LEARNING WHEN TO REVIEW: SELECTIVE MULTI-AGENT VERIFICATION FOR SURGICAL VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable surgical video understanding requires accurate and mutually consistent predictions across perception (*e.g.*, instrument recognition) and interaction understanding (*e.g.*, triplet recognition) to support procedure-level reasoning and summarization (*e.g.*, surgical report generation). Existing approaches use Vision–Language Models to infer these surgical states, yet prediction errors can propagate into downstream reasoning and compromise the factual accuracy of generated reports. Additional model review offers an opportunity to correct these errors, but reviewing every observation incurs redundant computation and can alter predictions that were already correct. Addressing this trade-off requires deciding when further review is useful and how its evidence should guide prediction updates. To this end, we introduce SURGROUTE, a selective multi-agent framework that couples evidence-guided review with controlled correction. Specifically, SURGROUTE comprises three coordinated components. First, Joint Probe Gate constructs candidate hypotheses using procedural priors and jointly probes interaction and phase alternatives, learning from the resulting evidence whether additional review is likely to reduce prediction errors. Second, Multi-Agent Review and Repair reuses the probe response within a heterogeneous reviewer panel and controls revisions through explicit candidate-admission and phase-replacement rules, terminating review when further ratings cannot change the prescribed decision. Finally, Tracker-Assisted Correction integrates instrument-tracking evidence, task-specific label constraints, and temporal phase aggregation to refine predictions from both the skip-review and review branches. Experiments on the benchmark demonstrate that SURGROUTE consistently improves upon base Vision–Language Models across all evaluated surgical tasks while improving inference efficiency.

1 INTRODUCTION

Surgical video understanding aims to describe a procedure at complementary semantic levels: which instruments are present, what actions they perform, which anatomical structures they interact with, and how these events fit within the surgical workflow. A coherent interpretation therefore requires more than independently accurate labels. Instrument, verb, and target predictions must support the instrument–verb–target triplet [31], while phase recognition must account for temporal context beyond an isolated action [25]. Developing models that recover these coupled states reliably is a central step toward comprehensive surgical scene understanding [21, 32, 49].

Recent progress has approached this goal through stronger visual representations and more expressive task modeling. Action-triplet recognition explicitly captures compositional surgical interactions [32], while surgical vision–language pretraining improves transfer through narrated video, hierarchical supervision, and procedural knowledge [53, 54, 55]. Temporally annotated tool trajectories provide a complementary source of evidence about instrument presence and persistence [34]. Together, these advances support increasingly detailed predictions from surgical video. Nevertheless, visual ambiguity remains unavoidable: occlusion or a change in viewpoint can obscure instrument–tissue contact, and visually similar frames can occur in different surgical phases [25].

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These limitations motivate an additional review stage that revisits an initial prediction using further model evidence. Feedback-based refinement, tool-interactive critique, and multiagent debate provide mechanisms for revisiting initial outputs [6, 11, 27]. Agentic methods additionally incorporate actions and remembered feedback [41, 52]. However, requesting a revision and trusting it are distinct decisions [42]: the opportunity to revise does not establish its benefit (*cf.* Fig. 1). Studies of intrinsic self-correction show that reasoning performance can degrade when revisions lack reliable external feedback [14]. Stable frames may gain little from repeated review, whereas an unconstrained revision can replace a correct component or introduce an incompatible interaction. Selective prediction and adaptive computation address the allocation of inference effort [3, 10, 35], but surgical understanding also requires control over the resulting edits. The key question is whether further review improves the coupled prediction under the chosen correction rules, beyond identifying uncertainty.

We argue that preliminary review evidence provides a direct connection between these two decisions. Assessing explicit candidates reveals whether the current labels are supported, whether a plausible interaction was omitted, and whether an alternative phase deserves consideration. This evidence can both inform a routing decision and become the first part of a larger review when needed. At the same time, separating model-based semantic review from explicit instrument and phase correction makes their effects easier to control. This perspective leads to our central question: *How can joint review evidence guide efficient and consistent surgical prediction updates?*

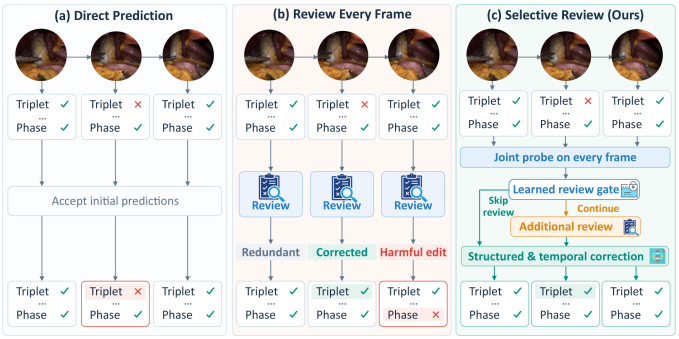

Figure 1: **Motivation for selective review.** (a) Direct prediction can retain errors. (b) Reviewing every frame adds cost and risks harmful edits. (c) SURGROUTE selectively extends joint probing, then applies structured and temporal correction. Symbols show possible outcomes; ellipses indicate additional tasks.

To address this question, we introduce SURGROUTE, which organizes prediction refinement into three coordinated components. First, a *Joint Probe Gate* (JPG) constructs an explicit candidate pool from an initial joint prediction and procedural priors, then assesses interaction and phase alternatives in one probe request. A learned router uses this feedback to identify observations for which further review is likely to reduce total prediction errors. Second, *Multi-Agent Review and Repair* (MAR) reuses the probe within a shared heterogeneous reviewer panel. Candidate admission, phase replacement, and ambiguity-preserving updates are determined by explicit rules, while exact early stopping avoids requests that cannot change the prescribed decision. Third, *Tracker-Assisted Correction* (TC) combines instrument evidence, label rules, and phase aggregation to refine the selected state.

We evaluate SURGROUTE on SurAgent-Bench using six vision–language backbones to assess its contribution to surgical recognition and downstream report generation. Across these backbones, triplet F1 improves by 4.33–11.88 points, alongside higher report-generation scores (*cf.* Table 1). These findings suggest that structured refinement of surgical predictions can benefit their subsequent use in language generation. Complementary diagnostic analyses examine the contributions of review allocation, controlled repair, and tracker-assisted correction, and assess how selective review and decision-preserving stopping balance prediction quality with computational cost (Section 4.4).

Our main contributions are as follows:

- We introduce a Joint Probe Gate that learns when review benefits coupled predictions from interaction and phase evidence, using net error reduction across all tasks as supervision.
- We develop Multi-Agent Review and Repair to reuse probe evidence and control revisions through component-aware admission and phase-update rules, with exact early stopping that avoids extra reviewer calls when remaining ratings cannot change the full-panel decision.
- We combine instrument tracking, label constraints, and temporal phase aggregation in Tracker-Assisted Correction to refine both review branches without additional model calls.

2 Related Work

Surgical Video Understanding. General-purpose vision–language models, including Gemini 3.8 Flash [4], Grok 4.6 [45], Qwen 3.8 Max [47], GLM 5.3 Flash [57], GPT 5.6 Sol [36], and Kimi K3 [46], provide general multimodal interfaces. Surgical pretraining adds domain-specific supervision: SurgVLP aligns narrated video and language [55], HecVL introduces hierarchical alignment [53], and PeskaVLP incorporates procedural knowledge [54]. Compositional methods model instrument–verb–target associations [31, 32, 33]; temporal convolution and attention model workflow [2, 8, 25, 37]; and tracking provides persistent tool identities [34]. RASO uses weak supervision for surgical-object recognition [19]. Surgical-VQA and SurgicalGPT study visual question answering [39, 40], while LLaVA-Surg and Surgical-LLaVA develop instruction-oriented surgical assistants [16, 18]. Scene graphs support structured report generation [22], and SurAgent-Bench links visual recognition with language annotations [21]. Building on these representations and backbones, SurgRoute focuses on selectively acquiring review evidence and controlling revisions to coupled surgical states before downstream reporting.

Selective Computation and Agentic Refinement. Selective prediction and learning to defer study when to abstain or seek expert input [9, 30]. SelectiveNet jointly learns prediction and rejection [10], routing and cascades balance model quality against cost [1, 35], and input-adaptive allocation predicts the benefit of additional computation [3]. For response refinement, Self-Refine uses self-feedback [27], CRITIC incorporates tool-based feedback [11], and ART learns when to refine and whether to trust a revision [42]. ReAct connects reasoning with actions [52]; Reflexion and LongMem retain feedback or historical context [41, 50]; and multiagent debate and dynamic agent networks coordinate model instances [6, 26]. SurgRoute couples reusable probe evidence with controlled candidate updates and shared temporal correction, tying review allocation to the surgical-label changes it can authorize.

3 Method

3.1 Problem Formulation and Overview

SurgRoute combines selective evidence acquisition (JPG), controlled revision (MAR), and shared temporal correction (TC) (*cf.* Fig. 2). For target frame x_f, a base vision–language model f_θ predicts within ontology Ω from three current and past observations:

$$W_f = (x_{f-2\Delta}, x_{f-\Delta}, x_f), \qquad h_f^0 = f_\theta(W_f; \Omega), \tag{1}$$

Here, $\Delta = 25$ corresponds to one second at 25 fps. H0 denotes $h_f^0 = (\mathcal{I}_f^0, \mathcal{V}_f^0, \mathcal{T}_f^0, \mathcal{R}_f^0, p_f^0)$: instrument, verb, target, and triplet label sets, followed by a phase $p_f^0 \in \mathcal{P} = \{0, \dots, 6\}$. For triplet r, $\pi_q(r)$ extracts its component $q \in \mathcal{Q}_{\text{cmp}} = \{\text{I}, \text{V}, \text{T}\}$.

