

ProActIS: Boosting LLMs’ Proactive Information Seeking in Dynamic Clinical Workflows with a Lightweight Exam Proposer

Ruiyang Zhang¹, Yushi Feng¹, Liang Peng¹, Lequan Yu¹

¹School of Computing and Data Science,
The University of Hong Kong, Hong Kong SAR, China

Correspondence: lqyu@hku.hk

Abstract

Large language models (LLMs) have shown strong promise as clinical reasoning agents, yet most benchmarks assume complete evidence availability before reasoning begins. However, real clinical practice unfolds through dynamic clinical workflows with progressive disclosure of patient evidence, requiring clinicians to decide what evidence to acquire next from partial and evolving patient states. This mismatch exposes a bottleneck in proactive evidence acquisition for LLM-based clinical agents. To address this gap, we introduce **ProActIS**, a lightweight state-conditioned examination proposer for proactive evidence acquisition. Given an evolving patient state, ProActIS ranks standardized laboratory, microbiology, and radiology examinations, together with STOP, through structured patient-state decomposition and action-card encoding. We construct an MIMIC-IV benchmark with 81,960 hospital admissions and 860,099 turn-level patient-state/examination pairs, using explicit diagnosis/result filtering and retrospective leakage sensitivity analysis over 129 standardized examination actions plus STOP (130 actions). On full-bank next-examination ranking, ProActIS achieves 0.6028 Recall@1, 0.8868 Recall@5, and 0.7258 MRR, substantially outperforming frequency and LLM baselines. In LLM-assisted dynamic workflows, ProActIS substantially improves benchmark-level evidence coverage and documented-trajectory alignment for downstream agents, with downstream diagnostic accuracy approaching the corresponding all-information controls.

1 Introduction

Clinical reasoning rarely begins with a complete view of the patient (Kassirer, 1983). A clinician first observes an initial presentation, forms tentative hypotheses, and then gathers additional evidence through laboratory tests, microbiology studies, and imaging (Elstein and Schwarz, 2002). This iterative

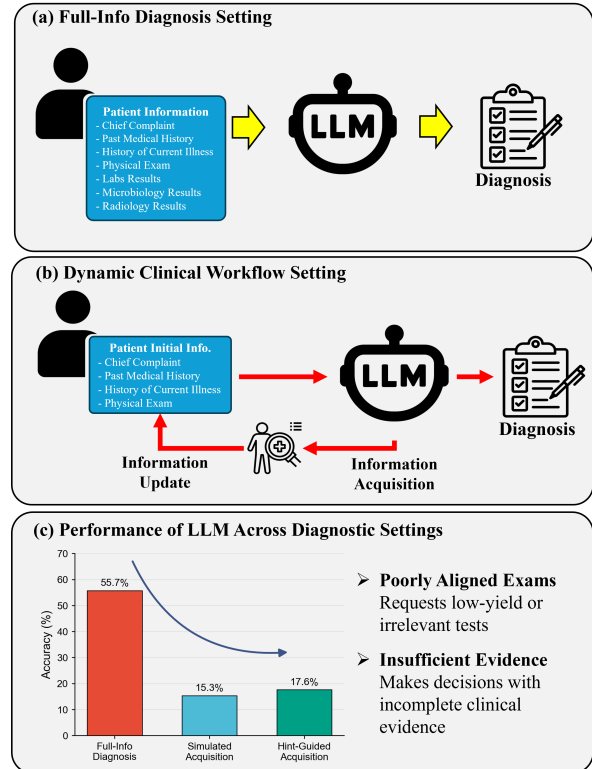


Figure 1: Static full-information diagnosis versus dynamic clinical evidence acquisition. Dynamic settings require the agent to request evidence before final diagnosis and expose a substantial drop in performance.

process is central to diagnosis: each new finding can confirm or rule out possible explanations (National Academies of Sciences, Engineering, and Medicine, 2015). While LLMs have shown strong performance on medical question answering and patient-record reasoning, these evaluations usually measure reasoning over evidence that has already been collected, rather than the ability to decide which evidence should be collected next.

When LLMs are placed in this dynamic setting, the difficulty shifts from only interpreting available evidence to actively acquiring evidence under a partial and evolving patient state. Specifically, the

same diagnostic LLM performs much better when complete patient information is provided than when it must request examinations sequentially. This gap suggests that dynamic failures are not only failures of final diagnostic reasoning, but also failures of proactive evidence acquisition. An agent may select plausible examinations while failing to recover examinations documented in the reference trajectory or stopping before the recorded endpoint.

Our reference trajectories represent examinations documented in retrospective EHRs. They provide behaviorally grounded supervision, but they are not physician-validated or counterfactual optimal policies. Accordingly, this paper studies observed next-examination prediction and recovery of documented examination trajectories, rather than clinical appropriateness or optimal evidence acquisition.

