SURGCOUPLE: COUPLING GLOBAL CONTEXT WITH LOCAL CUES VIA CHAIN OF SPATIO-TEMPORAL EVIDENCE FOR SURGICAL VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Similar surgical phases and instrument configurations can involve distinct instrument–tissue interactions, so global context alone cannot discriminate local events. Existing multimodal LLMs, however, lack a mechanism for selectively attending to sparse, spatial-temporal cues. We introduce **SurgCouple**, the first surgical video reasoning framework to **couple global context with hypothesis-guided spatial-temporal evidence discovery**. An Evidence Chain progressively discovers temporal, spatial, and fine-grained evidence, with intermediate reasoning states providing hypothesis-guided feedback to the latter two stages to focus on the informative regions. A Reasoning Chain maintains an evolving interpretation of the surgical event, refining the hypothesis space as evidence accumulates. A Consistency Verifier ensures that local judgments remain grounded in the evidence actually selected and coherent across stages. With a frozen Qwen3-VL-8B, SurgCouple achieves 89.31% final-answer accuracy and a 58.92% valid chain rate, surpassing 81.94% and 34.01% for Qwen3-VL-8B, and 88.92% and 52.63% for GPT-5.5.

1 INTRODUCTION

Surgical videos encode rich procedural knowledge that supports intraoperative decision-making, surgical training, and postoperative outcome analysis (Hu et al., 2024; Liu et al., 2025). Realizing these applications requires not only recognizing the global procedural context but also identifying the local evidence that substantiates each clinical judgment, as the video interpretation depends on localized spatiotemporal cues, while similar procedural contexts can involve different instrument–tissue interactions and local tissue changes (Wang et al., 2026b). However, existing Multimodal Large Language Models (MLLMs) lack an explicit mechanism for selectively attending to the sparse, decisive local cues, and hence cannot ensure that generated explanations are grounded in the relevant visual evidence (Zhang et al., 2024; Guan et al., 2024; Chen et al., 2024b).

MLLMs commonly employ two strategies: 1) task-oriented fine-tuning, which optimizes models on surgical VQA datasets to improve task-specific predictions (Wang et al., 2026a; Hao et al., 2025); and 2) prompt-guided reasoning, which elicits structured CoT through instructions or in-context examples (Man et al., 2025; Dong et al., 2025). Although these strategies can improve answer generation, answer supervision and reasoning instructions alone do not ensure that intermediate conclusions are grounded in localized visual evidence (Guan et al., 2024; Chen et al., 2024b).

Bridging global procedural context and local visual evidence presents three related challenges (Fig. 1, Left). First, similar surgical phases can involve different instrument–tissue interactions, so global context alone cannot disambiguate them without localized spatiotemporal cues (Wang et al., 2026b). Second, decisive cues are often sparse and fleeting, and interpreting them requires conditioning on the current global hypothesis; yet the discovered evidence must in turn be allowed to confirm or revise that hypothesis (Lin et al., 2023). Third, locating relevant evidence does not guarantee its support for the hypothesis. Hallucination and grounding failures in MLLMs illustrate this gap (Chen et al., 2024b), and maintaining coherence between successive judgments requires explicit verification.

To tackle these, we introduce **SurgCouple**, which couples global context with local visual cues for evidence-grounded surgical reasoning. Unlike prior methods that treat localized evidence as a fixed input, SurgCouple **bidirectionally couples** evidence discovery with reasoning: a hierarchical

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.

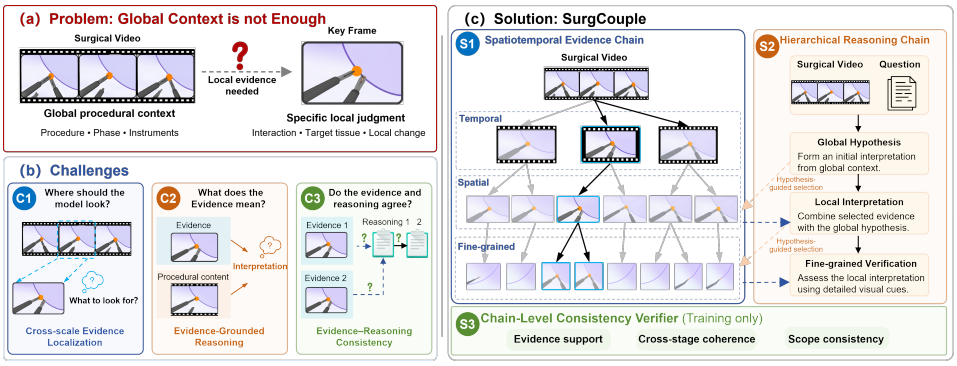


Figure 1: **(a) Problem:** Global context alone cannot resolve local events. **(b) Challenges:** cross-scale evidence localization, evidence-grounded reasoning, and cross-stage consistency. **(c) SurgCouple:** The three modules jointly refine hypotheses and evidence to ground local judgments.

Reasoning Chain (RC) forms an initial hypothesis from the global video context and refines it into evidence-conditioned judgments; a spatiotemporal **Evidence Chain (EC)** discovers, under the guidance of each current hypothesis, the increasingly localized evidence that the refined judgments require; and a chain-level **Consistency Verifier (CV)** regularizes their alignment. Each judgment beyond the initial global one is thus both plausible and consistent with the observed evidence.

The Evidence Chain enables **dynamic, hypothesis-dependent** evidence discovery by progressively injecting reasoning states into evidence selection, unlike prior methods that select evidence from the question alone across all stages (Zhang et al., 2025). Temporal windows are first identified from the question and dataset-provided knowledge, defining the scope for later refinement. This initial selection is independent of the global hypothesis. Spatial localization scores candidate regions via gated cross-attention, where a per-candidate gate conditioned on both the candidate and the reasoning state controls how much hypothesis-conditioned information each candidate contributes. The same gating injects the global hypothesis and the localized interpretation at the fine-grained stage.

Unlike prior methods that directly aggregate evidence into judgments, the Reasoning Chain forms a **progressively refined hierarchical cascade** in which each subsequent stage is conditioned on both reasoning-guided evidence and preceding reasoning states. It proceeds from a global hypothesis formed over the global video representation and question context, to a local interpretation refined from the selected spatiotemporal evidence, and finally to a verification against fine-grained cues. Later stages modulate incoming evidence by preceding states through the same gating mechanism, carrying global procedural context forward while incorporating progressively finer detail. Each stage is supervised to produce an answer prediction, so that every intermediate state remains answer-relevant.

Lastly, a Consistency Verifier performs **explicit chain-level verification** over the reasoning and evidence streams and scores chain-level consistency along three aspects: *evidence grounding*, whether the evidence-conditioned states are supported by the evidence; *cross-step coherence*, whether adjacent reasoning states are logically compatible; and *progressive scope consistency*, whether spatial regions lie within the selected temporal clips and fine-grained cues within those regions.

Our contributions are threefold: 1) an Evidence Chain that enables **dynamic, hypothesis-dependent evidence discovery** by progressively injecting reasoning states into selection; 2) a Reasoning Chain that forms a **progressively refined hierarchical cascade**, with each stage conditioned on reasoning-guided evidence and preceding states under stage-wise answer supervision; and 3) a Consistency Verifier that provides **explicit chain-level verification** of evidence grounding, cross-step coherence, and scope consistency, offering auxiliary supervision beyond the answer loss. Experiments on SurgCoT show improvements in final answer accuracy, valid chain rate, and evidence quality.

2 Related Work

Surgical Video Understanding and Multimodal Reasoning. Research in surgical video understanding has progressed from frame and clip recognition to visual question answering (VQA) and

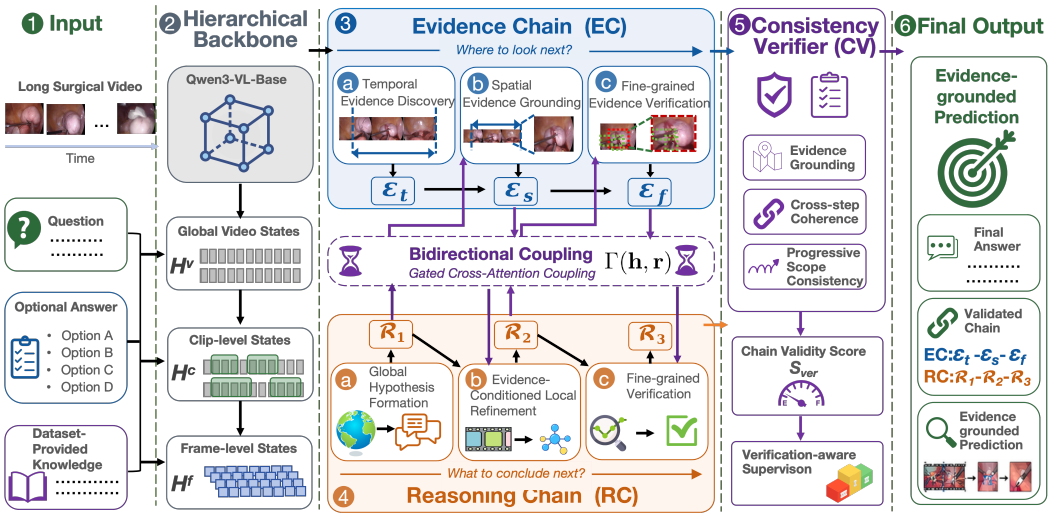


Figure 2: **SurgCouple** couples global context and local evidence bidirectionally. 1) **Reasoning Chain** forms an initial hypothesis from the global context and refines subsequent hypotheses into evidence-conditioned judgments; 2) **Evidence Chain** progressively localizes temporal, spatial, and fine-grained evidence under the guidance of each refined hypothesis; 3) **Consistency Verifier** scores evidence–state alignment, cross-step coherence, and scope consistency for auxiliary supervision.