JPG constructs prior-guided candidates and a preliminary state h_f^{c} (the cheap prediction), obtains Qwen probe evidence, and routes selected frames to MAR. TC follows branch selection:

$$\widetilde{h}_f = \text{Select}_{a_f}(h_f^{\text{c}}, h_f^{\text{r}}), \qquad h_f = \text{TC}(\widetilde{h}_f; e_f, \mathcal{H}_f^-), \tag{2}$$

Select_a returns h_f^{c} for $a = 0$ and h_f^{r} otherwise. TC uses tracker/probe evidence e_f and raw-phase history $\mathcal{H}_f^-$. Processing is chronological, with history reset at each video boundary.

3.2 Joint Probe Gate (JPG)

Prior-guided candidates. JPG expands H0 into a pool $\mathcal{C}_f$ using prior-conditioned base-model proposals and their triplet components. Candidate $c = (q, \ell)$ pairs an ontology-valid task and label. Let $\rho_f(r) = \rho_{-v}(r \mid p_f^0)$ denote triplet frequency under the H0 phase, estimated from training videos excluding query v. Prior rules update the preliminary state:

$$\mathcal{R}_f^{\text{c}} = (\mathcal{R}_f^0 \setminus \mathcal{R}_f^-) \cup \mathcal{R}_f^+, \qquad h_{f,q}^{\text{c}} = h_{f,q}^0 \cup \{\pi_q(r) : r \in \mathcal{R}_f^+\}, \tag{3}$$

Here, $h_{f,q}$ denotes the label set for task q. The removed and added sets are $\mathcal{R}_f^- = \{r \in \mathcal{R}_f^0 \setminus \mathcal{N} : \rho_f(r) < 0.01\}$ and $\mathcal{R}_f^+ = \{r \in \mathcal{C}_f^{\text{R}} \setminus (\mathcal{R}_f^0 \cup \mathcal{N}) : \rho_f(r) \geq 0.70\}$, where $\mathcal{C}_f^{\text{R}}$ contains

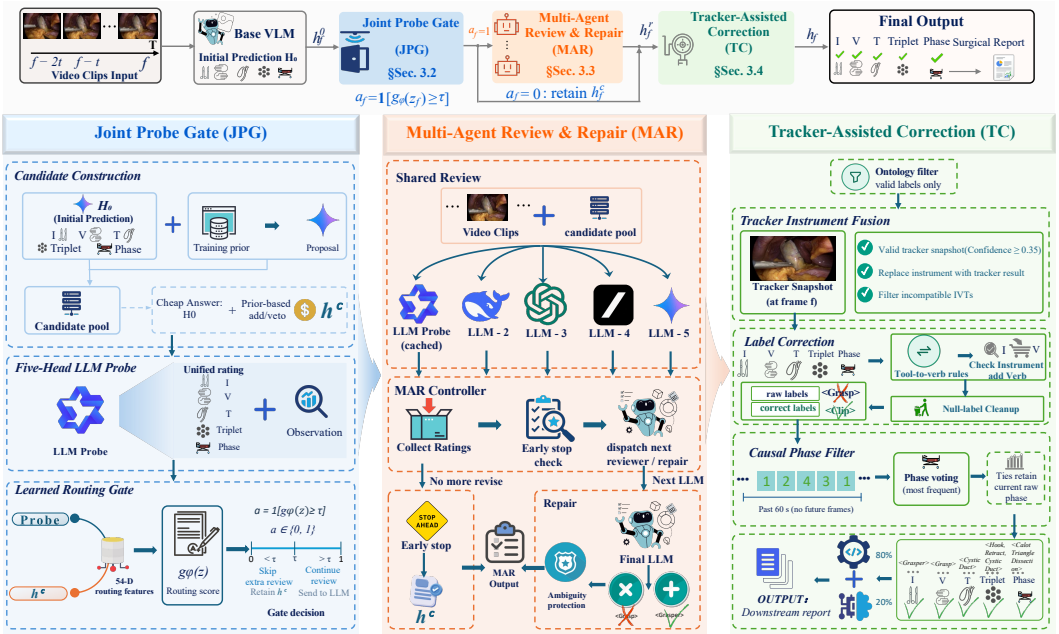

Figure 2: **Overview of SURGROUTE. (a) JPG** builds prior-guided candidates, jointly probes interaction and phase alternatives, and selects frames for further review. **(b) MAR** reuses the probe within a shared reviewer panel, controls candidate and phase updates, and stops when further ratings cannot affect the decision. **(c) TC** applies instrument evidence, label rules, and temporal phase aggregation to either branch before report generation.

candidate triplets and $\mathcal{N}$ contains null triplets. Additions insert their components; removals preserve component labels. Null triplets are exempt, and review retains the H0 phase bucket.

Joint evidence probing. A Qwen joint probe rates $\mathcal{C}_f$ and all seven phases: scores 1–2 indicate contradiction, 3 uncertainty, and 4–5 support. Its validated response yields $z_f = [z_f^{\mathrm{H0}}; z_f^{\mathrm{prior}}; z_f^{\mathrm{int}}; z_f^{\mathrm{phase}}] \in \mathbb{R}^{54}$, combining H0 structure, prior/proposal statistics, and interaction/phase evidence. Tracker observations and later ratings are excluded; implementation details appear in Appendix A.1.

Using a shared candidate vocabulary allows the probe to expose disagreement between an interaction and its components, while phase ratings describe competing workflow interpretations. These complementary signals inform routing and are reused by MAR on selected frames, connecting the decision to acquire more evidence with the subsequent repair of the coupled prediction.

Learning when to review. We train JPG on fully annotated frames from the designated training split, using the net error reduction achieved by full review. Both branches undergo the same instrument and label corrections, yielding $\bar{h}_f^{\mathrm{skip}}$ and $\bar{h}_f^{\mathrm{rev}}$ before temporal phase filtering:

$$d_f = \sum_{q \in \mathcal{Q}} \left(\left| \bar{h}_{f,q}^{\mathrm{skip}} \triangle y_{f,q} \right| - \left| \bar{h}_{f,q}^{\mathrm{rev}} \triangle y_{f,q} \right| \right), \qquad b_f = \mathbb{I}[d_f > 0], \tag{4}$$

Here, $\mathcal{Q} = \{\mathrm{I}, \mathrm{V}, \mathrm{T}, \mathrm{R}, \mathrm{P}\}$ indexes the five tasks, $y_{f,q}$ is ground truth, $\triangle$ denotes symmetric difference, and $\mathbb{I}$ is the indicator. Phase is a singleton set. Thus, d_f counts corrected minus introduced errors across the coupled prediction. Correcting one task is insufficient if errors introduced elsewhere cancel its benefit. The label $b_f = 1$ therefore identifies net improvement under the prescribed repair procedure, rather than merely detecting whether a prediction was edited.

A gradient-boosted classifier g_ϕ learns this target and routes by threshold τ:

$$\mathcal{L}_{\mathrm{gate}}(\phi) = \frac{1}{N} \sum_{f \in \mathcal{D}_{\mathrm{tr}}} w_{b_f} \ell_{\mathrm{bin}}(b_f, g_\phi(z_f)), \qquad a_f = \mathbb{I}[g_\phi(z_f) \geq \tau], \tag{5}$$

Here, $N = |\mathcal{D}_{\mathrm{tr}}|$ counts fully annotated training frames; ℓ_{bin} is binary log loss and w_{b_f} balances the classes. On the validation split, threshold selection minimizes review frequency while keeping

pooled errors no higher than skip review and retaining at least 90% of always review's positive mean-F1 gain. These aggregate constraints use final outputs after TC (Appendix A.1). Ground truth is used only for offline fitting and validation.

3.3 Multi-Agent Review and Repair (MAR)

MAR reuses the probe as the first of $J = 5$ heterogeneous reviewers. Each receives the same H0, candidate pool, and context W_f, without other reviewers' judgments. The shared panel supports both interaction and phase decisions.

Representing judgments as bounded, candidate-specific ratings separates evidence collection from the repair decision. Each potential edit is evaluated under its prescribed admission rule, and the remaining ratings have known limits that certify when completing the panel cannot change the output. Triplet and component support are checked separately, so a high triplet score cannot compensate for a weak component when authorizing a score-based addition. This makes compositional agreement an explicit condition of the repair rule.

Component-aware admission. A novel triplet qualifies for score-based addition only when it and all three components receive sufficient support:

$$\mu_f(c) = \frac{1}{J}\sum_{j=1}^{J} s_{f,j}(c), \qquad A_f(r) = \prod_{c\in\Gamma(r)} \mathbb{I}[\mu_f(c) \geq \alpha], \qquad \alpha = 4, \tag{6}$$

Here, $s_{f,j}(c) \in \{1,\dots,5\}$ is reviewer j's rating, $\mu_f(c)$ is its complete-panel mean, and $\Gamma(r) = \{(\mathrm{R}, r)\} \cup \{(q, \pi_q(r)) : q \in \mathcal{Q}_{\mathrm{cmp}}\}$. Invalid means cannot authorize edits. Starting from H0, component labels are added at $\mu_f \geq 4$ and existing labels removed at $\mu_f \leq 2$, except components required by retained triplets. Other labels retain their membership. JPG's prior rules then apply as a separate admission path, followed by restoring <*Grasp*>/<*Retract*> verbs and triplets to their memberships in h_f^{c} to preserve ambiguous predictions (Appendix A.1).

Shared phase selection. With all seven phase means valid, define $\mathcal{B}_f = \arg\max_{p\in\mathcal{P}} \mu_f(\mathrm{P}, p)$, writing $\mu_f(q,\ell) = \mu_f((q,\ell))$. MAR accepts a unique, supported winner only if it exceeds H0:

$$p_f^{\mathrm{r}} = \begin{cases} p^*, & \mathcal{B}_f = \{p^*\}, \quad \mu_f(\mathrm{P}, p^*) \geq \alpha, \quad \mu_f(\mathrm{P}, p^*) > \mu_f(\mathrm{P}, p_f^0), \\ p_f^0, & \text{otherwise}, \end{cases} \tag{7}$$

Missing phase evidence or early stopping retains p_f^0; no separate phase query is needed.