Prior work leaves this middle step underexplored. Medical QA and patient-record benchmarks emphasize reasoning from provided evidence (Schmidgall et al., 2026; Fleming et al., 2023), while interactive diagnosis methods often optimize dialogue behavior, disease prediction, or end-to-end task success (Fan et al., 2025). We instead focus on state-conditioned observed next-examination proposal before final diagnosis. We formulate this problem as turn-level ranking over a standardized examination action bank: the model observes the initial patient context at each turn, previous examination requests, and returned findings, then ranks candidate examinations together with the decision to stop evidence acquisition. This formulation isolates examination selection from downstream diagnosis generation while preserving the sequential and state-dependent nature of clinical evidence acquisition.

We introduce **ProActIS**, a lightweight state-conditioned examination proposer for LLM clinical agents. ProActIS encodes the evolving patient state and standardized action cards, then scores candidate examinations with a structured state-action ranking model. It handles STOP within the same action interface and uses trajectory-aware training and evaluation to accommodate ambiguity in valid examination order. Across offline next-examination alignment and assisted dynamic workflows, ProActIS improves evidence acquisition over frequent-exam and LLM baselines within the benchmark. As a standalone ranker, it better identifies reference-supported next examinations from the full action bank; as plug-in support for

LLM agents, it improves the coverage and alignment of dynamically requested examinations without replacing downstream diagnosis or stopping control. These results establish proactive information seeking as a distinct benchmark capability for clinical language agents, not as a validated clinical policy.

Our contributions are threefold:

1. We formulate proactive examination proposal as state-conditioned ranking over standardized examination actions and STOP, separating observed evidence acquisition from downstream diagnosis.
2. We build a patient-level-split MIMIC-IV observed-trajectory benchmark with 81,960 admissions and 860,099 turn-level state/action pairs, with explicit diagnosis/result filtering and retrospective leakage sensitivity analysis.
3. We propose ProActIS and evaluate its benchmark-level utility with matched offline baselines and assisted dynamic workflows, without treating observed trajectories as clinically optimal policies.

2 Related Work

Medical QA and fixed-context clinical reasoning. A growing body of research on medical LLMs and clinical agents has relied on question answering and examination-style benchmarks, including MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), MultiMedQA (Singhal et al., 2023), etc., to evaluate clinical and biomedical reasoning. These benchmarks assess medical knowledge, biomedical evidence interpretation, and reasoning over provided records, passages, or patient-specific contexts. However, they generally adopt a fixed-context formulation in which the evidence needed to answer the question is assumed to be available at inference time. Consequently, they evaluate a model’s ability to reason over given information rather than its ability to determine which additional evidence should be acquired. ProActIS addresses this complementary setting by studying proactive examination selection under partial and evolving patient information.

Contrastive learning for clinical representation and recommendation. Contrastive learning has been effective for longitudinal and heterogeneous patient modeling. EBCL (Oufattole