reasoning. Early work established surgical phase recognition via EndoNet (Twinanda et al., 2017) and temporal modeling through TeCNo (Czempiel et al., 2020), Trans-SVNet (Gao et al., 2021), and LoViT (Liu et al., 2025); Rendezvous (Nwoye et al., 2022) further modeled instrument–verb–target relationships via surgical triplet recognition. Vision-language methods extended these capabilities to surgical question answering through Surgical-VQA (Seenivasan et al., 2022), language-model-based VQA via SurgicalGPT (Seenivasan et al., 2023), and joint answering with spatial localization through Surgical-VQLA (Bai et al., 2023); general biomedical models such as LLaVA-Med (Li et al., 2023a) have also supported medical VQA and instruction following.

Chain-of-thought (CoT) prompting introduced intermediate reasoning steps for language tasks (Wei et al., 2022), with variants like Self-Consistency (Wang et al., 2023) and Tree of Thoughts (Yao et al., 2023) exploring multiple reasoning paths. In multimodal learning, Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b), and LLaVA (Liu et al., 2023) connect visual representations with language models, and extensions such as Multimodal-CoT (Zhang et al., 2024), Video-LLaVA (Lin et al., 2024), and InternVL (Chen et al., 2024c) incorporate visual information into reasoning and instruction following. CoT methods produce free-form, unsupervised intermediate steps, with visual evidence consumed once and isolated from the evolving rationale; how to maintain structured, supervised reasoning states that actively guide evidence selection remains underexplored.

Medical Evidence Grounding. Temporal grounding localizes query-relevant intervals in videos. Early methods (Gao et al., 2017; Zhang et al., 2020) localize query-relevant intervals through cross-modal matching, yet lack fine-grained interactions between visual and textual features. Moment-DETR (Lei et al., 2021) addresses this by casting moment retrieval as set prediction, and QD-DETR (Moon et al., 2023) further conditions moment queries on the text query to strengthen alignment. TimeChat (Ren et al., 2024) and UniVTG (Lin et al., 2023) extend the paradigm to timestamp-aware representations and multi-task unification. In medical imaging, visual grounding similarly connects observations with clinical descriptions, either by detecting anatomical regions and generating region-specific descriptions (Tanida et al., 2023) or by grounding reports through memory-driven generation (Chen et al., 2020); yet they do not support multi-step reasoning.

These methods retrieve evidence from a fixed query in a single pass, leaving selection independent of the reasoning it supports. SurgCoT (Wang et al., 2026b) provides staged questions and evidence-clue annotations for progressive surgical video reasoning. Building on this, SurgCouple couples evidence selection with successive reasoning stages, connecting global hypotheses with local visual evidence.

3 PROPOSED SURGCOUPLE

3.1 PROBLEM FORMULATION

Surgical video reasoning is formulated as a coupled process of evidence discovery and reasoning, connecting global context with local visual evidence. Formally, let $\mathcal{I} = (\boldsymbol{H}, \boldsymbol{Q}, \boldsymbol{K}, \{(\boldsymbol{Q}_j, \boldsymbol{O}_j)\}_{j=1}^{3})$, where $\boldsymbol{H}$ denotes hierarchical video states (Fig. 2), $\boldsymbol{Q}$ is the main question, $\boldsymbol{K}$ is SurgCoT-provided knowledge (Wang et al., 2026b), and $(\boldsymbol{Q}_j, \boldsymbol{O}_j)$ is the question and answer candidates for reasoning stage j. The main question has separate answer candidates $\boldsymbol{O} = \{\boldsymbol{o}_m\}_{m=1}^{M}$.

The goal is to predict a final answer via a *visually grounded reasoning chain* of three states $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$, corresponding to global hypothesis formation, localized spatiotemporal interpretation, and fine-grained verification. $\mathcal{R}_1$ captures a global hypothesis from the video, question, and knowledge context; subsequent states incorporate progressively discovered evidence $\mathcal{E}_t, \mathcal{E}_s, \mathcal{E}_f$ (temporal, spatial, and fine-grained cues). **Evidence Chain** and **Reasoning Chain** are defined as:

$$\textbf{EC:}\quad \mathcal{E}_t = g_t(\mathcal{I}), \qquad \mathcal{E}_s = g_s(\mathcal{I}, \mathcal{R}_1, \mathcal{E}_t), \qquad \mathcal{E}_f = g_f(\mathcal{I}, \mathcal{R}_1, \mathcal{R}_2, \mathcal{E}_t, \mathcal{E}_s); \tag{1}$$

$$\textbf{RC:}\quad \mathcal{R}_1 = f_1(\mathcal{I}, \boldsymbol{O}), \quad \mathcal{R}_2 = f_2(\mathcal{I}, \boldsymbol{O}, \mathcal{E}_t, \mathcal{E}_s, \mathcal{R}_1), \quad \mathcal{R}_3 = f_3(\mathcal{I}, \boldsymbol{O}, \mathcal{E}_t, \mathcal{E}_s, \mathcal{E}_f, \mathcal{R}_1, \mathcal{R}_2), \tag{2}$$

where g_t, g_s, g_f are evidence discovery functions and f_1, f_2, f_3 are reasoning functions. A shared trainable answer decoder D_{ans} produces a stage-wise prediction $\hat{A}_j$ from each R_j ($j = 1, 2, 3$) and final answer $\hat{A}$ from the full evidence–reasoning chain. $\hat{A}_3$ and $\hat{A}$ use different decoding contexts for the third-stage and final decisions, respectively. This couples evidence discovery and reasoning across stages: preceding states guide evidence selection, and selected evidence conditions next reasoning.

3.2 EVIDENCE CHAIN

The Evidence Chain progressively discovers temporal, spatial, and fine-grained evidence to refine global hypotheses into local judgments. A frozen Qwen3-VL backbone (Bai et al., 2025a) provides hierarchical video states $\boldsymbol{H}$, comprising a global matrix $\boldsymbol{H}^v \in \mathbb{R}^{L_v \times d_h}$, clip-level features $\boldsymbol{H}^c = \{\boldsymbol{h}_i^c\}_{i=1}^{N}$ with $\boldsymbol{h}_i^c \in \mathbb{R}^{d_h}$, and frame-level features $\boldsymbol{H}^f = \{\boldsymbol{h}_j^f\}_{j=1}^{N_f}$; the question $\boldsymbol{Q}$ and knowledge $\boldsymbol{K}$ are embedded into $\boldsymbol{q}$ and $\boldsymbol{k}$ by the same backbone. For each of the three stages below, a scoring function ϕ_ℓ ($\ell \in \{t, s, f\}$) evaluates candidate visual features under the question, knowledge, and preceding reasoning states where applicable, selecting evidence for subsequent reasoning.

Temporal Evidence Discovery. The model first scores each clip feature $\boldsymbol{h}_i^c$ against the question $\boldsymbol{q}$ and knowledge $\boldsymbol{k}$ as $s_i^t = \phi_t([\boldsymbol{h}_i^c; \boldsymbol{q}; \boldsymbol{k}])$. The scores are normalized as $p_i^t = \exp(s_i^t) / \sum_{j=1}^{N} \exp(s_j^t)$, and the top-$K_t$ clips form the evidence set $\mathcal{E}_t$. Temporal discovery is driven solely by $\boldsymbol{q}$ and $\boldsymbol{k}$, deliberately excluding $\mathcal{R}_1$ to avoid narrowing the temporal scope to fragments that merely confirm it.