Decision-preserving early stopping. For a candidate with k valid prefix ratings, let $S_{f,k}(c) = \sum_{j=1}^{k} s_{f,j}(c)$. Every completed-panel mean lies within

$$L_{f,k}(c) = \frac{S_{f,k}(c) + (J-k)}{J}, \qquad U_{f,k}(c) = \frac{S_{f,k}(c) + 5(J-k)}{J}, \tag{8}$$

Thus, $L_{f,k}(c) > 2$ rules out removal and $U_{f,k}(c) < 4$ rules out addition. A phase alternative remains viable only if its upper bound reaches 4 and exceeds the current phase's lower bound. MAR stops and retains h_f^{c} when these bounds rule out every editable interaction and phase change; otherwise it completes the panel. This preserves full-panel selection; correctness still depends on the candidate pool and reviewer evidence (Appendix A.2).

3.4 Tracker-Assisted Correction (TC)

Instrument and label correction. TC applies to both branches. An aligned frozen tracker provides instrument classes $\mathcal{D}_f$ with confidence at least 0.35, filtered to the triplet ontology. For valid evidence within output capacity,

$$\mathcal{I}_f^{\mathrm{TC}} = \mathcal{D}_f, \qquad \mathcal{R}_f^{\mathrm{TC}} = \{r \in \widetilde{\mathcal{R}}_f : \pi_{\mathrm{I}}(r) \in \mathcal{D}_f\}, \tag{9}$$

where $\widetilde{\mathcal{R}}_f$ is the selected triplet set. Unavailable or over-capacity evidence bypasses fusion; a valid empty set does not. TC then applies instrument–verb rules and probe-based null-label cleanup without additional model calls (Appendix A.1).

Temporal phase aggregation. TC votes over raw selected phases in a causal 60-second window:

$$n_f(p) = \sum_{j \in \mathcal{F}_{v,f}} \mathbb{I}[\widetilde{p}_j = p], \qquad p_f = \begin{cases} p_f^*, & \mathcal{M}_f = \{p_f^*\}, \\ \widetilde{p}_f, & |\mathcal{M}_f| > 1, \end{cases} \tag{10}$$

Here, $\mathcal{F}_{v,f} = \{j \in \mathcal{F}_v : f - \nu W < j \leq f\}$ contains processed targets of video v, with $\nu = 25$ fps and $W = 60$ seconds. The raw phase $\widetilde{p}_j$ is the accepted MAR phase or H0; $\mathcal{M}_f = \arg\max_{p \in \mathcal{P}} n_f(p)$ contains the modes. A unique mode p_f^* is selected; ties retain $\widetilde{p}_f$. The buffer records the branch-selected phase before temporal aggregation. A smoothed phase affects the current output but is never inserted as new evidence for later votes; subsequent decisions use only the original selected phases available within the same video's trailing window.

The corrected state supports report generation. H0, proposal, and probing require three calls; up to four additional reviewers give $3 + K_f$ calls per frame, $K_f \in \{0, \ldots, 4\}$. TC is deterministic.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate SurgRoute on SurAgent-Bench [21], which contains 34,499 frames and 275,992 question–answer pairs spanning eight interconnected tasks. The benchmark integrates public surgical datasets with a private clinical cohort of 100 patients. We use a study-specific 10:2:8 training/validation/test split by frame count. We select six representative tasks: five visual tasks—instrument, verb, target, instrument–verb–target triplet, and surgical phase recognition—and one language task, surgical report generation. Together, they cover surgical perception, interaction and workflow understanding, and language generation, enabling evaluation of how recognition quality affects downstream reporting. Appendix A.3 describes benchmark composition and annotations.

Base Models and Competitors. We implement SurgRoute on six general-purpose vision–language models: Gemini 3.8 Flash [4], Grok 4.6 [45], Qwen 3.8 Max [47], GLM 5.3 Flash [57], GPT 5.6 Sol [36], and Kimi K3 [46]. We compare each SurgRoute variant with its corresponding standalone backbone, yielding six paired comparisons (*cf.* Table 1). We also evaluate HecVL [53], PeskaVLP [54], and SurgVLP [55] using their authors' released weights pretrained on large-scale surgical data. The framework accesses its general-purpose backbones through API-based inference. In each paired comparison, the standalone backbone produces the joint first-pass prediction, while the corresponding SurgRoute variant applies selective review and correction to that prediction. Both use the same three-frame visual context and label ontology, allowing us to assess the contribution of the proposed pipeline across different backbones. Further implementation and input details are provided in Appendices A.1 and A.4.

Evaluation Metrics. We report micro precision, micro F1, and exact-set accuracy for instrument, verb, target, and triplet recognition, and accuracy for phase recognition. All model rows share the report generator, judge, prompts, and scoring protocol: $\text{Total} = 0.8\,\text{Rule} + 0.2\,\text{Judge}$ combines rule-based and LLM-based scores. Recognition metrics are percentages; report scores range from 0 to 100. Diagnostics additionally measure review allocation, token consumption , introduced errors, decision agreement, and phase-transition accuracy (Appendix A.4).

4.2 Quantitative Results

Visual Understanding Tasks. Results on the five visual tasks highlight the difficulty of recovering surgical interactions from otherwise recognizable components (*cf.* Table 1). Surgical LVLMs achieve limited fine-grained recognition performance, while general-purpose models improve instrument recognition substantially but remain less effective at triplet recognition. SurgRoute improves all four recognition F1 scores and phase accuracy across the six backbones. For instance, triplet F1 increases by 11.14, 8.49, and 11.88 percentage points for Qwen 3.8 Max, GLM 5.3 Flash, and Kimi K3, respectively (*cf.* Fig. 3). With Gemini 3.8 Flash, it reaches 46.40, alongside 87.45 phase accuracy. These gains show that selective review benefits both compositional recognition and workflow understanding across diverse backbones, although exact-set accuracy does not improve for every backbone–task pair.

Table 1: **Surgical recognition and report-generation performance.** Recognition metrics are percentages; report scores use a 0–100 scale. $\text{Total} = 0.8\,\text{Rule} + 0.2\,\text{Judge}$. Bold/underline mark the best/second-best distinct column values, including ties. Metric definitions appear in Appendix A.4.

Model	Instrument Rec.			Verb Rec.			Target Rec.			Triplet Rec.			Phase Rec.	Surgical Report Gen.		
	Prec.	F1	Acc.	Prec.	F1	Acc.	Prec.	F1	Acc.	Prec.	F1	Acc.	Acc.	Rule	Judge	Total
Surgical LVLMs																
HecVL [53]	20.32	22.62	0.58	27.21	37.09	0.02	15.68	20.51	0.72	1.97	2.18	0.00	27.25	9.02	53.76	17.97
PeskaVLP [54]	33.47	31.62	3.32	19.06	26.83	0.03	17.49	26.90	0.00	1.34	2.21	0.00	13.92	8.63	49.24	16.75
SurgVLP [55]	26.45	31.15	0.59	17.74	24.28	0.75	8.96	13.43	0.32	1.98	3.25	0.05	23.74	12.29	51.17	20.07
General-Purpose LVLMs																
Gemini 3.8 Flash [4]	89.92	86.75	72.77	63.26	62.14	39.42	57.36	57.28	30.05	40.65	39.99	18.21	74.37	44.54	69.55	49.54
Grok 4.6 [45]	81.55	79.00	58.44	62.26	60.80	35.65	57.46	52.62	26.44	37.67	37.19	15.95	52.51	43.03	68.59	48.14
Qwen 3.8 Max [47]	82.29	82.66	61.87	51.69	54.90	21.80	46.23	50.34	14.96	25.17	27.64	5.55	56.64	32.37	63.88	38.67
GLM 5.3 Flash [57]	80.00	75.86	54.20	59.07	56.76	30.78	44.69	47.05	15.36	30.56	30.38	10.27	51.47	32.89	64.76	39.26
GPT 5.6 Sol [36]	77.26	74.58	52.15	60.08	58.15	31.80	48.26	46.97	20.03	31.26	30.82	11.23	49.10	33.92	65.50	40.24
Kimi K3 [46]	74.24	73.70	47.55	45.77	47.34	18.76	36.48	41.10	9.74	16.69	18.84	5.78	46.85	24.63	61.10	31.92
Our Model																
SURGROUTE$^{\text{Gemini 3.8 Flash}}$	95.61	94.03	86.89	69.30	70.12	45.42	62.05	**62.34**	**31.56**	**54.44**	**46.40**	**20.55**	**87.45**	**50.35**	**71.12**	**54.50**
SURGROUTE$^{\text{Grok 4.6}}$	95.45	94.87	86.03	**70.80**	**71.97**	**49.39**	**67.30**	59.01	26.50	46.54	41.97	19.05	63.45	44.16	69.84	49.29
SURGROUTE$^{\text{Qwen 3.8 Max}}$	95.76	94.48	**88.21**	57.74	61.65	27.97	50.20	53.94	18.67	45.27	38.78	12.65	66.30	43.53	68.58	48.54
SURGROUTE$^{\text{GLM 5.3 Flash}}$	**96.01**	**94.96**	86.01	62.30	66.81	34.31	45.88	48.28	13.56	51.96	38.87	12.51	55.30	39.01	67.56	44.72
SURGROUTE$^{\text{GPT 5.6 Sol}}$	95.86	94.84	85.73	61.84	65.83	32.16	50.20	47.94	18.81	50.83	35.15	10.70	58.82	35.02	66.08	41.23
SURGROUTE$^{\text{Kimi K3}}$	95.75	94.80	85.68	54.89	61.40	27.11	40.87	46.68	9.60	45.18	30.72	6.27	62.78	31.30	63.57	37.75

Open-Language Task. The report-generation results further examine whether improved visual predictions translate into better procedural summaries (*cf.* Table 1). SURGROUTE improves Rule and Judge scores across all six backbones, indicating benefits under both structured-content checks and model-based assessment. These gains yield an average increase of 4.71 points in Total, with Qwen 3.8 Max improving from 38.67 to 48.54. The Gemini 3.8 Flash variant achieves the strongest overall report-generation performance, raising Total from 49.54 to 54.50. Together, these results support refining structured surgical states before using them for downstream generation.

