benchmarking multimodal large language models in surgical action planning. *arXiv preprint arXiv:2506.07196*.

Mengya Xu, Zhongzhen Huang, Jie Zhang, Xiaofan Zhang, and Qi Dou. 2025b. Surgical action planning with large language models. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 563–572. Springer.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025c. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025b. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. 2023. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*.