Spatial Evidence Grounding. Within the temporal scope, the model uses the global hypothesis $\mathcal{R}_1$ to guide spatial search, *i.e.*, it scores region r in clip i as $s_{i,r}^s = \phi_s([\Gamma(\boldsymbol{h}_{i,r}^c, \mathcal{R}_1); \boldsymbol{q}; \boldsymbol{k}; \mathcal{R}_1])$, where $\boldsymbol{h}_{i,r}^c$ is the feature of clip i in region r. Here Γ is a gated cross-attention block:

$$\Gamma(\boldsymbol{h}_{i,r}^c, \mathcal{R}_1) = \sigma(\boldsymbol{W}_g[\boldsymbol{h}_{i,r}^c; \mathcal{R}_1]) \odot \mathrm{CrossAttn}(\boldsymbol{h}_{i,r}^c, \mathcal{R}_1), \tag{3}$$

where $\boldsymbol{W}_g$ is a learnable gate projection, $\sigma(\cdot)$ is the sigmoid, $\odot$ is element-wise multiplication, and $\mathrm{CrossAttn}$ is multi-head cross-attention. The gate modulates the visual feature by the global hypothesis, amplifying hypothesis-relevant cues while preserving raw evidence for later verification. The highest-scoring regions form the spatial evidence set $\mathcal{E}_s$.

Fine-grained Evidence Selection. Within the spatiotemporal scope defined by $\mathcal{E}_t$ and $\mathcal{E}_s$, the final stage scores candidate patch p in frame j as $s_{j,p}^f = \phi_f([\Gamma(\boldsymbol{h}_{j,p}^f, \mathcal{R}_2); \boldsymbol{q}; \boldsymbol{k}; \mathcal{R}_1; \mathcal{R}_2])$, where $\boldsymbol{h}_{j,p}^f$ is the patch feature and the top-scoring patches form $\mathcal{E}_f$. Conditioning on $\mathcal{R}_1$ and $\mathcal{R}_2$ lets the global hypothesis and localized interpretation jointly determine which visual details to examine, so the selection verifies the current interpretation rather than merely localizing objects. This coarse-to-fine structure, $\mathcal{E}_t \to \mathcal{E}_s \to \mathcal{E}_f$, confines fine-grained scrutiny to an already localized scope, avoiding costly dense matching over the full video.

3.3 REASONING CHAIN

The Reasoning Chain implements a *hierarchically conditioned cascade* that addresses a core challenge in surgical video reasoning: local evidence is inherently ambiguous without its global procedural context, *e.g.*, the same tissue or instrument can carry different meanings at different surgical phases. To resolve this ambiguity, each stage is conditioned on both the selected evidence and the preceding reasoning state, carrying global procedural context forward while incorporating progressively finer detail. This yields the sequence $\mathcal{R}_1 \to \mathcal{R}_2 \to \mathcal{R}_3$, where each stage refines the inherited context with local evidence. Each $\mathcal{R}_j$ is supervised through the shared answer decoder to produce a stage-wise prediction $\hat{A}_j$, anchoring intermediate states to the task and preventing drift into task-irrelevant abstraction. As visual evidence salience is hypothesis-dependent, $\mathcal{R}_1$ and $\mathcal{R}_2$ are fed back to guide evidence selection (Sec. 3.2).

Global Hypothesis Formation. Let $\boldsymbol{q}_j$ and $\boldsymbol{o}_j$ denote the backbone embeddings of the stage question $\boldsymbol{Q}_j$ and its answer candidates $\boldsymbol{O}_j$, respectively. Interpreting local observations in a long surgical video requires an initial understanding of the procedural context. We therefore form the first reasoning state from the global video representation, the main-question context, and the first-stage question and candidates: $\mathcal{R}_1 = f_1([\boldsymbol{H}^v; \boldsymbol{q}; \boldsymbol{o}; \boldsymbol{k}; \boldsymbol{q}_1; \boldsymbol{o}_1])$, where f_1 is a trainable multimodal Transformer module (the same architecture is used for f_2 and f_3). As a revisable global hypothesis, $\mathcal{R}_1$ guides subsequent evidence selection and is refined by local evidence.

Evidence-Conditioned Local Refinement. $\mathcal{R}_1$ captures the procedure at a coarse level but remains ungrounded in the selected video content. This stage grounds it by aggregating the spatiotemporal evidence $\tilde{\boldsymbol{H}}_2$ from $\mathcal{E}_t$ and $\mathcal{E}_s$, and computing $\mathcal{R}_2 = f_2([\Gamma(\tilde{\boldsymbol{H}}_2, \mathcal{R}_1); \boldsymbol{q}; \boldsymbol{o}; \boldsymbol{k}; \boldsymbol{q}_2; \boldsymbol{o}_2; \mathcal{R}_1])$, where Γ is the gated cross-attention block of Eq. 3. The gate conditions the aggregated evidence on $\mathcal{R}_1$, so that $\mathcal{R}_2$ becomes a video-grounded refinement of the global hypothesis.

Fine-grained Verification. $\mathcal{R}_2$ refines the global hypothesis with spatiotemporal evidence, but its interpretation remains open to error at the finest granularity. This final stage verifies it against the selected fine-grained evidence by aggregating $\tilde{\boldsymbol{H}}_3$ from $\mathcal{E}_f$ and computing $\mathcal{R}_3 = f_3([\Gamma(\tilde{\boldsymbol{H}}_3, \mathcal{R}_2); \boldsymbol{q}; \boldsymbol{o}; \boldsymbol{k}; \boldsymbol{q}_3; \boldsymbol{o}_3; \mathcal{R}_1; \mathcal{R}_2])$. Retaining $\mathcal{R}_1$ and $\mathcal{R}_2$, this stage can confirm or revise the local interpretation against fine-grained cues, yielding the verified reasoning state $\mathcal{R}_3$.

3.4 CONSISTENCY VERIFICATION FOR COUPLED REASONING

Coupling the two chains through the gated cross-attention mechanism of Eq. 3 makes evidence selection and reasoning interdependent, yet does not by itself enforce that the resulting chain remains internally consistent. We therefore introduce a lightweight training-time Consistency Verifier $\mathcal{V}$ that jointly attends to both streams:

$$s_v = \mathcal{V}(\mathcal{R}_{1:3},\, \mathcal{E}_{t:s:f},\, \boldsymbol{H}) \in [0, 1]. \tag{4}$$

It targets three failure modes of a coupled chain: 1) *evidence–state alignment*, whether each evidence-conditioned state remains matched to the evidence used to produce it; 2) *cross-step coherence*, whether adjacent states are compatible, so that refinement does not contradict what it inherits; and 3) *scope consistency*, whether the selected regions lie within the selected clips and the fine-grained cues within those regions, preserving the hierarchical inclusion the Evidence Chain assumes. It is trained to distinguish original chains from corrupted ones:

$$\mathcal{L}_{\text{ver}} = -y \log s_v - (1 - y) \log(1 - s_v), \tag{5}$$

where $y = 1$ for original chains and $y = 0$ for corrupted ones obtained by shuffling evidence, permuting reasoning steps, or breaking stage dependencies. The verifier is used only during training.

To align local reasoning states with the selected evidence, we introduce a consistency loss between their projected representations:

$$\mathcal{L}_{\text{couple}} = \text{DIS}\big(\psi_2(\mathcal{R}_2), \eta_2(\tilde{\boldsymbol{H}}_2)\big) + \text{DIS}\big(\psi_3(\mathcal{R}_3), \eta_3(\tilde{\boldsymbol{H}}_3)\big), \tag{6}$$

where $\tilde{\boldsymbol{H}}_2, \tilde{\boldsymbol{H}}_3$ are detached features aggregated from $(\mathcal{E}_t, \mathcal{E}_s)$ and $\mathcal{E}_f$, respectively, ψ_i, η_i are projection heads, and DIS is a distance metric. Temporal evidence identifies relevant clips and spatial evidence localizes regions within them, jointly providing the spatiotemporal context for $\mathcal{R}_2$. $\mathcal{R}_1$ is excluded as it derives from the global representation rather than selected evidence, and detaching the evidence features prevents this loss from backpropagating through them on the target branch.

3.5 Training Objective

The framework is optimized jointly for answer prediction, evidence selection, representation alignment, and chain consistency:

$$\mathcal{L} = \lambda_{\text{ans}}\mathcal{L}_{\text{ans}} + \lambda_{\text{evi}}\mathcal{L}_{\text{evi}} + \lambda_{\text{couple}}\mathcal{L}_{\text{couple}} + \lambda_{\text{ver}}\mathcal{L}_{\text{ver}}, \tag{7}$$

where $\mathcal{L}_{\text{ans}} = \mathcal{L}_A + \sum_{j=1}^{3}\mathcal{L}_{A_j}$ supervises the final and stage-wise answers via token-level cross-entropy: $\mathcal{L}_A = -\sum_{u=1}^{|A|}\log P_\theta(a_u \mid a_{<u}, \mathcal{C}_A)$, with $\mathcal{C}_A$ the full decoding context (question, knowledge, reasoning states, and selected evidence); $\mathcal{L}_{A_j} = -\sum_{u=1}^{|A_j|}\log P_\theta(a_u \mid a_{<u}, \mathcal{C}_{A_j})$, where $\mathcal{C}_{A_j}$ is the context available up to stage j. Evidence selection is supervised by $\mathcal{L}_{\text{evi}} = -\sum_{\ell\in\{t,s,f\}}\sum_{i\in\mathcal{E}_\ell} y_i^\ell \log p_i^\ell$, where $\boldsymbol{p}^\ell = \text{softmax}(\boldsymbol{s}^\ell)$ over the level-ℓ candidates $\mathcal{E}_\ell$, and $y_i^\ell = m_i^\ell / \sum_{k\in\mathcal{E}_\ell} m_k^\ell$ normalizes over candidates matched to the SurgCoT evidence annotations.