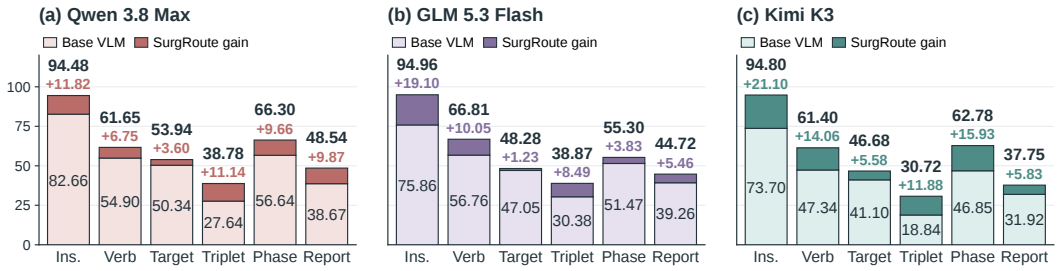

Figure 3: **Performance gains on selected backbones.** Pale segments show base-model scores and dark segments show gains from SURGROUTE. The first four tasks use F1, phase uses accuracy, and reports use Total. Recognition gains are percentage points; report gains are score points.

4.3 CASE STUDY

We compare four recognition tasks and surgical report generation (*cf.* Fig. 4). Each visual row shows a distinct example; the lower panel compares Gemini 3.8 Flash with SURGROUTE on a separate report input. Two findings emerge from these examples.

Finding I: Correct component recognition does not guarantee accurate interactions. Gemini 3.8 Flash identifies the correct instruments and targets in their respective examples, yet substitutes <*Grasp*> for <*Dissect*> in the verb case. In the triplet case, it predicts <*Grasper, Dissect, Cystic Artery*> rather than <*Hook, Dissect, Cystic Duct*>. SURGROUTE corrects both errors while retaining the valid <*Grasper, Retract, Gallbladder*> interaction and the correct component labels. These examples illustrate how controlled refinement can improve fine-grained recognition without overwriting predictions already supported by the image.

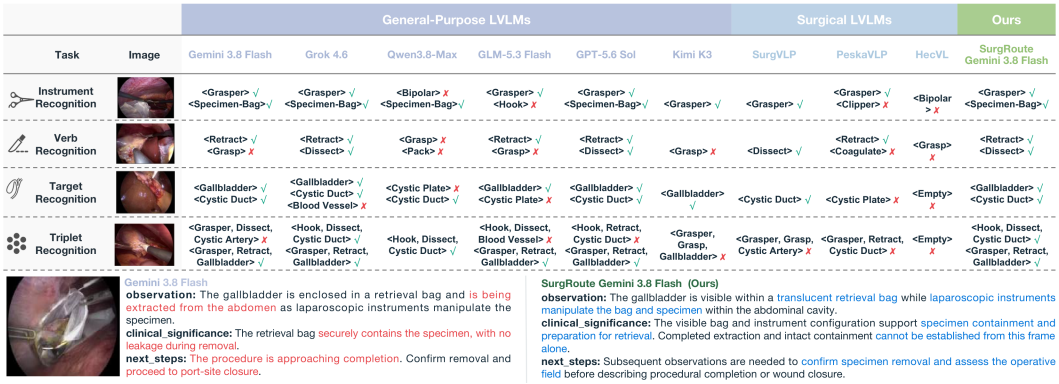

Figure 4: **Illustrative recognition and report-generation examples.** Checks and crosses mark correct and incorrect labels. The lower panel compares Gemini 3.8 Flash and SURGROUTE on a separate input; red highlights unsupported claims and blue highlights grounded refinements.

Finding II: Grounded reporting limits unsupported procedural conclusions. Recognizing the retrieval bag does not establish that extraction is underway, containment is intact, or closure is imminent. Nevertheless, the baseline report makes these procedural claims. SURGROUTE describes the visible bag and instrument manipulation, relates them to preparation for retrieval, and leaves completed extraction and intact containment unconfirmed. It also identifies the further observations needed before asserting completion. Such qualification connects visual observations to procedural context without overstating what has occurred.

4.4 DIAGNOSTIC EXPERIMENTS

We perform a series of ablation studies on Gemini 3.8 Flash to analyze the contributions of our proposed components. Appendices A.4 and A.5 detail the controls and metrics.

Table 2: **Component contributions.** Recognition F1 and phase accuracy (%).

Components			Recognition F1↑				Phase
JPG	MAR	TC	Ins.	Verb	Tar.	Triplet	Acc.↑
×	×	×	86.75	62.14	57.28	39.99	74.37
×	✓	×	86.97	62.88	59.38	41.57	75.27
×	✓	✓	**94.71**	67.55	**62.60**	43.77	85.33
✓	✓	×	87.25	62.96	61.22	40.23	74.97
✓	✓	✓	94.03	**70.12**	62.34	**46.40**	**87.45**

Table 3: **Review policies.** Metrics (%); Token Consumption (M).

Metric	Skip	Always	**JPG**
Review	0.00	100.00	32.93
ER↑	0.00	100.00	33.02
BVR↑	–	2.78	**6.29**
Triplet F1↑	46.21	46.43	46.40
Phase Acc.↑	87.63	86.86	87.45
Tokens↓	331.79	653.74	450.07

Component Contributions. We compare JPG, MAR, and TC as functional bundles (*cf.* Table 2). The full pipeline improves triplet F1 by 6.41 points and phase accuracy by 13.08 points over H0, while also improving all component scores. Removing TC reduces instrument F1 by 6.78 points and phase accuracy by 12.48 points. Removing JPG reduces triplet F1 by 2.63 points; this jointly disables prior-guided candidates, preliminary correction, and selective routing. The full model leads on verb, triplet, and phase recognition; omitting JPG slightly improves instrument and target F1.

Review Allocation and Cost. Candidates, probing, MAR rules, and TC are fixed (*cf.* Table 3). Error recall (ER) measures selection of erroneous cheap predictions; beneficial verification rate (BVR)

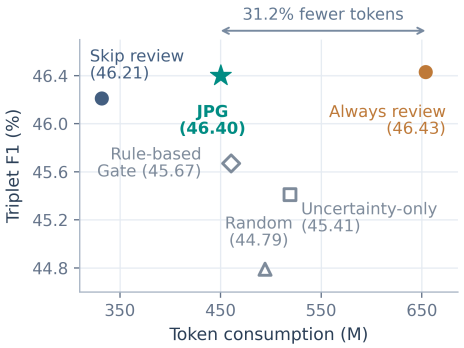

Figure 5: **Triplet F1 versus tokens.** Random, Uncertainty-only, and Rule-based Gate are simulated settings; token totals are measured from model usage and reported in millions.

measures how often MAR reduces net errors on reviewed, scorable frames. Both are computed before TC; final recognition scores include its corrections. JPG reviews 32.93% of frames, achieving 33.02% ER and 6.29% BVR, compared with 2.78% BVR for always review. The cost comparison shows 31.2% fewer tokens at a cost of 0.03 triplet F1 points and includes three simulated routing settings (*cf.* Fig. 5). Relative to skip review, JPG gains 0.19 F1 points with 35.6% more tokens. Its phase accuracy lies 0.59 points above always review and 0.18 below skip review (*cf.* Table 3).

Table 4: **Joint-probe evidence.** Review target: 50%. Review, gain recall, and gain precision are percentages. Net reports the number of errors removed per 100 frames.

Probe	Review	Recall↑	Prec.↑	Net↑
H0 + prior	50.60	72.06	48.42	37.10
+ Interaction	49.80	85.00	58.03	53.40
+ Phase	50.30	81.47	55.07	48.80
Joint probe	49.50	**89.41**	**61.41**	**59.20**

Table 5: **MAR rules and stopping.** Triplet F1 and agreement (%); harm per 100 routed frames; additional calls per routed frame. Phase accuracy: 87.45%.

Variant	F1↑	Harm↓	Calls↓	Agree.↑
Full-panel	**46.40**	**7.85**	4.00	**100.00**
Triplet-only	44.92	13.26	4.00	90.41
No rollback	45.37	11.58	4.00	87.88
Exact stopping	**46.40**	**7.85**	**2.85**	**100.00**

Joint-Probe Evidence. We add interaction and phase evidence at a common review target (*cf.* Table 4). Gains follow Equation 4, distinct from pre-TC BVR. Joint probing raises gain recall from 72.06% to 89.41% and precision from 48.42% to 61.41%, increasing net errors removed from 37.10 to 59.20 per 100 frames. Both probes contribute at similar review rates.

Controlled Repair and Early Stopping. Routing, candidates, and reviewer responses are fixed (*cf.* Table 5). Relative to full-panel MAR, triplet-only admission lowers triplet F1 from 46.40 to 44.92 and raises introduced errors from 7.85 to 13.26 per 100 routed frames. Removing rollback yields 45.37 F1 and 11.58 introduced errors. These controls support component checks and ambiguity preservation. Exact stopping retains 100.00% output agreement while reducing additional calls from 4.00 to 2.85 per routed frame (28.8%), complementing JPG's token savings.

Instrument and Temporal Correction. We vary label rules and phase history independently (*cf.* Table 6). Label rules raise triplet F1 from 46.18 to 46.40. Using raw rather than filtered history raises phase accuracy from 79.86% to 87.45% and transition accuracy from 18.20% to 27.60%. Raw history avoids recirculating earlier filtered decisions, although transitions remain difficult. With pre-TC states and routing fixed, these changes arise within the correction stage. The controls distinguish label constraints from temporal aggregation, both of which reuse available evidence without additional model calls.

Table 6: **TC controls.** Label rules and phase history vary separately with fixed pre-TC states and routing. Instrument F1 is 94.03 throughout.