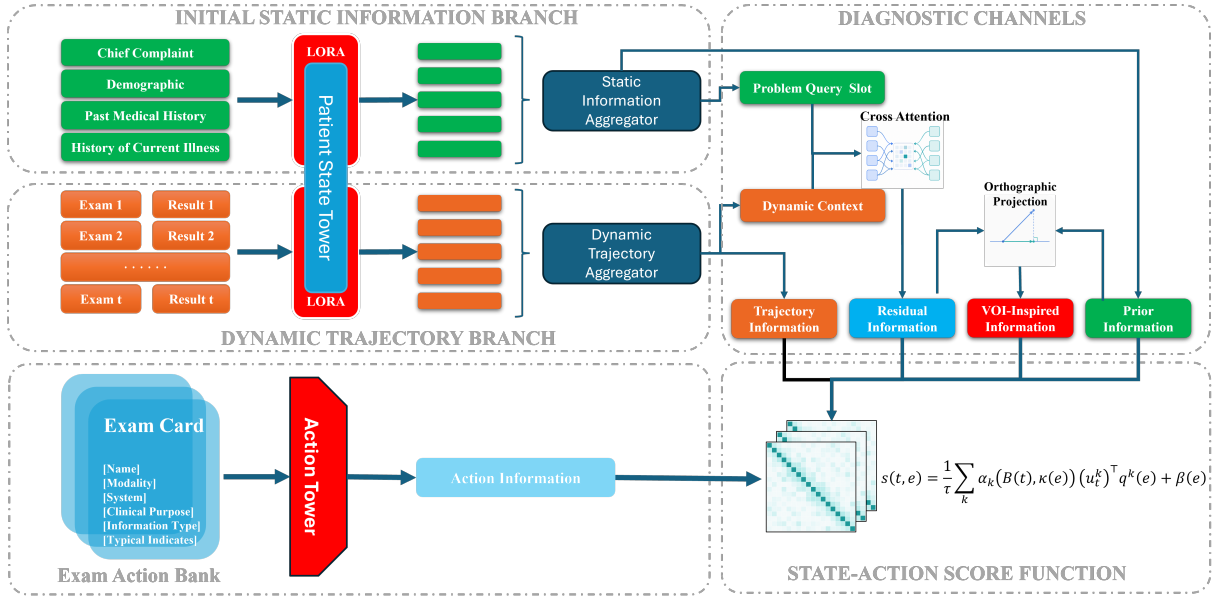


Figure 2: ProActIS architecture: structured state encoding and action-card encoding for unified exam/STOP ranking.

et al., 2024) shows that event-centered contrastive objectives can learn patient embeddings sensitive to temporal clinical progression, while ConMEHR (Yin et al., 2025) aligns structured and unstructured EHR signals in a shared latent space. ARCI (Hadizadeh Moghaddam et al., 2024) and DGCL (Li et al., 2023) apply contrastive learning to sequential prescription or medication recommendation, where models learn not only patient states but also their compatibility with candidate clinical actions. However, their action spaces mainly concern medications rather than diagnostic evidence acquisition. ProActIS addresses this gap by aligning evolving patient states with standardized examination concepts, making patient-state-to-exam matching the core learning problem.

Sequential diagnosis and evidence acquisition. Recent work has also studied diagnosis as an interactive or sequential decision process. AgentClinic (Schmidgall et al., 2026) provides a multimodal tool-using clinical-agent benchmark, AMIE (Tu et al., 2025) evaluates diagnostic dialogue, while LA-CDM (Bani-Harouni et al., 2025) and DiagAgent (Qiu et al., 2025) model multi-step clinical decision making through examination requests, result observation, and diagnosis in constrained or simulated environments. These approaches highlight the importance of evidence acquisition, but often optimize full dialogue behavior, simulated environment policies, or end-to-end diagnostic success. As a result, the compatibility

between a specific evolving patient state and each concrete examination action is usually implicit in the agent policy or constrained by a limited action space. ProActIS focuses on this explicit turn-level matching problem, enabling direct evaluation of examination selection and lightweight integration with LLM-based clinical agents without replacing downstream diagnosis.

3 Dataset Construction

We construct the dataset from MIMIC-IV structured records (Johnson et al., 2024) included laboratory, microbiology, and radiology records and MIMIC-IV-Note discharge summaries (Johnson et al., 2023). Each case corresponds to one hospital admission. To approximate information available near admission, the model-visible Patient Medical Record is limited to chief complaint, age at admit, race, past medical history, history of present illness, and physical exam. Because these admission-facing sections are extracted from discharge summaries, individual sentences may not be contemporaneous with the index time. Diagnosis fields are used only for cohort filtering and reference analysis.

We filter explicit diagnosis and direct-result leakage by removing discharge-oriented and result-revealing sections, including discharge diagnoses, hospital course, discharge medications, and reported laboratory or imaging results. Cases are further excluded when the final diagnosis is explicitly mentioned in visible patient fields, using string

matching, clinical synonym matching, and LLM-assisted review. This procedure does not guarantee that every admission-facing sentence was written contemporaneously or that all retrospective reconstruction has been removed.

As an exploratory residual-leakage audit, we reviewed all 5,998 HPI and physical-examination sentences in 100 test cases with a single LLM. The audit marked 151 sentences (2.52%) as result-revealing, 89 (1.48%) as post-index, and 335 (5.59%) as temporally ambiguous; 14 cases contained at least one post-index or result-revealing sentence. In a conservative admission-available-only sensitivity test on the same cases, Top-3 Micro Hit fell from 66.85% to 61.25% and downstream diagnostic accuracy from 57.00% to 54.79%. We treat this masking result as a robustness stress test rather than proof that retrospective leakage was eliminated; detailed screening rules and metrics are given in Appendix B.1.

Reference trajectories are derived from time-stamped laboratory, microbiology, and radiology records within a 24-hour window around the case index time. Raw item-level records were normalized into order-level examination actions using LLMs guided by the LOINC terminology (Regenstrief Institute, Inc., 2025); the resulting action set was manually reviewed and frozen across data splits.