During curriculum training, we gradually introduce model-predicted evidence. At inference, SurgCouple performs evidence selection and reasoning in the prescribed stage order.

4 Experimental Results

4.1 Experimental Setup

Dataset. SurgCouple is evaluated on SurgCoT (Wang et al., 2026b), a benchmark for CoT reasoning in surgical videos, with 2,841 videos across 7 specialties. Each question is annotated $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{C}, \boldsymbol{O}, \boldsymbol{A})$, where $\boldsymbol{A}$ denotes the answer, and $\boldsymbol{C}$ the evidence clues, namely temporal anchors, visual landmarks, and region descriptions, supervising temporal, spatial, and fine-grained evidence discovery. Stage-wise answer annotations evaluates progressive reasoning. We split the videos in a ratio of 7:2:1 (1,989/568/284 for training/validation/test), assigning all clips and question-answer pairs from the same source video to the same split, with every split covering all 7 specialties.

Compared Methods. We compare SurgCouple with four baseline groups: 1) *closed-source commercial MLLMs:* GPT-5.5 (OpenAI, 2026), GPT-5 (OpenAI, 2025), Gemini-3.1-Pro (Google DeepMind, 2026), Gemini-2.5-Pro (Gemini Team, Google, 2025), Claude-Opus-4.7 (Anthropic, 2026), Claude-Sonnet-4.5 (Anthropic, 2025); 2) *open-source general MLLMs:* Qwen3.5-9B (Qwen Team, 2026), Qwen3-VL-8B (Bai et al., 2025a), Qwen2.5-VL-7B (Bai et al., 2025b), InternVL3.5 (Wang et al., 2025); 3) *medical-domain MLLMs:* MedGemma-27B-IT (Sellergren et al., 2025), Lingshu-7B (Xu et al., 2025), LLaVA-Med-7B (Li et al., 2023a), HuatuoGPT-Vision-7B (Chen et al., 2024a); 4) Qwen3-VL-8B-based baselines. Two baselines share SurgCouple's backbone and answer supervision while ablating its evidence mechanisms. Fixed-Aggregation Fine-Tuning retains the SurgCouple architecture but replaces learned evidence selection with fixed visual aggregation. CoT Fine-Tuning drops the evidence chain and instead fine-tunes Qwen3-VL-8B with LoRA to emit a textual reasoning step before each stage-wise answer.

Metrics. We report three sets of metrics: 1) *Answer correctness:* **Final Acc.** on the final answer $\hat{A}$ and stage-wise accuracy $(\textbf{Acc}_1, \textbf{Acc}_2, \textbf{Acc}_3)$ on the stage-wise answers $\hat{A}_1, \hat{A}_2, \hat{A}_3$. 2) *Joint stage correctness:* **Valid Chain Rate (VCR)**, the fraction of samples for which all three stage answers are correct; 3) *Evidence alignment:* **Evidence Score**, the average of temporal, spatial, and fine-grained evidence scores, each measuring the alignment of selected evidence with the ground-truth clues.

Implementation Details. We instantiate SurgCouple with a frozen Qwen3-VL-8B backbone (Bai et al., 2025a). Videos are sampled as 16 clips, and the Evidence Chain selects top-$K_t = 4$ clips, top-$K_s = 8$ regions, and top-$K_f = 16$ fine-grained candidates. Training uses AdamW (lr 2×10^{-5}, 5 epochs) with loss weights $\lambda_{\text{ans}} = 1.0$, $\lambda_{\text{evi}} = 0.5$, $\lambda_{\text{couple}} = 0.5$, and $\lambda_{\text{ver}} = 0.2$. The loss weights were selected based on validation-set performance. Inference employs greedy decoding. More implementation details are provided in the supplementary materials.

4.2 Comparison with State-of-the-Art Models

Table 1 compares SurgCouple with closed-source, open-source, and medical MLLMs, as well as two Qwen3-VL-8B-based supervised baselines. SurgCouple couples global context with localized

Table 1: **Main results on SurgCoT.** We report final-answer accuracy (**Final Acc.**), stage-wise accuracy (**Acc**$_1$/**Acc**$_2$/**Acc**$_3$), and Valid Chain Rate (**VCR**), which requires all three stage-wise answers to be correct for the same sample. Best in **bold**; second-best underlined.

	Model	Final Acc. (%)	Acc$_1$ (%)	Acc$_2$ (%)	Acc$_3$ (%)	VCR (%)
Closed	GPT-5.5 (OpenAI, 2026)	<u>88.92</u>	<u>76.42</u>	<u>68.37</u>	<u>61.68</u>	<u>52.63</u>
	GPT-5 (OpenAI, 2025)	81.67	53.87	57.61	46.38	46.27
	Gemini-3.1-Pro (Google DeepMind, 2026)	86.83	74.21	64.21	61.51	49.18
	Gemini-2.5-Pro (Gemini Team, Google, 2025)	81.83	34.15	44.04	43.06	30.84
	Claude-Opus-4.7 (Anthropic, 2026)	87.58	72.81	64.69	57.34	46.48
	Claude-Sonnet-4.5 (Anthropic, 2025)	77.64	50.91	40.35	37.82	32.11
Open	Qwen3.5-9B (Qwen Team, 2026)	82.26	74.89	52.70	52.84	41.76
	Qwen3-VL-8B (Bai et al., 2025a)	81.94	67.11	55.02	49.36	34.01
	Qwen2.5-VL-7B (Bai et al., 2025b)	69.37	34.12	25.41	27.86	20.73
	InternVL3.5-8B (Wang et al., 2025)	74.61	67.12	65.43	52.08	32.17
Med.	MedGemma-27B-IT (Sellergren et al., 2025)	76.49	54.73	56.92	76.18	42.84
	Lingshu-7B (Xu et al., 2025)	69.48	36.11	47.93	43.52	25.16
	LLaVA-Med-7B (Li et al., 2023a)	75.21	31.74	36.48	36.39	28.31
	HuatuoGPT-Vision-7B (Chen et al., 2024a)	71.26	69.84	67.91	55.61	18.73
FT.	Fixed-Aggregation Fine-Tuning	82.74	66.47	55.82	49.58	35.43
	CoT Fine-Tuning	84.03	69.18	59.64	52.13	38.68
	SurgCouple	**89.31**	**76.83**	**72.84**	**79.57**	**58.92**

evidence, and its gains concentrate where global context alone is insufficient: fine-grained and chain-level correctness. Five findings support this: it attains the best chain-level accuracy; the gains depend on coupling rather than on any single component; they concentrate where coupling operates; they grow with evidence granularity; and they reverse the fine-grained degradation.

1) SurgCouple attains the best chain-level and fine-grained accuracy. It reaches 89.31% final accuracy and 58.92% VCR, the only VCR above 53% among all methods. Against GPT-5.5 (OpenAI, 2026), the strongest baseline, the final-accuracy margin is only 0.39%, whereas those on **Acc**$_3$ and VCR reach 17.89% and 6.29%; relative to its own Qwen3-VL-8B backbone, it gains 7.37% in final accuracy and 24.91% in VCR. The advantage is therefore concentrated where global context alone cannot settle the prediction, the fine-grained stage and the joint correctness of the chain, rather than in the endpoint answer that global information already supports.

2) The gains depend on coupling global context with local evidence. The two supervised baselines form an ablation ladder that isolates this contribution. Fixed-Aggregation Fine-Tuning keeps SurgCouple's frozen backbone, Reasoning Chain, and answer decoder but replaces the learned evidence chain with fixed visual aggregation; it stays close to the backbone, with final accuracy changing by $+0.80\%$, **Acc**$_1$ by -0.64%, and VCR by $+1.42\%$. Restoring the coupling under otherwise identical settings raises **Acc**$_3$ by 29.99% and VCR by 23.49%, whereas CoT Fine-Tuning, which supervises textual reasoning without selecting evidence, remains 27.44% behind on **Acc**$_3$. The gains thus do not survive the removal of coupling and originate in its joint operation: evidence discovery coupled with evidence-conditioned reasoning.

3) The gains concentrate where coupling operates. The gains are highly uneven along the chain, rising from +9.72% at stage 1 to +17.82% and +30.21% at stages 2 and 3 over the backbone, with VCR gaining +24.91%. This matches the design: $\mathcal{R}_1$ derives from the global representation and is excluded from the coupling objective, whereas $\mathcal{R}_2$ and $\mathcal{R}_3$ condition on the selected temporal, spatial, and fine-grained evidence; the Consistency Verifier, which enforces cross-step coherence and scope consistency along the coupled chain, accounts for the VCR gain. The improvement thus concentrates where global context is coupled with local evidence and where that coupling is kept coherent.