(a) Label rules

Setting	Triplet F1↑
Disabled	46.18
Enabled	**46.40**

(b) Phase history

History	Phase Acc.↑	Transition Acc.↑
Recursive	79.86	18.20
Raw	**87.45**	**27.60**

5 Conclusion and Discussion

Conclusion. We presented SurgRoute to connect review allocation with controlled correction in joint surgical video understanding. The framework uses the expected benefit of revision to guide evidence acquisition and explicit update rules to govern changes to coupled predictions. Results on the evaluated benchmark support this design, suggesting that coordinating when to review and how to revise can help balance prediction quality and computational cost. Diagnostic analyses further suggest that the value of review depends on both evidence allocation and update control, since revisions can correct existing errors while introducing new ones across coupled tasks.

Discussion. Exact early stopping reduces additional calls while preserving the full-panel decision, although this agreement does not guarantee correctness. Our ongoing work explores reviewer scheduling and validation-based window selection to address the current fixed reviewer order and temporal window, with the aim of reducing redundant computation and improving responsiveness near phase transitions while maintaining stable predictions during sustained surgical activities.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and code debugging. All technical content, experimental design, results, and interpretations were verified by the authors.

REFERENCES

[1] Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

[2] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 343–352, 2020.

[3] Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. Learning how hard to think: Input-adaptive allocation of lm computation. In *International Conference on Learning Representations (ICLR)*, pp. 102783–102802, 2025.

[4] Google DeepMind. Gemini 3.8 flash: Model card. Official model card, 2026. Available at https://deepmind.google/models/model-cards/gemini-3-8-flash/.

[5] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, 2024.

[6] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

[7] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[8] Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 593–603, 2021.

[9] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[10] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, pp. 2151–2159, 2019.

[11] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations (ICLR)*, pp. 57734–57811, 2024.

[12] Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330. PMLR, 2017.

[14] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations (ICLR)*, pp. 32808–32824, 2024.

[15] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, 2023.

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

[17] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 28541–28564, 2023.

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

[19] Jiajie Li, Brian Quaranto, Chenhui Xu, Ishan Mishra, Ruiyang Qin, Dancheng Liu, Peter Kim, and Jinjun Xiong. Recognize any surgical object: unleashing the power of weakly-supervised data. In *International Conference on Learning Representations (ICLR)*, pp. 33811–33823, 2025.

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023.

[21] Pengpeng Li, Chenghao Hu, Guikun Chen, Liulei Li, Hongyu Qu, Yifei Feng, Xiangbo Shu, Wenguan Wang, and Jinhui Tang. SurAgent: Benchmarking multi-agent collaboration for vision-language reasoning in surgical video. Manuscript. Project URL: https://github.com/GeorgeHuLite/SurAgent, 2026.

[22] Chen Lin, Shuai Zheng, Zhizhe Liu, Youru Li, Zhenfeng Zhu, and Yao Zhao. SGT: Scene graph-guided transformer for surgical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 507–518, 2022.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 34892–34916, 2023.

[24] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, 2023.

[25] Yang Liu, Maxence Boels, Luis C. Garcia-Peraza-Herrera, Tom Vercauteren, Prokar Dasgupta, Alejandro Granados, and Sébastien Ourselin. Lovit: Long video transformer for surgical phase recognition. *Medical Image Analysis*, 99:103366, 2025.

[26] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic llm-powered agent network for task-oriented agent collaboration. In *Conference on Language Modeling*, 2024.

[27] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 46534–46594, 2023.

[28] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9004–9017, 2023.

[29] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367, 2023.

[30] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning (ICML)*, pp. 7076–7087, 2020.

[31] Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 364–374, 2020.

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

[33] Chinedu Innocent Nwoye, Deepak Alapatt, Tong Yu, Armine Vardazaryan, Fangfang Xia, Zixuan Zhao, Tong Xia, Fucang Jia, Yuxuan Yang, Hao Wang, et al. Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. *Medical Image Analysis*, 86:102803, 2023.

[34] Chinedu Innocent Nwoye, Kareem Elgohary, Anvita Srinivas, Fauzan Zaid, Joël L Lavanchy, and Nicolas Padoy. Cholectrack20: A multi-perspective tracking dataset for surgical tools. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8942–8952, 2025.

[35] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In *International Conference on Learning Representations (ICLR)*, pp. 34433–34448, 2025.

[36] OpenAI. Gpt-5.6: System card. Official system card, 2026. Available at https://deploymentsafety.openai.com/gpt-5-6.

[37] Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, and Nicolas Padoy. Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. *International Journal of Computer Assisted Radiology and Surgery*, 16:1111–1119, 2021.

[38] Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 17456–17472, 2022.

[39] 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 (MICCAI)*, pp. 33–43, 2022.

[40] Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. Surgicalgpt: End-to-end language-vision GPT for visual question answering in surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 281–290, 2023.

[41] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8634–8652, 2023.

[42] Kumar Shridhar, Koustuv Sinha, Andrew Cohen, Tianlu Wang, Ping Yu, Ramakanth Pasunuru, Mrinmaya Sachan, Jason Weston, and Asli Celikyilmaz. The art of llm refinement: Ask, refine, and trust. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5872–5883, 2024.

[43] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

[44] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[45] SpaceXAI. Grok 4.6: Model card. Official model card, 2026. Available at https://media.x.ai/v1/website/card-4p6-4cd2dc57.pdf.

[46] Kimi Team. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

[47] Qwen Team. Qwen3.8-max: A new bar for coding and cowork. Official model release, August 2026. Available at https://qwen.ai/blog?id=qwen3.8.

[48] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5433–5442, 2023.

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

[50] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 74530–74543, 2023.

[51] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[52] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[53] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 306–316, 2024.

[54] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 122952–122983, 2024.

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

[56] Z.ai. Glm-4.5v. Official model card, 2025. Available at https://huggingface.co/zai-org/GLM-4.5V.

[57] Z.ai. Glm-5.3-flash. Official model card, 2026. Available at https://huggingface.co/zai-org/GLM-5.3-Flash.

[58] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 46595–46623, 2023.

A APPENDIX

Overview. This appendix provides additional implementation details, analysis of the review and correction rules, and dataset and evaluation information. It is organized as follows:

- Implementation and Routing Protocol (Section A.1).
- Review Gain and Correction Analysis (Section A.2).
- Dataset and Task Selection (Section A.3).
- Evaluation Protocol and Metrics (Section A.4).
- Additional Diagnostic Results (Section A.5).
- Inference and Review Procedures (Section A.6).

A.1 IMPLEMENTATION AND ROUTING PROTOCOL

Configuration and evidence access. The selected configuration uses tracker output modules v2.2 and identifies JPG as `five_head_probe`. In the diagnostic experiments, Gemini 3.8 Flash provides H0. Processing follows H0, prior-guided candidates and the cheap prediction, Qwen probing, gate selection, MAR on selected frames, and shared TC. The panel order is Qwen, GPT, Gemini, Grok, and DeepSeek, with Qwen's probe ratings reused as the first review. Candidate proposals allow two new labels per component task and four new triplets, with at most 64 interaction candidates plus seven phases. Reviewers receive H0 and the pool, excluding the Cheap Answer, tracker observations, other reviewers' ratings, and ground truth.

Candidate construction and preliminary correction. The preliminary state h_f^c is called the Cheap Answer in the implementation. It is formed before Qwen probing and remains the pre-gate prediction against which ER and BVR are measured. Priors use empirical triplet frequencies conditioned on the H0 phase; global frequencies replace phase buckets without eligible entries. Pool membership only makes a candidate available for review, rather than accepting it as a prediction. Null triplets are exempt from prior removal and addition, and h_f^c retains the H0 phase. Prior estimation follows the query-video exclusion in Section 3.2. Subsequent prior correction always uses that same phase bucket, so a reviewed phase cannot modify its own prior evidence.

Probe validation and feature encoding. Responses are checked for candidate identity and visual-evidence references. Invalid scores receive explicit sentinel values and validity summaries rather than being interpreted as support. The 54 features comprise 20 H0-structure, 12 prior/proposal, 10 interaction-probe, and 12 phase-probe features (*cf.* Table 7). Phase features include all seven ratings, their valid count, current-phase support, the strongest alternative, its margin, and a phase-settled indicator. Features are computed before additional review and exclude tracker observations and later reviewer scores. The three distinct input frames end at the target; no future observation enters H0 or review.

Table 7: Routing features grouped by evidence available before additional review.

Group	Count	Contents
H0 structure	20	Four label counts; seven phase indicators; interaction-empty flag; causal image count; null-triplet count; six component-missing/extra-label counts.
Prior/proposal	12	Veto/addition counts and change flag; five novel-candidate counts; three triplet-prior summaries; proposal-rule indicator.
Interaction probe	10	Non-ambiguous and invalid counts; support summaries for present/absent labels; settled fraction and all-settled flag.
Phase probe	12	Seven phase ratings; valid count; current-phase support; best alternative; alternative margin; phase-settled flag.

Router fitting and threshold selection. The router is fitted on fully annotated frames from the designated training split. It uses histogram-based gradient boosting, within Friedman's additive-tree framework [7], with 200 iterations, learning rate 0.05, at most 15 leaves per tree, and seed 3407. In Equation 5, $\ell_{\text{bin}}(b, u) = -b \log u - (1-b) \log(1-u)$, and a positive fraction ϖ gives class weights

$w_1 = 1/(2\varpi)$ and $w_0 = 1/[2(1-\varpi)]$. These weights balance class contributions; g_ϕ is used as a routing score rather than an assumed calibrated probability. Calibration of neural classifiers and elicited language-model confidence is studied separately [13, 48]. The `harm` training setting uses $b_f = \mathbb{I}[d_f > 0]$ from Equation 4: a review is positive only when it removes more errors than it introduces. The `overall` threshold-selection setting minimizes the number of reviewed validation frames subject to aggregate quality constraints:

$$\tau^* \in \arg\min_{\tau \in \mathcal{T}} \sum_{f \in \mathcal{D}_{\text{val}}} \mathbb{I}[g_\phi(z_f) \geq \tau], \quad \text{subject to} \quad E(\tau) \leq E_0, \qquad F(\tau) - F_0 \geq 0.9 \max(0, F_1 - F_0), \tag{11}$$

where $\mathcal{D}_{\text{val}}$ is the designated validation split, $\mathcal{T}$ is the candidate threshold set, $E(\tau)$ counts pooled false positives and false negatives, and $F(\tau)$ is the mean of the five dataset-level micro-F1 scores. Both metrics use final outputs after TC, including temporal phase filtering. Subscripts 0 and 1 denote skip review and always review under the same surrounding pipeline. The first constraint keeps total errors no higher than skip review; the second retains at least 90% of always review's positive mean-F1 gain. Among feasible thresholds, the one reviewing the fewest frames is selected. Candidate thresholds are score quantiles targeting review fractions from 5% to 100% in 5-point steps; ties can change the realized fraction. If none is feasible, the routine returns a skip-all fallback and records infeasibility. These validation constraints do not guarantee non-regression on every task or test sequence.