The final action bank contains 129 standardized examination actions: 31 laboratory, 64 microbiology, and 34 radiology actions. We add a terminal STOP action, yielding 130 actions. Physical-examination findings may appear in the initial context, but physical-examination actions are not requestable actions in our benchmark because MIMIC-IV does not provide sufficiently standardized and temporally resolved bedside action-result pairs. Each trajectory is converted into turn-level examples where the state contains the initial patient context and prior examination-result history, and the target is the next reference action. The final dataset contains 81,960 cases from 60,039 patients and 860,099 turn-level pairs, split at the patient level into train, validation, and test sets with a 0.90/0.05/0.05 ratio. These trajectories are used as observed-practice references rather than gold-standard clinical policies.

Table 1: Dataset statistics after leakage filtering, action normalization, and turn-level conversion.

Statistic	Count
Patients	60,039
Cases / notes	81,960
Patient-state/action pairs	860,099
Lab action pairs	528,535
Microbiology action pairs	132,716
Radiology action pairs	116,888
STOP action pairs	81,960
Total actions, incl. STOP	130

4 Methodology

ProActIS is a lightweight state-conditioned exam proposer that ranks standardized examination actions or STOP from patient state. It implements as a contrastive dual-encoder model (Radford et al., 2021) with structured patient-state channels and action-card representations, trained to prefer the reference next action over competing legal candidates.

4.1 Problem Definition

At turn t , the state is $S_t = (S_0, p_t)$, where S_0 is the initial patient context and p_t is the ordered prefix of examination-result pairs. ProActIS ranks the legal candidate set $\mathcal{C}_t$ from the 130-action bank, including laboratory, microbiology, radiology, and STOP actions. The target e_{t+1}^* is the next action documented in the observed trajectory, not a physician-validated optimal action. The proposer ranks actions but does not interpret results or produce the final diagnosis.

4.2 Model Architecture

ProActIS is a lightweight contrastive dual encoder. The patient encoder maps S_t to four structured channels,

$$u_t = \{u_t^{\text{prior}}, u_t^{\text{traj}}, u_t^{\text{res}}, u_t^{\text{voi}}\},$$

representing the static context, observed trajectory, evidence-conditioned state, and remaining evidence gap. The action encoder maps each frozen, patient-independent action card into matching channel vectors. These channels are inductive biases rather than disentangled clinical variables; the detailed architecture is given in Appendix D.

4.3 State–Action Scoring and STOP

For channel k , the compatibility logit is

$$\begin{aligned} z_k(t, e) &= (u_t^k)^\top q^k(e), \\ s(t, e) &= \tau^{-1} \sum_k \alpha_k(B(t), \kappa(e)) z_k(t, e) \\ &\quad + \beta(e). \end{aligned}$$

where the channel weights are normalized over k and may depend on the workflow bucket $B(t)$ and action kind $\kappa(e)$. All legal examinations and STOP share this score space. STOP is therefore a ranked candidate and also receives an auxiliary observed-terminal-decision loss; no separate stopping threshold is used at inference.

4.4 Future-Aware Training and Inference

The immediate next action remains the only positive. Later actions in the same reference trajectory form a future-valid set and are protected from strong negative pressure, while hard negatives are mined from the remaining legal candidates. We optimize the exact and future-relaxed cross-entropies together with the invalid-candidate margin loss and auxiliary STOP loss:

$$\mathcal{L} = \mathcal{L}_{\text{ranker}} + \lambda_{\text{HN}} \mathcal{L}_{\text{HN}} + \lambda_{\text{STOP}} \mathcal{L}_{\text{STOP}}.$$

At inference, only S_0 , the observed prefix, and $\mathcal{C}_t$ are available; future labels and final diagnoses are never used.

5 Experiments and Results

5.1 Experimental Setup and Metric

We evaluate ProActIS in two settings. The primary setting is offline turn-level examination proposal, where each instance contains the current patient context, previous examination-result history, and the legal candidate set. The model ranks all candidates, including STOP, against the immediate next reference action. This setting directly measures state-conditioned proactive examination selection. We also evaluate ProActIS as plug-in support for an LLM-based diagnostic workflow, where the LLM receives either no proposer support or ProActIS Top-1/Top- K suggestions before requesting the next examination. This tests whether the proposer can improve dynamic evidence acquisition without producing the final diagnosis itself.

The main evaluation target is examination selection quality. For offline ranking, we report Recall@ K , MRR, and exact-or-future@1 over

turn-level states, with STOP included as a candidate and as the target at terminal turns. Recall@ K and MRR measure strict agreement with the immediate next reference action, but diagnostic trajectories are not always uniquely ordered: an examination predicted one step early may still appear later in the same documented reference trajectory. We therefore use exact-or-future@1 to separate clearly unsupported top predictions from temporally misordered but trajectory-consistent examinations. All offline metrics measure agreement with documented reference trajectories: exact Recall and MRR target the immediate recorded action, while exact-or-future measures whether a rank-1 action appears later in the same recorded trajectory. None of these metrics measures clinical appropriateness, necessity, safety, or cost-effectiveness. For dynamic workflows, we report hit rate, case-level recall, Jaccard overlap, ordered trajectory similarity, and matched reference examinations. Diagnosis accuracy is reported only as a downstream contextual outcome, not as clinical accuracy of the proposer.