4) Gains grow with evidence granularity as the design predicts. The gains increase monotonically from stage 1 to stage 3 across all four reference points, reaching $+9.72\%/+17.82\%/+30.21\%$ over the backbone, with the same ordering holding against Fixed-Aggregation Fine-Tuning, CoT Fine-Tuning, and GPT-5.5. This pattern follows from the coarse-to-fine design of the Evidence Chain: as it proceeds from temporal to spatial to fine-grained cues, later stages depend more on precisely localized evidence and benefit more from explicit selection, although the lower baseline accuracy at finer stages also contributes to the larger margins.

Table 2: Ablation study of SurgCouple's core components.

SN	Variant	Final Acc. (%)	$\mathbf{Acc}_1$ (%)	$\mathbf{Acc}_2$ (%)	$\mathbf{Acc}_3$ (%)	Evidence Score (%)	VCR (%)
1	Full SurgCouple	**89.31**	**76.83**	**72.84**	**79.57**	**74.62**	**58.92**
2	w/o Learned Evidence Selection	85.16	72.48	62.37	66.84	51.73	39.86
3	w/o Bidirectional Coupling	86.42	73.61	66.58	72.49	64.35	45.27
4	w/o Consistency Verifier	87.46	75.12	70.34	67.28	72.18	47.63
5	w/o Stage-wise Supervision	85.83	70.26	63.74	69.31	65.47	42.58

5) Evidence selection reverses the fine-grained degradation. Across Qwen3-VL-8B-based variants sharing the backbone, architecture, and supervision, those without learnable evidence selection degrade as stages become finer, with $\mathbf{Acc}_3$ lowest in every case (backbone: 67.11% at stage 1 versus 49.36% at stage 3). SurgCouple reverses this trend, making $\mathbf{Acc}_3$ (79.57%) its highest stage-wise result (76.83/72.84/79.57). Since these variants differ only in whether evidence is selected learnably, the reversal is attributable to evidence discovery rather than to the backbone or supervision; the fine-grained stage ceases to be the weakest link, though it remains imperfect.

4.3 Ablation Study

Table 2 ablates learned evidence selection, bidirectional coupling, consistency verification, and stage-wise supervision, by removing each core component in turn. It reveals four key findings.

1) Every component is necessary. Removing any component lowers every metric, with VCR dropping by 11.29%–19.06% and final accuracy by 1.85%–4.15%, so none is redundant. Each ablated variant remains below the full framework on all six metrics.

2) Learned selection governs whether evidence is found. Evidence Score drops most when the learned selector and its supervision are removed, from 74.62% to 51.73%, a loss of 22.89%, whereas the other three ablations lose at most 10.27%. Evidence quality thus depends chiefly on the Evidence Chain and its supervision, not on the components that later consume the evidence.

3) The Consistency Verifier governs whether evidence is used correctly. Removing it barely changes evidence quality, with Evidence Score down by only 2.44%, and leaves the early stages almost intact, with final accuracy, $\mathbf{Acc}_1$, and $\mathbf{Acc}_2$ down by 1.85%, 1.71%, and 2.50%, the smallest losses in the table; it nevertheless costs 12.29% on $\mathbf{Acc}_3$ and 11.29% on VCR. The Verifier thus governs whether the selected evidence is converted into correct later-stage and joint judgments.

4) Coupling improves evidence selection. Removing bidirectional coupling degrades both the consuming stages, with $\mathbf{Acc}_2$ and $\mathbf{Acc}_3$ down by 6.26% and 7.08%, and Evidence Score, down by 10.27%, even though the Evidence Chain itself is intact. Coupling is thus not a one-way conduit passing fixed evidence to the Reasoning Chain; it feeds reasoning outcomes back into selection, and the resulting Evidence Score drop of 10.27% is attributable specifically to this bidirectional feedback, since a unidirectional pipeline would not affect evidence quality.

4.4 Qualitative Analysis

Fig. 3 visualizes the global-to-local evidence trajectory on a laparoscopic pedunculated myomectomy case, where the question asks for the current risk and decision, and the correct answer is bleeding at the myoma base, requiring hemostasis before proceeding. SurgCouple narrows the evidence scope from the temporal window where the risk emerges, to the myoma-base region, and finally to the fine-grained bleeding point, committing to the decision only after this cue is verified; all three stage-wise predictions are correct.

The two supervised baselines fail differently. Fixed-Aggregation Fine-Tuning aggregates visual tokens uniformly, resolving some stages while erring on others. CoT Fine-Tuning produces a textual chain that is incomplete, describing the decision in words without grounding it in the visible cue. Both reach the fine-grained stage without the evidence that determines the answer.

This example illustrates what the stage-wise and VCR gains in Table 1 measure at the level of a single case: the decision depends on a localized cue that global context alone cannot resolve, which SurgCouple locates before answering and neither baseline does.

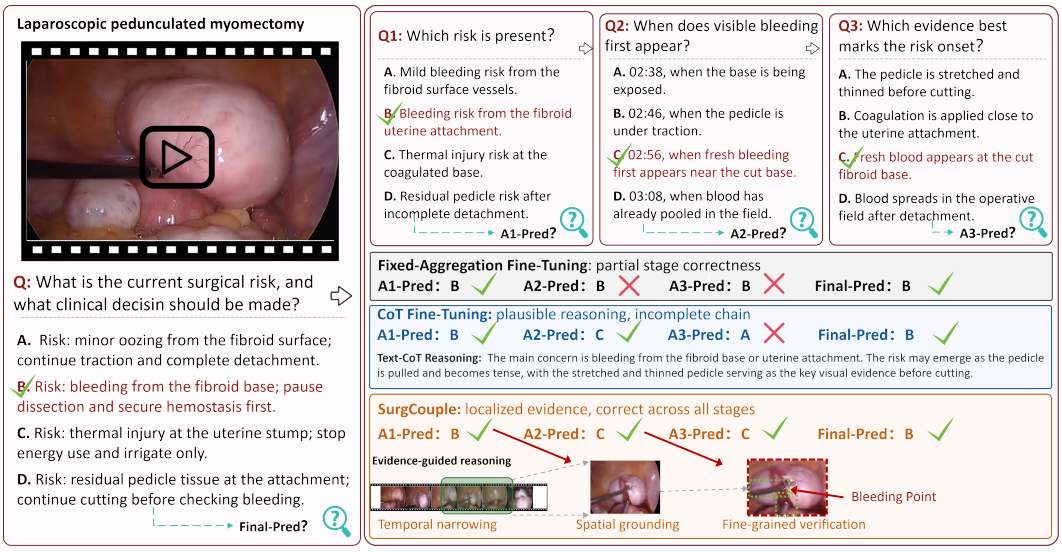


Figure 3: **Qualitative example of evidence-grounded reasoning.** SurgCouple narrows the evidence scope from temporal windows to spatial regions to fine-grained cues, making the visual evidence behind each reasoning stage explicit.

4.5 Generalization and Limitations

Generalization beyond SurgCoT. Neither Surgical-VQA (Seenivasan et al., 2022) nor Video-MME (Fu et al., 2025) provides stage questions, so at inference we set $(\boldsymbol{Q}_j, \boldsymbol{O}_j) = (\boldsymbol{Q}, \boldsymbol{O})$ for $j = 1, 2, 3$ and evaluate only the final answer, without updating parameters. The evidence chain still narrows from temporal to spatial to fine-grained cues; only the questions are shared across stages.

On Surgical-VQA, SurgCouple reaches 76.84%, exceeding Qwen3-VL (65.92%) and CoT Fine-Tuning (67.63%) by 10.92% and 9.21%. The advantage thus extends beyond SurgCoT's staged questions to an independent surgical setting.

On Video-MME, SurgCouple reaches 56.73%, vs. 52.71% for Qwen3-VL and 53.50% for CoT Fine-Tuning, gains of 4.02% and 3.23%, providing initial evidence of transfer beyond surgical video, though the smaller margin is expected because the evidence types are defined for surgery.

Limitations. SurgCouple relies on SurgCoT's stage-wise answers and evidence clues, which the ablation confirms drive its gains: removing evidence supervision lowers Evidence Score by 22.89%, and removing stage-wise supervision lowers the stage accuracies by 6.57%, 9.10%, and 10.26%. Extending it to weaker or sefl-generated supervision, so that it can be *trained* on datasets with only final-answer labels, is future work rather than a defect. Second, zero-shot transfer to Surgical-VQA and Video-MME covers a modest range of domains and task formats, and broader validation would help characterize when coupling global context with local evidence is most beneficial.