Gate fitting and threshold selection follow the training/validation/test split in Section A.4. The fitted classifier and validation-selected threshold are frozen before test evaluation. Test annotations do not enter gate fitting or threshold selection, and candidate construction and correction rules remain fixed across the compared routing policies.

Interaction selection and invalid evidence. Complete-panel means require a valid rating from every reviewer. An invalid or missing rating makes the corresponding mean unavailable and its admission predicate false; intermediate or unavailable scores preserve existing membership. Score-based interaction selection starts from H0. A component referenced by a surviving triplet is protected against removal. JPG prior veto/addition is applied after score-based selection as a distinct admission path. Finally, the <*Grasp*> and <*Retract*> verbs and all associated triplets are restored to their memberships in h_f^c. Thus, ambiguity rollback covers both verbs and relations, without reconstructing every component head from the triplet set. Phase replacement requires valid means for all seven alternatives, a unique maximum, support at least 4, and strict superiority over H0; otherwise H0 is retained. Early-stop tests exclude interaction memberships fixed by rollback, and an invalid observed rating cannot authorize an edit through a later completed aggregate.

Tracker alignment and label cleanup. A tracker snapshot is valid only when its video, target index, and final observed frame match the current target. Classes from that final frame are thresholded at confidence 0.35 and filtered to the triplet instrument ontology. If the snapshot is invalid or unavailable, or its class set exceeds output capacity, TC retains the ontology-filtered model state. A valid empty class set instead replaces the instrument head with an empty set and filters triplets accordingly. After successful fusion, the instrument–verb rules are <*Clipper*>→<*Clip*>, <*Hook*>→<*Dissect*>, and <*Scissors*>→<*Cut*>; the verb update is skipped if it exceeds capacity. TC removes null triplets validly rated 1 by the existing probe and removes isolated null-verb/null-target labels when no null triplet remains. Raw selected phases are appended to video-specific history after selection, before voting; the current raw phase participates in its own vote. If the phase vote is tied, TC retains the current raw phase even when it is not a mode. These corrections use available evidence without another model request.

Relation to multimodal modeling and review. General visual instruction tuning and frozen-encoder interfaces connect visual representations to language generation [20, 23]; biomedical assistants adapt this capability through instruction tuning or few-shot medical examples [17, 29]. These works concern the underlying model interface. Our intervention begins after a backbone has produced a structured surgical state and governs which additional evidence is acquired and which labels are revised.

Adaptive computation can change recurrent computation depth or exit language-model layers early [12, 38]; test-time scaling studies how to allocate extra reasoning effort across inputs [43]. Our stopping rule instead operates over a fixed ordered reviewer panel: it skips only requests whose

possible scores cannot change the prescribed decision. Self-consistency and model ensembling aggregate multiple candidate outputs [15, 51]. Verification chains and sample-consistency checks target unsupported statements [5, 28]. We use these as conceptual comparisons, while the reported pipeline evaluates explicit surgical candidates and applies the deterministic admission, rollback, and temporal rules specified above.

A.2 Review Gain and Correction Analysis

Net review gain as corrected versus introduced errors. Equation 4 has a direct interpretation in terms of the edits made by review. For task q, let $\mathcal{A}_{f,q} = \bar{h}^{\mathrm{rev}}_{f,q} \setminus \bar{h}^{\mathrm{skip}}_{f,q}$ contain added labels and $\mathcal{D}_{f,q} = \bar{h}^{\mathrm{skip}}_{f,q} \setminus \bar{h}^{\mathrm{rev}}_{f,q}$ contain deleted labels. Then

$$\begin{aligned} C_f &= \sum_{q \in \mathcal{Q}} \left(|\mathcal{A}_{f,q} \cap y_{f,q}| + |\mathcal{D}_{f,q} \setminus y_{f,q}| \right), \\ H_f &= \sum_{q \in \mathcal{Q}} \left(|\mathcal{A}_{f,q} \setminus y_{f,q}| + |\mathcal{D}_{f,q} \cap y_{f,q}| \right), \\ d_f &= \underbrace{C_f}_{\text{corrected errors}} - \underbrace{H_f}_{\text{introduced errors}}, \qquad b_f = \mathbb{I}[C_f > H_f], \end{aligned} \tag{12}$$

where C_f counts true labels added and false labels deleted, while H_f counts false labels added and true labels deleted. Labels whose membership is unchanged contribute equally to the two symmetric differences in Equation 4 and cancel. Every remaining membership change either corrects one error or introduces one, giving the identity above. The target therefore rewards net correction rather than the number of edits alone. It can be positive despite a regression in an individual task. This decomposition explains the implemented supervision; it introduces no additional reward model or training stage.

Why early stopping preserves the selector. The unobserved score sum lies between $J - k$ and $5(J - k)$, yielding Equation 8 for every completion of the panel. The runtime stops only when the bounds certify that every admissible completion of the remaining ratings yields h^{c}_f under the complete selector, including component protection, the prior pass, ambiguity rollback, and phase admission. If this equality cannot be certified, MAR completes the panel. Invalid evidence prevents the corresponding aggregate from authorizing an edit; ambiguity rollback and prior-based rules are independent of the remaining scores. Consequently, additional responses cannot change the prescribed output, although the output itself can be incorrect.

Correction boundaries. Candidate closure and admission checks constrain particular edits; the implementation does not globally reconstruct every component task from the final triplet set. Tracker filtering, ambiguity rollback, and independently retained components can leave extra component predictions. Phase voting uses raw selected phases in a finite past window, not recursively filtered outputs, but can still delay a true transition. These boundaries distinguish controlled updates from a guarantee of globally consistent or error-free predictions.

Error-count interpretation. The gate target counts label errors rather than task-level correctness. For a singleton phase, replacing an incorrect phase with the correct one removes one false positive and one false negative. A positive net gain may therefore coexist with a regression in another head; it does not impose per-task non-regression. The common post-review instrument and label corrections are applied before constructing this target, whereas temporal phase filtering is included only when selecting and evaluating the routed sequence.

A.3 Dataset and Task Selection

Benchmark provenance and composition. SurAgent-Bench [21] combines Cholec80 [49], CholecT50 [33], and CholecTrack20 [34] with a private clinical cohort. Source videos are downsampled from 25 fps to 1 fps. Its 34,499 frames comprise 19,499 public and 15,000 private frames, with eight tasks per frame (275,992 question–answer pairs; *cf.* Table 8). The project is available at https://github.com/GeorgeHuLite/SurAgent.

Private clinical cohort. Curated with partner hospitals, the private cohort comprises raw endoscopic surgery videos from 100 patients: 69 male and 31 female, aged 22–91 years (predominantly 50–69).

Table 8: Composition of the complete SurAgent-Bench benchmark [21]. These are benchmark-level counts, not partition-specific evaluation counts.

Source	Annotated frames
Public: Cholec80, CholecT50, and CholecTrack20	19,499
Private: clinical cohort of 100 patients	15,000
Total	34,499

After downsampling from 25 fps to 1 fps, the extracted frames were annotated by clinical experts. The source manuscript reports anonymization of out-of-body frames and restricted access to the clinical videos.

Annotation and reference-answer construction. The source benchmark reports that visual annotations and their task dependencies were annotated and verified by clinical experts. Questions for visual understanding are generated from templates, with answers derived from the annotated labels. For open description and surgical report generation, GLM-4.5V [56] produces reference answers conditioned on both the image and ground-truth visual annotations. These outputs are then organized into observations, clinical significance, and next procedural steps. Thus, the benchmark combines expert-verified visual labels with model-assisted language references. These construction details are inherited from SurAgent-Bench.

Task selection and representativeness. We select six representative tasks from SurAgent-Bench: five visual understanding tasks and one language-generation task. Instrument, verb, and target recognition assess the entities and actions underlying surgical perception. Triplet recognition evaluates their composition into instrument–tissue interactions, while phase recognition captures the procedural context. Surgical report generation assesses the use of these predictions in structured language output. Together, these tasks cover the progression from local visual evidence to interaction and workflow understanding and downstream reporting, supporting evaluation of both recognition quality and its implications for generated reports. Instrument grounding and open description are outside the reported evaluation.

A.4 Evaluation Protocol and Metrics

Study-specific split and temporal inputs. The study-specific training/validation/test split follows a 10:2:8 ratio by total frame count (50%/10%/40%). The implementation uses label spaces of 7 instruments, 10 verbs, 15 targets, 100 triplets, and 7 phases. The three-frame temporal input follows Equation 1; tracker and phase histories are reset at each video boundary. The input stride uses original source-frame indices, while the 60-second history uses timestamps rather than a fixed count of retained targets.

Model inputs and diagnostic backbone. The general-purpose backbones receive the three-frame context defined in Equation 1 and return the same structured recognition schema. For each backbone, the standalone row reports its joint first-pass prediction, while the corresponding SurgRoute row applies the complete pipeline to the same inputs and ontology. HecVL, PeskaVLP, and SurgVLP use their authors' released checkpoints pretrained on large-scale surgical datasets. All model rows (*cf.* Table 1), including surgical models, standalone general-purpose backbones, and SurgRoute variants, use the same downstream report generator and evaluation protocol. Unless otherwise stated, diagnostics use Gemini 3.8 Flash as the common backbone.