5.2 Dynamic Workflow Evaluation

We evaluate three downstream LLM agents, DeepSeek V4 Pro (DeepSeek-AI et al., 2026), Qwen3.5-27B (Qwen Team, 2026), and Claude Sonnet 5 (Anthropic, 2026), under five settings: All information, Initial + Simulated Acquisition, Initial + Hint-Guided Acquisition, Initial + ProActIS Top-1, and Initial + ProActIS Top-3. All settings use the same cases, examination environment, acquisition prompt, request budget, and stopping interface. All-information supplies the complete patient record before diagnosis; the four dynamic settings begin with initial information and require sequential examination requests. When a requested examination is recorded, the environment returns its recorded result. Otherwise, Simulated Acquisition uses a World Model to synthesize feedback, whereas Hint-Guided Acquisition informs the downstream LLM that the examination was not performed. In ProActIS settings, the proposer supplies ranked candidates, but the downstream LLM retains final control over examination requests and diagnosis.

Table 2 first compares the two unguided dynamic baselines with the All-information control for each agent. Diagnosis accuracy falls from 55.7%, 50.50%, and 70.70% with complete information to 15.3%/17.6%, 11.00%/13.20%, and 19.90%/14.60% under Simulated/Hint-Guided Ac-

Table 2: Dynamic workflow performance under a fixed evaluation interface. For requests lacking recorded results, Simulated Acquisition uses a World Model to synthesize feedback, whereas Hint-Guided Acquisition signals that the examination was not performed. The recorded reference trajectories contain 7.97 examination requests per case on average.

Agent	Setting	Avg. requests	Micro Hit	Diag. accuracy	Judge
DeepSeek V4 Pro	All information	—	—	55.7%	0.599
	Initial + Simulated Acquisition	3.17	39.90%	15.3%	0.334
	Initial + Hint-Guided Acquisition	2.11	42.16%	17.6%	0.342
	Initial + ProActIS Top-1	5.93	64.55%	55.2%	0.595
	Initial + ProActIS Top-3	5.05	71.68%	56.5%	0.597
Qwen3.5-27B	Full information	—	—	50.50%	0.579
	Initial + Simulated Acquisition	4.64	29.37%	11.00%	0.263
	Initial + Hint-Guided Acquisition	2.56	27.98%	13.20%	0.213
	Initial + ProActIS Top-1	4.71	68.30%	46.80%	0.546
	Initial + ProActIS Top-3	3.53	85.79%	46.90%	0.568
Claude Sonnet 5	Full information	—	—	70.70%	0.700
	Initial + Simulated Acquisition	6.49	32.05%	19.90%	0.366
	Initial + Hint-Guided Acquisition	2.76	28.81%	14.60%	0.303
	Initial + ProActIS Top-1	7.28	61.54%	65.30%	0.646
	Initial + ProActIS Top-3	5.00	74.40%	64.80%	0.655

quisition for DeepSeek V4 Pro, Qwen3.5-27B, and Claude Sonnet 5, respectively. These baselines also request fewer examinations than the 7.97-request reference trajectories. Thus, under the common interface, interaction feedback without proposer guidance is insufficient to sustain evidence acquisition and downstream diagnosis.

Adding ProActIS changes both acquisition and diagnostic outcomes. Across all three agents, Top-1 and Top-3 increase Micro Hit substantially over the unguided baselines: from 39.90%/42.16% to 64.55%/71.68% for DeepSeek, from 29.37%/27.98% to 68.30%/85.79% for Qwen, and from 32.05%/28.81% to 61.54%/74.40% for Claude (Simulated/Hint-Guided versus Top-1/Top-3). Their diagnostic accuracies also recover to 55.2%/56.5%, 46.80%/46.90%, and 65.30%/64.80%, respectively, approaching the corresponding All-information controls. Top-3 generally provides higher Micro Hit than Top-1 while often using fewer requests, indicating that the coverage gain is not simply a consequence of longer trajectories.