5 Conclusion

We introduce SurgCouple to bridge global procedural context and local visual evidence for surgical video reasoning. The Evidence Chain narrows evidence from temporal windows to spatial regions to fine-grained cues, while intermediate reasoning states guide increasingly localized selection. The Reasoning Chain then refines global hypotheses using the selected evidence, and the Consistency Verifier supervises evidence–reasoning consistency during training. Experiments show that SurgCouple improves intermediate-stage and joint-chain correctness, not merely the final answer: its VCR reaches 58.92%, surpassing 52.63% for GPT-5.5 and 34.01% for its own backbone, narrowing the gap between correct final predictions and reliable evidence-grounded reasoning. Ablation studies confirm that every component contributes, with learned evidence selection contributing most.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, code development, and figure refinement. The authors reviewed and verified the resulting text, code, and figures. The MLLMs evaluated in this study were used as experimental models, as described in the experimental setup. The authors are responsible for the method design, experimental implementation, result analysis, and conclusions.

ETHICS STATEMENT

This study uses existing publicly available surgical video reasoning benchmarks and does not involve the recruitment of human participants or the collection of new personal data. All experiments are conducted on previously released datasets under their respective usage conditions. The study focuses on methodological development and benchmark evaluation and does not provide clinical recommendations or support real-time clinical decision-making.

REPRODUCIBILITY STATEMENT

Section 3 presents the SurgCouple formulation and training objectives, while Section 4.1 reports the dataset splits, evaluation metrics, and implementation settings. Appendix A details the inference, training, and ablation procedures; Appendix B defines the Evidence Score and Valid Chain Rate; and Appendix C specifies the baseline prompting, answer parsing, and decoding protocols. The source code will be made publicly available upon acceptance.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23716–23736, 2022.

Anthropic. Introducing claude sonnet 4.5. `https://www.anthropic.com/news/claude-sonnet-4-5`, September 2025. Accessed: 2026-05-02.

Anthropic. What's new in claude opus 4.7. `https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-7`, 2026. Claude API Docs. Accessed: 2026-04-26.

Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. Surgical-vqla:transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pp. 6859–6865. IEEE, 2023. doi: 10.1109/ICRA48891.2023.10160403. URL `https://doi.org/10.1109/ICRA48891.2023.10160403`.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal LLMs at scale. *arXiv preprint arXiv:2406.19280*, 2024a.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:44393–44418, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, June 2024c.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449, 2020.

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. Springer, 2020.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9062–9072, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5267–5275, 2017.

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. Springer, 2021.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Google DeepMind. Gemini 3.1 pro - model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026. Published 19 February 2026. Accessed: 2026-04-26.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Pengfei Hao, Hongqiu Wang, Shuaibo Li, Zhaohu Xing, Guang Yang, Kaishun Wu, and Lei Zhu. Surgical-mamballm: Mamba2-enhanced multimodal large language model for vqla in robotic surgery. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 573–583. Springer Nature Switzerland, 2025.

Ming Hu, Peng Xia, Lin Wang, Siyuan Yan, Feilong Tang, Zhongxing Xu, Yimin Luo, Kaimin Song, Jurgen Leitner, Xuelian Cheng, et al. Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. In *European Conference on Computer Vision (ECCV)*, pp. 481–500. Springer, 2024.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

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. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023a.

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. PMLR, 2023b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2794–2804, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 34892–34916, 2023.

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

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14268–14280, 2025.

WonJun Moon, Sangeek Hyun, Sang-uk Park, Dongheon Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23023–23033, 2023.

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.

OpenAI. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`, August 2025. Accessed: 2026-05-02.

OpenAI. Introducing gpt-5.5. `https://openai.com/index/introducing-gpt-5-5/`, April 2026. OpenAI product release page. Accessed: 2026-04-26.

Qwen Team. Qwen3.5-9b. `https://huggingface.co/Qwen/Qwen3.5-9B`, March 2026. Accessed: 2026-05-02.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14313–14323. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01357. URL `https://doi.org/10.1109/CVPR52733.2024.01357`.

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. Springer, 2022.

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. Springer, 2023.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7433–7442, 2023.

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

Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, Hongbin Liu, Jiazheng Wang, Fan Zhang, Nicolas Padoy, Nassir Navab, and Hongliang Ren. Endochat: Grounded multimodal large language model for endoscopic surgery. *Medical Image Analysis*, 107:103789, 2026a.

Gui Wang, YongSong Zhou, Kaijun Deng, Wooi Ping Cheah, Rong Qu, Jianfeng Ren, and Linlin Shen. Surgcot: Advancing spatiotemporal reasoning in surgical videos through a chain-of-thought benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17011–17021, June 2026b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12870–12877, 2020.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024, 2024.

A Detailed Inference and Training Procedures

This section provides the complete inference and training procedures of SurgCouple. As described in the main paper, SurgCouple couples an Evidence Chain (EC), which progressively selects temporal, spatial, and fine-grained cues, with a Reasoning Chain (RC), which forms a global hypothesis and refines subsequent judgments using the selected evidence. A lightweight Consistency Verifier (CV) provides auxiliary training supervision for consistency between the selected evidence and reasoning states and coherence across stages.

A.1 Inference Procedure

At inference time, the model receives a surgical video V, a main question Q, dataset-provided knowledge K, optional main-question answer candidates $O = \{o_m\}_{m=1}^{M}$, and three stage question–candidate pairs $\{(Q_j, O_j)\}_{j=1}^{3}$. The model first encodes the multimodal input using the frozen Qwen3-VL backbone and then executes the coupled evidence–reasoning cascade. Importantly, ground-truth evidence clues are not used during inference. Instead, the Evidence Chain progressively selects temporal, spatial, and fine-grained cues, with preceding reasoning states guiding spatial and fine-grained selection. The selected evidence then informs subsequent judgments.

Algorithm 1: Inference Procedure of SurgCouple

Require: Surgical video V, main question and candidates (Q, O), stage questions and candidates $\{(Q_j, O_j)\}_{j=1}^{3}$, dataset-provided knowledge K

Require: Frozen multimodal backbone $\mathcal{B}$; Evidence Chain modules g_t, g_s, g_f; Reasoning Chain modules f_1, f_2, f_3; shared answer decoder D_{ans}

Require: Evidence budgets K_t, K_s, K_f

Ensure: Final answer $\hat{A}$, stage-wise answers $(\hat{A}_1, \hat{A}_2, \hat{A}_3)$, reasoning states $(\hat{R}_1, \hat{R}_2, \hat{R}_3)$, and selected evidence $(\hat{E}_t, \hat{E}_s, \hat{E}_f)$

Encode multimodal input.

Obtain video and textual features from the frozen backbone:

$$(H_v, H_c, H_f, q, k, o, \{q_j, o_j\}_{j=1}^{3}) = \mathcal{B}\big(V, Q, K, O, \{(Q_j, O_j)\}_{j=1}^{3}\big),$$

where H_v, H_c, H_f denote global, clip-level, and frame/patch-level video features; q, k, o denote embeddings of the main question, knowledge, and main candidates; and q_j, o_j denote embeddings of stage question Q_j and its candidates O_j.

Global hypothesis formation.

Form the first-stage reasoning state:

$$\hat{R}_1 = f_1([H_v; q; o; k; q_1; o_1]).$$

Decode $\hat{A}_1$ for (Q_1, O_1) from $\hat{R}_1$ using D_{ans}.

Temporal evidence discovery.

Score each clip using the main question and knowledge:

$$s_i^t = \phi_t([h_i^c; q; k]).$$

Normalize scores over the temporal candidates:

$$\boldsymbol{p}^t = \mathrm{softmax}(\boldsymbol{s}^t).$$

Select the top-K_t clips:

$$\hat{E}_t = \mathrm{TopK}(\boldsymbol{p}^t, K_t).$$

Spatial evidence grounding.

For each region candidate r within $\hat{E}_t$, compute its reasoning-modulated score:

$$s_{i,r}^s = \phi_s([\Gamma(h_{i,r}, \hat{R}_1); q; k; \hat{R}_1]).$$

Normalize scores over the spatial candidates:

$$\boldsymbol{p}^s = \mathrm{softmax}(\boldsymbol{s}^s).$$

13: Select the top-K_s spatial candidates:

$$\hat{E}_s = \mathrm{TopK}(\boldsymbol{p}^s, K_s).$$

14: **Evidence-conditioned local reasoning.**
15: Aggregate the selected temporal and spatial evidence:

$$\widetilde{H}_2 = \mathrm{Agg}(H_c, H_f, \hat{E}_t, \hat{E}_s).$$

16: Form the second-stage reasoning state:

$$\hat{R}_2 = f_2([\Gamma(\widetilde{H}_2, \hat{R}_1); q; o; k; q_2; o_2; \hat{R}_1]).$$

17: Decode $\hat{A}_2$ for (Q_2, O_2) from $\hat{R}_2$ using D_{ans}.