Recognition metrics and valid annotations. Following standard distinctions between classification measures [44], scores for task q pool label counts over its valid annotation set $\mathcal{V}_q$. Let $\widehat{Y}_{f,q}$ and $Y_{f,q}$ be the predicted and reference label sets. We compute $\mathrm{TP}_q = \sum_{f\in\mathcal{V}_q} |\widehat{Y}_{f,q} \cap Y_{f,q}|$, $\mathrm{FP}_q = \sum_{f\in\mathcal{V}_q} |\widehat{Y}_{f,q} \setminus Y_{f,q}|$, and $\mathrm{FN}_q = \sum_{f\in\mathcal{V}_q} |Y_{f,q} \setminus \widehat{Y}_{f,q}|$. The reported percentages are

$$\begin{aligned} \mathrm{Prec}_q &= 100\frac{\mathrm{TP}_q}{\mathrm{TP}_q + \mathrm{FP}_q}, \qquad \mathrm{F1}_q = 100\frac{2\mathrm{TP}_q}{2\mathrm{TP}_q + \mathrm{FP}_q + \mathrm{FN}_q}, \\ \mathrm{Acc}_q &= \frac{100}{|\mathcal{V}_q|}\sum_{f\in\mathcal{V}_q} \mathbb{I}[\widehat{Y}_{f,q} = Y_{f,q}]. \end{aligned} \tag{13}$$

Precision and F1 use pooled counts rather than averages of per-frame scores. Exact-set accuracy requires every predicted label to match the reference set. Valid null labels are included; a missing annotation is masked rather than interpreted as an empty set. A zero precision or F1 denominator yields zero, and a task with no valid annotations is unscored. Phase accuracy is the fraction of valid targets with the correct single phase. Consequently, micro F1 can increase while exact-set accuracy decreases, and F1 may exceed precision when recall is higher.

Report generation and scoring. The implemented quantitative reporting interface generates a short event summary from a canonical structured prediction. The generator receives predicted interactions and phase, without reference annotations, and returns one JSON `report` field containing at most 60 words. This structured-state-conditioned task assesses how recognition outputs support reporting; the three-part reports in the qualitative comparison illustrate broader narrative behavior and do not define this scoring format.

The rule-based evaluator extracts directed instrument–action–target relations from the generated text and compares them with the annotated relations. Its score is

$$\text{Rule} = 0.60S_{\text{int}} + 0.20S_{\text{phase}} + 0.15S_{\text{content}} + 0.05S_{\text{format}}, \tag{14}$$

where S_{int} is the extracted-relation F1 on a 0–100 scale. Unresolved action claims count as false positives. When both relation sets are empty, full interaction credit requires an explicit statement that no interaction is identified. The other terms are binary 0/100 checks: S_{phase} requires the extracted phase set to equal the annotated singleton; S_{content} requires nonempty text without unsupported ontology terms or terms matched by the evaluator's fixed risk-term list; and S_{format} requires the prescribed JSON schema and length. The content check is a closed-vocabulary heuristic, not a clinical-safety assessment.

Model-based evaluation has been studied through preference judging and rubric-based scoring [24, 58]. Here, the model-based judge compares the candidate report against ground-truth interactions and phase, without images or access to the prediction source. It scores factual consistency S_{factual}, clarity S_{clarity}, and coherence $S_{\text{coherence}}$ on a 0–100 scale:

$$\begin{aligned} \text{Judge} &= 0.50S_{\text{factual}} + 0.25S_{\text{clarity}} + 0.25S_{\text{coherence}}, \\ \text{Total} &= 0.8\,\text{Rule} + 0.2\,\text{Judge}. \end{aligned} \tag{15}$$

The evaluator recomputes the weighted scores from their components. The report generator, judge version, prompts, scoring rules, and eligible evaluation frames are shared across all model rows (*cf.* Table 1). Thus, reporting comparisons vary the predicted structured state while holding generation and assessment fixed. Scores are averaged over the common eligible frames; missing responses are not assigned zero or silently omitted. Total is computed from unrounded scores before display rounding.

Bundled component ablations. The component ablations switch complete functional bundles (*cf.* Table 2). Disabling JPG removes prior-guided candidate expansion, preliminary correction, and selective routing: MAR reviews H0 candidates on every frame, retaining its repair rules and early stopping. This differs from the always-review policy (*cf.* Fig. 5), which retains candidate construction, preliminary correction, and joint probing. Disabling TC jointly removes instrument fusion, label correction (instrument–verb rules and probe-based null cleanup), and causal phase filtering; its effect cannot be attributed to tracking alone. H0 applies none of these refinement bundles. MAR is assessed through its internal admission and stopping rules (*cf.* Table 5); the component ablations do not include a separate JPG+TC configuration with MAR disabled (*cf.* Table 2). H0 alone is not the skip-review policy, which retains JPG's candidate construction, preliminary correction, and probe.

Review policies and computation. Skip review fixes $a_f = 0$, always review fixes $a_f = 1$, and JPG applies Equation 5. All three retain candidate construction, preliminary correction, joint probing, and common TC; reviewed frames use the same MAR rules and exact stopping. Skip review therefore retains the fixed cost of evidence acquisition. The policy comparison reports total tokens in millions (*cf.* Fig. 5 and Table 10). Token differences describe volume within this comparison, not billing cost, latency, or request count. In contrast, if K_f counts additional reviewer requests and M is the number of evaluated frames, the recognition pipeline uses $3 + M^{-1}\sum_f K_f$ logical requests per frame; H0 alone uses one. This identity excludes downstream report generation and cannot recover request counts from token totals. Accordingly, JPG's 31.2% token saving and exact

stopping's 28.8% saving in additional calls use different reference budgets. They should be reported separately rather than added to estimate a combined reduction.

Review fraction, error recall, and BVR. Review fraction is $100 \sum_f a_f / M$ over all M evaluated frames. ER and BVR instead use scorable frames and valid task annotations. Let $\mathcal{Q}_f$ contain the valid tasks for frame f, and define the pre-gate cheap and post-MAR error counts as

$$\epsilon_f^{\mathrm{c}} = \sum_{q \in \mathcal{Q}_f} |h_{f,q}^{\mathrm{c}} \triangle y_{f,q}|, \qquad \epsilon_f^{\mathrm{r}} = \sum_{q \in \mathcal{Q}_f} |h_{f,q}^{\mathrm{r}} \triangle y_{f,q}|. \tag{16}$$

Both states precede TC. Valid-task masks are fixed across policies and before/after comparisons. Let $\mathcal{S}_{\mathrm{c}}$ contain frames with a scorable cheap prediction and $\mathcal{S}_{\mathrm{cr}} \subseteq \mathcal{S}_{\mathrm{c}}$ contain frames scorable before and after MAR. Then

$$\begin{aligned} \mathrm{ER} &= 100 \frac{\sum_{f \in \mathcal{S}_{\mathrm{c}}} a_f \mathbb{I}[\epsilon_f^{\mathrm{c}} > 0]}{\sum_{f \in \mathcal{S}_{\mathrm{c}}} \mathbb{I}[\epsilon_f^{\mathrm{c}} > 0]}, \\ \mathrm{BVR} &= 100 \frac{\sum_{f \in \mathcal{S}_{\mathrm{cr}}} a_f \mathbb{I}[\epsilon_f^{\mathrm{r}} < \epsilon_f^{\mathrm{c}}]}{\sum_{f \in \mathcal{S}_{\mathrm{cr}}} a_f}. \end{aligned} \tag{17}$$

ER is determined immediately after gate selection, before MAR: it asks whether erroneous cheap predictions were selected. BVR is measured after MAR and before TC: it asks whether review reduced the total valid-task error count. Partial net corrections count as beneficial; unchanged counts do not. BVR is undefined when no scorable frames are reviewed, including skip review. Neither metric credits tracker fusion, label cleanup, or temporal phase filtering. Final task scores in the same table are measured after TC.

These diagnostics differ from the gate-training target and probe gain metrics in Equation 4. Those compare skip/full-review branches after shared instrument and label correction but before temporal phase filtering; BVR compares the cheap and MAR states before any TC. Consequently, gain precision (*cf.* Table 4) and BVR (*cf.* Table 3) use different outcome definitions and are not numerically interchangeable.

A.5 Additional Diagnostic Results

Table 9: **Diagnostic controls and measurement stages.** Recognition scores use final outputs; gain supervision and repair diagnostics use the stages specified below.

Study	Varied	Fixed	Measurement
Components	JPG/MAR/TC bundles	H0 backbone and task ontology	Final recognition
Review policies	Skip/always/JPG	Candidate pool, probe, MAR rules, TC	ER at gate; BVR before TC; final quality/tokens
Probe features	Available feature groups	Router settings; 50% validation review target	Branch gain before phase filtering
MAR rules	Admission, rollback, stopping	Routed frames, candidates, ordered reviewer responses	Harm/agreement before TC; final F1
TC controls	Label rules; phase-history source	Pre-TC states, tracker observations, routing	Final recognition; 5-second transition accuracy

Policy-level task scores. The detailed results include phase scores for the three evaluated policies (*cf.* Table 10 and Fig. 5). JPG approaches always-review triplet performance at lower token consumption, while skip review has the highest phase accuracy. The figure additionally includes the supplied simulated Random, Uncertainty-only, and Rule-based Gate settings. Their triplet F1 scores are 44.79, 45.41, and 45.67 at 494.10, 518.98, and 460.46 million tokens, respectively. These points are displayed separately from the three evaluated policies and do not enter the ER/BVR comparison (*cf.* Table 3).

Probe-feature controls and review gain. The probe ablations compare the 32-feature H0/prior baseline with interaction-only (42 features), phase-only (44 features), and joint-probe (54 features)

Table 10: **Recognition scores and total token volume under the three review policies.** Task scores are percentages. Token Consumption is reported in millions and rounded to two decimals. It quantify the reported token volume rather than latency or monetary cost.