The cross-agent comparison separates model capacity from acquisition support. Claude Sonnet 5 has the highest All-information accuracy (70.70%) and retains the highest absolute accuracy with ProActIS, while Qwen3.5-27B has the lowest control accuracy (50.50%) but shows the largest Micro Hit improvement. For Qwen and Claude, the best ProActIS settings recover approximately 92.9% and 92.4% of their All-information

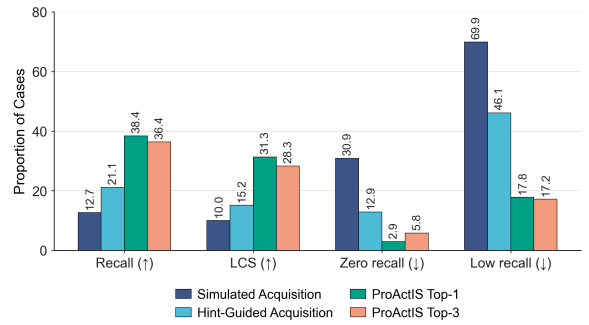


Figure 3: Case-level evidence recovery across dynamic workflow settings. ProActIS improves recall and ordered trajectory similarity while substantially reducing zero-recall and low-recall cases.

accuracies; DeepSeek reaches a comparable level with 55.2%/56.5% against a 55.7% control. ProActIS therefore improves evidence acquisition across downstream capability levels, while the final diagnostic level remains partly determined by the downstream model and by interpretation, confidence, and stopping decisions. These results support cross-agent utility under the fixed interface, not robustness to alternative prompts or stopping policies; full metrics and protocol details are provided in Appendix C.

Figure 3 further shows that ProActIS reduces case-level failures. Unguided settings leave many cases with zero or low recall, whereas ProActIS Top-1 achieves the best case-level recall, ordered trajectory similarity, and matched examinations, and Top-3 maintains low failure rates with stronger

Table 3: Offline full-bank next-exam ranking. E/F@1 denotes exact-or-future validity at rank 1.

Model	E/F@1	R@1	R@3	R@5	MRR
Frequent	0.1864	0.0954	0.2817	0.3776	0.2038
MedGemma-27B	0.4244	0.1329	0.3036	0.3866	0.2263
Baichuan-M2 32B	0.4123	0.1051	0.2753	0.3979	0.2046
Qwen3.5 27B	0.4642	0.1411	0.3459	0.4885	0.2598
DeepSeek V4 Pro	0.4411	0.1378	0.3450	0.4891	0.2582
ProActIS	0.8245	0.6028	0.8169	0.8868	0.7258

local hit rates. Overall, ProActIS makes dynamic workflows more aligned with recorded actions at each step and more complete relative to documented trajectories.

5.3 Turn-Level Next-Exam Alignment

We next evaluate whether each method can select the immediate next reference examination at the turn level. To make the comparison fair, all LLM baselines (Qwen Team, 2026; Sellergren et al., 2025; Baichuan-M2 Team et al., 2025; DeepSeek-AI et al., 2026) are given the standardized action bank and asked to choose within the same action space. This setting therefore tests next-action selection under a shared interface, rather than comparing constrained ranking with unconstrained free-form generation.

The top-frequency baseline performs poorly, showing that next-examination selection is not a global popularity problem. The LLM baselines perform better than frequency ranking, but their exact next-action recall remains low even with the standardized action bank provided. ProActIS substantially outperforms all baselines, reaching 0.6028 Recall@1, 0.8868 Recall@5, and 0.7258 MRR. Its Exact-or-future@1 of 0.8245 shows that its gains are not limited to strict immediate-action matching, but also improve trajectory-relative validity.

5.4 Matched Offline Baselines

To separate the effect of the overall ranking formulation from differences in data or action space, we additionally compare ProActIS with two supervised baselines under matched conditions. All three models use the same patient-level split, the same 860,099 turn-level states, the same 130-action inventory (including STOP), the same legal-action constraints, and the same observed next-action core supervision. The BioLinkBERT (Yasunaga et al., 2022) classifier encodes the state with a pooled representation and performs flat 130-way classification. SFT Qwen3.5-4B (Qwen Team, 2026) receives the immediate and future- K actions and

Table 4: Matched offline baselines under the shared patient split, trajectory states, 130-action inventory, legal-action constraints, and observed next-action supervision. STOP@1 is the accuracy of the observed terminal decision.

Model	R@1	R@3	STOP@1
BioLinkBERT classifier	0.3616	0.4943	0.7621
SFT Qwen3.5-4B	0.4934	0.5489	0.6823
ProActIS	0.6028	0.8169	0.9621

Table 5: Cumulative component ablation.

Ablation	R@1	R@5	MRR	STOP@1
Single-vector	0.3092	0.6183	0.4489	0.8948
Four-channel	0.5019	0.8073	0.6355	0.6327
+ Hard neg.	0.5427	0.8654	0.6817	0.7440
+ STOP-aware	0.5441	0.8647	0.6830	0.9093
+ Channel calib.	0.5907	0.8781	0.7152	0.9543
+ Future-aware	0.6028	0.8868	0.7258	0.9621

directly generates a ranked Top- K list, without repeated sampling, beam search, or answer merging.