18: **Fine-grained evidence selection.**
19: For each fine-grained candidate p within $(\hat{E}_t, \hat{E}_s)$, compute its score:

$$s^f_{f,p} = \phi_f([\Gamma(h_{f,p}, \hat{R}_2); q; k; \hat{R}_1; \hat{R}_2]).$$

20: Normalize scores over the fine-grained candidates:

$$\boldsymbol{p}^f = \mathrm{softmax}(\boldsymbol{s}^f).$$

21: Select the top-K_f fine-grained candidates:

$$\hat{E}_f = \mathrm{TopK}(\boldsymbol{p}^f, K_f).$$

22: **Fine-grained reasoning and final prediction.**
23: Aggregate the selected fine-grained evidence:

$$\widetilde{H}_3 = \mathrm{Agg}(H_f, \hat{E}_f).$$

24: Form the third-stage reasoning state:

$$\hat{R}_3 = f_3([\Gamma(\widetilde{H}_3, \hat{R}_2); q; o; k; q_3; o_3; \hat{R}_1; \hat{R}_2]).$$

25: Decode $\hat{A}_3$ for (Q_3, O_3) from $\hat{R}_3$ using D_{ans}.
26: Decode the final answer $\hat{A}$ for (Q, O) from the full evidence–reasoning chain using D_{ans}.
return $\hat{A}$, $(\hat{A}_1, \hat{A}_2, \hat{A}_3)$, $(\hat{R}_1, \hat{R}_2, \hat{R}_3)$, $(\hat{E}_t, \hat{E}_s, \hat{E}_f)$

A.2 Training Procedure

During training, SurgCouple is optimized with four complementary objectives. The answer loss $\mathcal{L}_{\mathrm{ans}}$ supervises the final and stage-wise answers; the evidence loss $\mathcal{L}_{\mathrm{evi}}$ supervises temporal, spatial, and fine-grained evidence discovery; the coupling loss $\mathcal{L}_{\mathrm{couple}}$ aligns reasoning states with the corresponding evidence features; and the verifier loss $\mathcal{L}_{\mathrm{ver}}$ trains the Consistency Verifier to distinguish original evidence–reasoning chains from corrupted ones.

We use a progressive curriculum for evidence conditioning. Early in training, local reasoning mainly uses visual evidence selected according to the ground-truth clues. As training proceeds, ground-truth evidence is gradually replaced by model-predicted evidence to reduce the mismatch between training and inference.

Algorithm 2: Training Procedure of SurgCouple

Require: Training set $\mathcal{D} = \{(V, Q, K, O, \{(Q_j, O_j, A^*_j)\}_{j=1}^3, A^*, C^*)\}$
Require: Evidence clues $C^* = (E^*_t, E^*_s, E^*_f)$
Require: Frozen backbone $\mathcal{B}$; Evidence Chain g_t, g_s, g_f; Reasoning Chain f_1, f_2, f_3; shared answer decoder D_{ans}; Consistency Verifier $v(\cdot)$
Require: Evidence budgets K_t, K_s, K_f; loss weights $\lambda_{\mathrm{ans}}, \lambda_{\mathrm{evi}}, \lambda_{\mathrm{couple}}, \lambda_{\mathrm{ver}}$
Require: Total epochs T; teacher-forcing schedule $\alpha(t) \in [0, 1]$
Ensure: Trained parameters of EC, RC, D_{ans}, and CV

for epoch $t = 1, \ldots, T$ **do**
  for mini-batch $\mathcal{B}_t \subset \mathcal{D}$ **do**
    Encode multimodal inputs.
    Obtain video and textual features from the frozen backbone:

$$(H_v, H_c, H_f, q, k, o, \{q_j, o_j\}_{j=1}^{3}) = \mathcal{B}\big(V, Q, K, O, \{(Q_j, O_j)\}_{j=1}^{3}\big)\,.$$

    Predict evidence and reasoning chains.
    Form the first-stage reasoning state:

$$\hat{R}_1 = f_1([H_v; q; o; k; q_1; o_1]).$$

    Predict temporal evidence using the main question and knowledge:

$$\hat{\boldsymbol{p}}^t = \mathrm{softmax}(\phi_t([H_c; q; k])), \qquad \hat{E}_t = \mathrm{TopK}(\hat{\boldsymbol{p}}^t, K_t).$$

    Select temporal evidence for reasoning by stochastic teacher forcing:

$$\bar{E}_t = \begin{cases} E_t^*, & u_t < \alpha(t), \\ \hat{E}_t, & \text{otherwise}, \end{cases} \qquad u_t \sim \mathcal{U}(0,1).$$

    Score spatial candidates within the selected temporal scope $\bar{E}_t$:

$$s_{i,r}^s = \phi_s([\Gamma(h_{i,r}, \hat{R}_1); q; k; \hat{R}_1]).$$

    Normalize spatial scores and select evidence:

$$\hat{\boldsymbol{p}}^s = \mathrm{softmax}(\boldsymbol{s}^s), \qquad \hat{E}_s = \mathrm{TopK}(\hat{\boldsymbol{p}}^s, K_s).$$

    Let $E_s^{*,\mathrm{valid}}$ be the spatial candidates matched to the ground-truth clues within $\bar{E}_t$.
    Select spatial evidence for reasoning:

$$\bar{E}_s = \begin{cases} E_s^{*,\mathrm{valid}}, & u_s < \alpha(t) \text{ and } E_s^{*,\mathrm{valid}} \neq \varnothing, \\ \hat{E}_s, & \text{otherwise}, \end{cases} \qquad u_s \sim \mathcal{U}(0,1).$$

    Aggregate localized evidence and form the second-stage reasoning state:

$$\widetilde{H}_2 = \mathrm{Agg}(H_c, H_f, \bar{E}_t, \bar{E}_s),$$

$$\hat{R}_2 = f_2([\Gamma(\widetilde{H}_2, \hat{R}_1); q; o; k; q_2; o_2; \hat{R}_1]).$$

    Score fine-grained candidates within the scope $(\bar{E}_t, \bar{E}_s)$:

$$s_{f,p}^f = \phi_f([\Gamma(h_{f,p}, \hat{R}_2); q; k; \hat{R}_1; \hat{R}_2]).$$

    Normalize fine-grained scores and select evidence:

$$\hat{\boldsymbol{p}}^f = \mathrm{softmax}(\boldsymbol{s}^f), \qquad \hat{E}_f = \mathrm{TopK}(\hat{\boldsymbol{p}}^f, K_f).$$

    Let $E_f^{*,\mathrm{valid}}$ be the fine-grained candidates matched to the ground-truth clues within $(\bar{E}_t, \bar{E}_s)$.
    Select fine-grained evidence for reasoning:

$$\bar{E}_f = \begin{cases} E_f^{*,\mathrm{valid}}, & u_f < \alpha(t) \text{ and } E_f^{*,\mathrm{valid}} \neq \varnothing, \\ \hat{E}_f, & \text{otherwise}, \end{cases} \qquad u_f \sim \mathcal{U}(0,1).$$

    Aggregate fine-grained evidence and form the third-stage reasoning state:

$$\widetilde{H}_3 = \mathrm{Agg}(H_f, \bar{E}_f),$$

$$\hat{R}_3 = f_3([\Gamma(\widetilde{H}_3, \hat{R}_2); q; o; k; q_3; o_3; \hat{R}_1; \hat{R}_2]).$$

    Use D_{ans} to produce each $\hat{A}_j$ for (Q_j, O_j) from $\hat{R}_j$, and the final $\hat{A}$ for (Q, O) from the full evidence–reasoning chain.

    Construct verifier training pairs.
    Use the uncorrupted evidence–reasoning chain as a positive training example:

$$\mathcal{Z}^+ = (\hat{R}_1, \hat{R}_2, \hat{R}_3, \bar{E}_t, \bar{E}_s, \bar{E}_f).$$

Construct corrupted chains:

$$\mathcal{Z}^- = \mathrm{Corrupt}(\mathcal{Z}^+),$$

where $\mathrm{Corrupt}(\cdot)$ includes evidence shuffling, reasoning-step permutation, and progressive-scope violation.

Compute verifier scores:

$$s_v^+ = v(\mathcal{Z}^+, H_v, H_c, H_f), \qquad s_v^- = v(\mathcal{Z}^-, H_v, H_c, H_f).$$

Compute training objectives.

Compute token-level final and stage-wise answer losses:

$$\mathcal{L}_{\mathrm{ans}} = \mathcal{L}_A + \sum_{j=1}^{3} \mathcal{L}_{A_j}.$$

Teacher-forced evidence is restricted to candidates within the selected parent scope. If no ground-truth match exists in that scope, use model-predicted evidence and omit the corresponding evidence loss.

Compute evidence discovery loss:

$$\mathcal{L}_{\mathrm{evi}} = -\sum_{\ell \in \{t,s,f\}} \sum_{i \in \mathcal{E}_\ell} y_i^\ell \log \hat{p}_i^\ell,$$

where y_i^ℓ is the normalized matching target defined in the main paper; levels without matched candidates are excluded.

Compute evidence–reasoning coupling loss:

$$\mathcal{L}_{\mathrm{couple}} = d\Big(\psi_2(\hat{R}_2), \eta_2(\mathrm{sg}(\widetilde{H}_2))\Big) + d\Big(\psi_3(\hat{R}_3), \eta_3(\mathrm{sg}(\widetilde{H}_3))\Big),$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient.