	Triplet recognition			Phase	Tokens Consumption
Policy	Acc.↑	Prec.↑	F1↑	Acc.↑	(M)↓
Skip review	20.32	54.06	46.21	87.63	331.79
Always review	20.60	54.50	46.43	86.86	653.74
JPG	20.55	54.44	46.40	87.45	450.07

variants (*cf.* Table 4). Classifiers share the designated data partitions, fully annotated training frames, fitting settings, and target b_f in Equation 4. A validation threshold targets 50% review and is frozen for testing; tied scores are retained, so the realized test fraction can differ. This common validation budget is separate from the default 32.93% test operating point. Gain recall is $100 \sum_f a_f b_f / \sum_f b_f$, gain precision is $100 \sum_f a_f b_f / \sum_f a_f$, and net gain is $100 \sum_f a_f d_f / M$, in label errors removed per 100 frames. Both branches receive common instrument/label correction before phase aggregation, unlike pre-TC BVR. The full probe improves gain recall from 72.06% to 89.41% and gain precision from 48.42% to 61.41%; either probe group alone yields intermediate gains. Net gain rises from 37.10 to 59.20 label errors removed per 100 frames.

MAR rules and exact stopping. The MAR ablations fix routing, candidates, and complete five-reviewer responses (*cf.* Table 5). Triplet-only admission replaces $A_f(r)$ by $\mathbb{I}[\mu_f(\mathrm{R}, r) \geq 4]$ only for score-based admission; ontology checks, component updates, procedural priors, ambiguity rollback, and phase rules remain active. The rollback ablation changes only restoration of *<Grasp>*/*<Retract>* memberships. Complete-panel variants use all four additional reviewers after the reused probe; exact stopping evaluates the same ordered responses until the decision is determined. Agreement requires exact equality with full-panel MAR across all five tasks before TC. Exact stopping achieves 100.00% agreement and reduces additional requests per routed frame from 4.00 to 2.85, a 28.8% reduction. This saving applies to additional reviewers on the fixed routed subset, not total pipeline tokens. Removing component support or ambiguity rollback reduces triplet F1 from 46.40 to 44.92 or 45.37, respectively, while increasing introduced errors. Phase accuracy is unchanged because these controls leave phase selection fixed.

Introduced label errors. For MAR variant v, let $\mathcal{A}^v_{f,q}$ and $\mathcal{D}^v_{f,q}$ be labels added to and removed from $h^{\mathrm{c}}_{f,q}$ before TC. The introduced-error count is $H^v_f = \sum_q (|\mathcal{A}^v_{f,q} \setminus y_{f,q}| + |\mathcal{D}^v_{f,q} \cap y_{f,q}|)$. Harm reports $100 \sum_f H^v_f / M_{\mathrm{routed}}$ over the common routed subset; recognition scores use the complete evaluated sequence after the same TC. This is a label-error count per 100 routed frames, not the percentage of fields changed from an exactly correct set to an incorrect set.

TC controls and transition accuracy. The TC ablations fix selected pre-TC states, tracker snapshots, probe responses, and routing (*cf.* Table 6). Removing label rules disables instrument–verb additions and probe-based null cleanup while retaining instrument replacement and triplet filtering. This lowers triplet F1 from 46.40 to 46.18 without changing instrument or phase scores. The recursive-history control stores filtered phases instead of raw selected phases, with the same 60-second window and tie rule. Histories reset at video boundaries and never use future frames. Raw history raises overall phase accuracy from 79.86% to 87.45% and transition accuracy from 18.20% to 27.60% relative to recursive history. Transition accuracy pools targets in $[t, t+5\,\mathrm{s})$ after annotated phase changes at time t, counting overlapping windows once. Transition accuracy remains low, indicating limited accuracy around phase changes even when recursive feedback is avoided. The history comparison changes feedback construction while preserving tracker inputs, visual evidence, and review allocation.

A.6 Inference and Review Procedures

Algorithms 1 and 2 summarize inference with a frozen router and threshold. Candidate construction, validation, and output capacities follow Section A.1; no reference annotations are inputs to either procedure.

Algorithm 1 Chronological JPG–MAR inference with shared TC

Require: Videos with eligible target indices; ontology Ω; frozen tracker
Require: Base model f_θ, priors ρ_{-v}, router g_ϕ, threshold τ
Ensure: Corrected states h_f for all processed targets

for each video v **do**
  Reset tracker state and raw-phase history $\mathcal{H} \leftarrow \varnothing$
  for each target f in chronological order **do**
    Form causal context W_f using Equation 1
    $h_f^0 \leftarrow f_\theta(W_f; \Omega)$
    Fix ρ_f to the H0 phase bucket (global fallback if empty)
    Obtain prior-conditioned proposals with the base model
    Build $\mathcal{C}_f$ from H0, proposals, and triplet components
    $h_f^{\mathrm{c}} \leftarrow$ prior correction of H0 (Equation 3)
    $s_{f,1} \leftarrow$ validated Qwen ratings of $\mathcal{C}_f$ and all phases
    $z_f \leftarrow$ the 54 H0, prior/proposal, and probe features
    $a_f \leftarrow \mathbb{I}[g_\phi(z_f) \geq \tau]$; $\widetilde{h}_f \leftarrow h_f^{\mathrm{c}}$
    if $a_f = 1$ **then**
      $\widetilde{h}_f \leftarrow$ Algorithm 2$(W_f, h_f^0, h_f^{\mathrm{c}}, \mathcal{C}_f, \rho_f, s_{f,1})$
    end if
    $h_f \leftarrow$ ontology-filtered $\widetilde{h}_f$; $\widetilde{p}_f \leftarrow$ phase of $\widetilde{h}_f$
    Read tracker snapshot for video v and target f
    If available, extract $\mathcal{D}_f$ at confidence ≥ 0.35 in the ontology
    if tracker evidence is valid and $\mathcal{D}_f$ is within capacity **then**
      Replace instruments and filter triplets (Equation 9)
      Apply instrument–verb rules if within capacity (Section A.1)
    end if
    Apply probe-based null-label cleanup (Section A.1)
    Append $(f, \widetilde{p}_f)$ to $\mathcal{H}$ before temporal aggregation
    Keep only entries with $f - 25 \cdot 60 < j \leq f$ in $\mathcal{H}$
    Count raw-phase votes n_f and modes $\mathcal{M}_f$ (Equation 10)
    $h_{f,\mathrm{P}} \leftarrow$ the unique mode if one exists; otherwise $\widetilde{p}_f$
    Emit h_f; retain only raw phases as history
  end for
end for

The phase window uses original frame indices at 25 fps, equivalently elapsed timestamps, rather than a fixed number of retained targets. An aligned, valid empty tracker class set still passes the fusion condition. Instrument–verb rules run after successful fusion; null-label cleanup reuses the probe. The corrected interactions and phase condition the shared report generator.

Algorithm 2 Ordered MAR with probe reuse and conservative stopping

Require: $W_f, h_f^0, h_f^{\mathrm{c}}, \mathcal{C}_f, \rho_f$, and validated probe $s_{f,1}$
Ensure: Reviewed state h_f^{r}, or h_f^{c} on certified early stopping

$J \leftarrow 5$; $k \leftarrow 1$; reuse Qwen as reviewer 1
Remaining order: GPT, Gemini, Grok, DeepSeek
while $k < J$ **do**
  $\mathcal{V}_k \leftarrow$ interaction/phase candidates with all k ratings valid
  Compute $L_{f,k}, U_{f,k}$ on $\mathcal{V}_k$ (Equation 8)
  Mark means outside $\mathcal{V}_k$ unavailable
  if CertCheap$(L_{f,k}, U_{f,k}, \mathcal{V}_k; h_f^0, h_f^{\mathrm{c}}, \mathcal{C}_f, \rho_f)$ **then**
    return h_f^{c}
  end if
  $k \leftarrow k + 1$
  Request reviewer k ratings with the same $(W_f, h_f^0, \mathcal{C}_f)$
  Validate ratings of interactions and all seven phases
end while
for each candidate c in $\mathcal{C}_f \cup (\{\mathrm{P}\} \times \mathcal{P})$ **do**
  if all J ratings of c are valid **then**
    $\mu_f(c) \leftarrow J^{-1} \sum_{j=1}^{J} s_{f,j}(c)$
  else
    $\mu_f(c) \leftarrow \perp$ {cannot authorize an edit}
  end if
end for
Compute $A_f(r)$ using all four means in $\Gamma(r)$ (Equation 6)
$h_f^{\mathrm{r}} \leftarrow$ ScoreRepair$(h_f^0, \mathcal{C}_f, \mu_f, A_f)$
Apply the prior pass with unchanged ρ_f (Section A.1)
Restore Grasp/Retract verbs and triplets to membership in h_f^{c}
$p_f^{\mathrm{r}} \leftarrow p_f^0$
if all seven phase means are valid **then**
  $\mathcal{B}_f \leftarrow \arg\max_{p \in \mathcal{P}} \mu_f(\mathrm{P}, p)$
  if $\mathcal{B}_f = \{p^*\}$, $\mu_f(\mathrm{P}, p^*) \geq 4$, and $\mu_f(\mathrm{P}, p^*) > \mu_f(\mathrm{P}, p_f^0)$ **then**
    $p_f^{\mathrm{r}} \leftarrow p^*$
  end if
end if
Set the phase of h_f^{r} to p_f^{r}; **return** h_f^{r}

Selection and stopping. ScoreRepair denotes the score-based interaction selector in Sections 3.3 and A.1: novel triplets require $A_f(r)$; component additions/removals use valid means at least 4/at most 2, with protection for components of surviving triplets. Intermediate or unavailable means preserve membership. Prior correction and ambiguity rollback follow this selector; component heads are not reconstructed globally.

CertCheap checks the complete selection sequence, including component protection, the H0-conditioned prior pass, rollback, and phase admission. It returns true only when every admissible completion yields h_f^{c}. An invalid observed rating keeps its mean unavailable; possible later invalid ratings are included in the certificate. Memberships fixed by rollback need no independent edit test, but still affect triplet protection. Unresolved dependencies continue review, even if a non-cheap outcome already appears inevitable. Thus, the procedure preserves the conservative stopping rule rather than replacing coupled selection with independent score thresholds.