Under these matched conditions, ProActIS leads both baselines on every reported metric. Relative to the BioLinkBERT classifier, it improves R@1, R@3, and STOP@1 by 24.12, 32.26, and 20.00 percentage points; relative to SFT Qwen3.5-4B, the corresponding gains are 10.94, 26.80, and 27.98 points. The Qwen baseline uses future- K supervision and direct ranked generation, yet its R@3 remains below ProActIS, indicating that generating multiple candidates alone does not ensure reliable ranking. These results support an end-to-end advantage for the structured ranking formulation under matched evaluation, without attributing the gains to any single component or implying clinical optimality.

5.5 Component Ablation

Table 5 shows the cumulative contribution of the main design choices. Replacing a single state vector with four channels gives the largest early gain; channel calibration and future-aware supervision then provide complementary improvements, while STOP-aware training mainly improves terminal-decision recovery.

Table 6 reports leave-one-channel-out ablations for the full model. Removing any channel reduces performance; the VOI-inspired channel has the largest average drop, especially in R@1 and MRR.

5.6 Calibration and Uncertain States

We assess confidence with respect to the observed next action rather than clinical correctness. A sin-

Table 6: Selected leave-one-channel-out results.

Removed	R@1	R@5	MRR
None	0.6028	0.8868	0.7258
Prior	0.5647	0.8612	0.6912
Trajectory	0.5316	0.8452	0.6652
Residual	0.5294	0.8423	0.6619
VOI-inspired	0.4957	0.8585	0.6490

Table 7: Confidence-stratified offline alignment on non-terminal test states. E/F denotes exact-or-future validity; E/F gain is the Top-3 minus Top-1 difference in percentage points.

Confidence group	MSP	Exact@1	Exact@3	E/F@1	E/F@3	E/F gain (pp)
Q1, highest	0.9951	99.33%	99.74%	99.82%	99.93%	+0.11
Q2	0.8571	83.05%	94.77%	96.10%	99.36%	+3.25
Q3	0.6041	58.64%	86.20%	88.41%	97.53%	+9.12
Q4	0.4101	39.65%	75.56%	76.72%	93.17%	+16.45
Q5, lowest	0.2343	23.38%	53.91%	51.30%	78.54%	+27.24

gle temperature was fitted on a held-out validation calibration split and then applied to the untouched test set. This reduces top-label exact-action ECE from 0.0408 to 0.0124 without changing the ranking. To examine when candidate expansion is useful, we divide non-terminal test states into equal-sized quintiles by calibrated maximum softmax probability (MSP). The calibration procedure and bootstrap details are given in Appendix C.2.

Exact@1 and E/F@1 decline monotonically as confidence decreases, while the Top-1-to-Top-3 E/F gain rises from 0.11 points in Q1 to 27.24 points in Q5. The Q5 gain has a patient-bootstrap 95% confidence interval of [26.12, 28.32]. Top-3 rescues 55.93% of states whose Top-1 prediction is unsupported, but 21.46% of Q5 states still contain no exact or future-valid action in their Top-3 set. Candidate expansion therefore reduces, but does not eliminate, unsupported proposals.

We also select confidence thresholds for an adaptive $K \in \{1, 2, 3, 4, 5\}$ policy using only the validation set and apply them unchanged to the test set. Adaptive- K obtains comparable E/F coverage to fixed Top-3 (93.78% versus 93.70%) while returning 2.30 rather than 3.00 candidates on average. It does not improve downstream diagnostic accuracy, indicating that efficient candidate allocation alone does not resolve downstream interpretation or stopping errors. These analyses establish calibration for observed next-action alignment, not calibration of clinical necessity or correctness.

Table 8: Common versus rare target examinations on the untouched test set, excluding terminal STOP decisions. Rare targets use the predefined frequency threshold of $\leq 5\%$. Unsupported@3 is one minus E/F@3.

Target frequency	Exam turns	Exact@1	E/F@3	Unsupported @3
Common	36,170	61.63%	95.88%	4.12%
Rare ($\leq 5\%$)	1,926	28.50%	76.85%	23.15%

5.7 Subgroup and Rare-Examination Analysis

We evaluate subgroup variation on the untouched test set after excluding terminal STOP decisions, yielding 38,096 examination decisions from 4,000 cases. The full discharge-service and mapped-diagnosis breakdowns are provided in Appendix E. Across discharge-service proxies, Exact@1 ranges from 52.30% to 64.00% and E/F@3 from 91.83% to 95.22%; across mapped diagnosis categories, the corresponding ranges are 53.62%–63.33% and 90.64%–95.61%. Emergency and non-emergency admissions have similar E/F@3 coverage (92.86% versus 93.42%), with Exact@1 of 55.69% and 58.74%, respectively.