Compute verifier loss:

$$\mathcal{L}_{\mathrm{ver}} = -\log s_v^+ - \log(1 - s_v^-).$$

Optimize the total objective:

$$\mathcal{L} = \lambda_{\mathrm{ans}}\mathcal{L}_{\mathrm{ans}} + \lambda_{\mathrm{evi}}\mathcal{L}_{\mathrm{evi}} + \lambda_{\mathrm{couple}}\mathcal{L}_{\mathrm{couple}} + \lambda_{\mathrm{ver}}\mathcal{L}_{\mathrm{ver}}.$$

Update EC, RC, D_{ans}, and CV parameters with AdamW; keep $\mathcal{B}$ frozen.

end for

end for

return trained SurgCouple model.

A.3 Ablation Settings

Without learned evidence selection. This variant replaces learned evidence scoring with deterministic uniform selection. Four clips are selected uniformly from the 16 temporal candidates. Spatial regions and fine-grained patches are selected on fixed grids within the selected clips and regions, respectively, using the same candidate sets and selection budgets as the full model. Selection is independent of the question, knowledge, reasoning states, and ground-truth clues. The evidence scoring heads and $\mathcal{L}_{\mathrm{evi}}$ are removed, while the Reasoning Chain, coupling loss, and Consistency Verifier are retained. This variant bypasses evidence teacher forcing: the fixed selections are supplied to the Reasoning Chain during both training and inference. Evidence Score is computed from the selections supplied to the Reasoning Chain using the same matching rules as for the full model.

Fixed-Aggregation Fine-Tuning baseline. This baseline uses the same frozen Qwen3-VL-8B backbone, Reasoning Chain architecture, and shared answer decoder as SurgCouple. Visual features are aggregated by a fixed procedure rather than learned evidence selection.

B Definitions of Evidence Score and Valid Chain Rate

We define the evidence alignment and joint stage correctness metrics used in the main paper. For each sample n, the model predicts a reasoning chain $(\hat{R}_{1,n}, \hat{R}_{2,n}, \hat{R}_{3,n})$, stage-wise answers $(\hat{A}_{1,n}, \hat{A}_{2,n}, \hat{A}_{3,n})$, and an evidence chain $(\hat{E}_{t,n}, \hat{E}_{s,n}, \hat{E}_{f,n})$. The ground-truth annotations provide stage-wise answers $(A^*_{1,n}, A^*_{2,n}, A^*_{3,n})$ and evidence clues $(E^*_{t,n}, E^*_{s,n}, E^*_{f,n})$.

Evidence Score. Evidence Score measures whether the predicted evidence chain is aligned with the annotated evidence clues at temporal, spatial, and fine-grained levels:

$$\text{Evidence Score} = \frac{1}{3}(S_t + S_s + S_f).$$

The temporal score S_t is defined as:

$$S_t = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left[\text{Match}_t(\hat{E}_{t,n}, E^*_{t,n}) = 1\right].$$

When the temporal clue is provided as a timestamp, $\text{Match}_t = 1$ if the timestamp falls into one of the selected top-K_t clips. When the temporal clue is provided as an interval, $\text{Match}_t = 1$ if the maximum temporal IoU between the predicted temporal evidence and the annotated interval exceeds a threshold τ_t.

The spatial score S_s is defined as:

$$S_s = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left[\text{Match}_s(\hat{E}_{s,n}, E^*_{s,n}) = 1\right].$$

For region-level annotations, Match_s is computed by region overlap or region-tag agreement. For textual region descriptions, we map the description to the normalized region category provided by the annotation protocol and evaluate whether at least one selected spatial candidate matches this category.

The fine-grained score S_f is defined as:

$$S_f = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left[\text{Match}_f(\hat{E}_{f,n}, E^*_{f,n}) = 1\right].$$

A fine-grained prediction is considered correct only if it lies within the selected temporal-spatial scope and matches the annotated discriminative cue, such as a local tissue change, instrument-tissue interaction, anatomical landmark, or risk-related visual pattern.

Evaluation procedure. For each level, a sample is counted as a hit if at least one model-selected candidate matches its annotated clue under the rules above. Evidence Score is computed from model-selected evidence, without teacher-forced evidence:

Require: Predictions $\{(\hat{E}_{t,n}, \hat{E}_{s,n}, \hat{E}_{f,n})\}_{n=1}^{N}$ and annotations $\{(E^*_{t,n}, E^*_{s,n}, E^*_{f,n})\}_{n=1}^{N}$
Ensure: Evidence Score

for $\ell \in \{t, s, f\}$ **do**
  $h_\ell \leftarrow 0$
  for $n = 1, \ldots, N$ **do**
    $b_{\ell,n} \leftarrow \text{Match}_\ell(\hat{E}_{\ell,n}, E^*_{\ell,n})$
    $h_\ell \leftarrow h_\ell + b_{\ell,n}$
  end for
  $S_\ell \leftarrow h_\ell / N$
end for
return $(S_t + S_s + S_f)/3$

Valid Chain Rate. Valid Chain Rate (VCR) measures the proportion of samples for which all three stage answers are correct. For each sample n, we define the stage-wise answer correctness indicator as:

$$C^{\text{ans}}_n = \prod_{j=1}^{3} \mathbb{I}\left[\hat{A}_{j,n} = A^*_{j,n}\right].$$

The Valid Chain Rate is then defined as:

$$\mathrm{VCR} = \frac{1}{N}\sum_{n=1}^{N} C_n^{\mathrm{ans}}.$$

Each $\hat{A}_{j,n}$ is evaluated against $A^*_{j,n}$ using the answer candidates $\boldsymbol{O}_j$ for stage question $\boldsymbol{Q}_j$; the main-question answer is evaluated separately using $\boldsymbol{O}$. Thus, a sample is counted as valid only when all three stage-wise answers are correct for the same sample. This metric is applicable to all models that output stage-wise answers, including external MLLMs and Text-CoT.

Evidence matching is evaluated separately through Evidence Score. We compute it for models whose selected temporal, spatial, and fine-grained evidence can be matched to the annotated clues. Evidence matching is not included in Table 1 VCR, which measures joint correctness of the three stage answers.

C Baseline Prompt Templates and Answer Parsing Rules

Inputs. For baselines requiring generated stage-wise answers, each sample provides the surgical video $\boldsymbol{V}$, the main question and its candidates $(\boldsymbol{Q}, \boldsymbol{O})$, three stage questions with their respective candidates $\{(\boldsymbol{Q}_j, \boldsymbol{O}_j)\}_{j=1}^{3}$, and dataset-provided knowledge $\boldsymbol{K}$ when available. Each candidate set is displayed separately with its own option labels. Ground-truth evidence clues $\boldsymbol{C}$ are never provided at inference.

CoT Fine-Tuning. We fine-tune Qwen3-VL-8B with LoRA on the same inputs, using three-stage textual reasoning and stage-wise and final answers as training targets. Only the adapter parameters are updated; the pretrained backbone weights remain frozen. At inference, the model is instructed to provide brief reasoning for each stage, answer its corresponding stage question, and then answer the main question. It does not explicitly select temporal, spatial, or fine-grained evidence.

Output schema. The prompts request the following labeled fields; Direct-answer prompting omits the three `Reasoning` fields:

```
Stage 1 Reasoning:  <brief explanation>
Stage 1 Answer:  <answer under O_1>
Stage 2 Reasoning:  <brief explanation>
Stage 2 Answer:  <answer under O_2>
Stage 3 Reasoning:  <brief explanation>
Stage 3 Answer:  <answer under O_3>
Final Answer:  <answer under O>
```

For SurgCouple, the same four answer fields represent the outputs of its shared answer decoder. The model also outputs selected temporal clips, spatial regions, and fine-grained candidates. Its reasoning states $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ are latent representations, not generated textual explanations. Only model-selected evidence is used at inference; annotated clues are used for training supervision and evaluation.

Answer parsing. For multiple-choice questions, we extract answers from the labeled fields. Each stage answer is checked against its own candidate set $\boldsymbol{O}_j$, and the final answer against $\boldsymbol{O}$. A field is valid only if it specifies one unambiguous option. If the `Final Answer` field is absent, we extract an answer from the last sentence only when it contains an unambiguous option from $\boldsymbol{O}$. Missing stage answers are not inferred from the final answer. For questions without candidates, generated answers are normalized and compared with the reference using the annotation protocol's equivalence rules. A missing, ambiguous, out-of-set, irrelevant, or refusal response is counted as incorrect for the corresponding question. A missing or incorrect stage answer also makes that sample invalid for VCR.

Evidence and decoding. Baselines are assigned an Evidence Score only if they produce evidence that can be matched to the annotated clues under the shared evaluation rules. Text-CoT's free-form reasoning is used for qualitative analysis, not treated as an explicit evidence chain. For open-source models, we use greedy decoding. For closed-source APIs, we set temperature to zero when supported; otherwise, we use the default setting and report it in the implementation details.