Rare targets lose 33.13 percentage points in Exact@1 and 19.03 points in E/F@3 relative to common targets, while their unsupported Top-3 rate rises from 4.12% to 23.15%. Broad subgroup coverage is therefore comparatively stable, but rare and indication-specific examinations remain the main weakness. Frequency imbalance and sparse supervision can cause these actions to be under-ranked, and the aggregate metrics are weighted toward common actions rather than representing uniform reliability across the action bank. Because service is only a specialty proxy and 29.5% of mapped diagnoses are Other/unspecified, these are descriptive benchmark findings rather than evidence of equal clinical validity across subgroups.

5.8 Error and Recommendation-Bias Audit

The strict Top-1 prediction differs from the immediate observed action on 39.72% of decisions. Among these errors, 55.83% are temporal-order errors: the recommended action appears later in the same observed trajectory. Strict next-action error therefore overstates unsupported recommendation error, although a later recorded action should not be treated as clinically interchangeable without physician review. Among genuinely trajectory-unsupported Top-1 errors, the observed action still appears in the Top-5 in 61.61% of cases and in the

Top-10 in 80.87%, locating much of the remaining failure in ranking rather than complete candidate retrieval.

Qualitative error types. The audit identifies three recurring patterns. First, fine-grained action confusion substitutes non-equivalent tests, such as Urine Panel for Urine Culture or one CT protocol for another. Second, long-tail under-prioritization favors a frequent XR over the indication-specific DXA, which appears only ten times in training. Third, premature STOP occurs in multimorbid cases with unresolved evidence needs, such as chronic kidney disease, cellulitis, and acute renal deterioration. These patterns suggest distinct remedies—action-card refinement, rare-action reweighting, and stronger STOP calibration—rather than a single ranking-noise explanation.

We also compare recommendation counts with the recorded trajectories to audit concentration and modality bias. ProActIS predicts 20,536 common-panel actions versus 18,693 in the reference trajectories (1.099 times the reference rate per turn), but case coverage is nearly unchanged (3,972 versus 3,940 cases; ratio 1.008). It predicts fewer unique common actions per case (4.153 versus 4.673; ratio 0.889), so the excess is more consistent with repetition across decision turns than with substantially broader patient coverage. Advanced-imaging predictions are lower than the reference both per turn (2,063 versus 2,386; ratio 0.865) and by covered cases (1,512 versus 1,858; ratio 0.814). Thus, the audit does not support systematic high-cost imaging over-recommendation relative to recorded practice, but the opposite pattern may reflect under-ranking of rare or indication-specific imaging. All comparisons are deviations from documented utilization, not measures of clinical overuse, necessity, or cost-effectiveness; counting definitions are given in Appendix C.4.

5.9 Premature Stopping Case Study

This held-out case illustrates premature stopping, not clinical adjudication or representative behavior. The admission-facing context describes a 78-year-old woman with peripheral arterial disease, chronic kidney disease, atrial fibrillation on anticoagulation, anemia, cirrhosis, and heart failure who presents with right-leg ischemia, erythema, severe pain, and ulceration. The recorded diagnosis and examination trajectory serve only as a post hoc comparison, not as physician-validated targets.

The trajectory contains nine standardized actions spanning vascular, laboratory, coagulation, metabolic/hepatic, urinary, and imaging categories. Unguided downstream agents requested Doppler followed by CT and stopped after two requests; CT received a simulated result in Simulated Acquisition but was unmatched in Hint-Guided Acquisition. These annotations do not imply clinical invalidity or unavailability. With ProActIS Top-1 or Top-3, the agent made six requests and overlapped five distinct recorded actions. Support extended acquisition beyond the chief complaint, but neither recovered the full trajectory.

The recorded diagnosis field includes peripheral arterial disease, cellulitis, acute-on-chronic renal failure, heart failure, and urinary incontinence. Unguided outputs emphasized lower-extremity ischemia and peripheral arterial disease and added venous-disease or negative-DVT terms absent from the recorded labels, whereas ProActIS outputs included peripheral arterial disease and cellulitis, plus renal dysfunction or tissue-loss/ulceration terms depending on Top- K . We report overlap with recorded labels, not clinical correctness or hallucination rates. This is consistent with aggregate evidence-coverage gains while showing that candidate expansion does not guarantee per-case superiority: Top-3 did not increase requests or recorded-action overlap here. ProActIS ranks examinations and STOP, while the downstream agent controls requests, interpretation, diagnosis, and termination; the example therefore demonstrates evidence and trajectory alignment only, not improved diagnosis, appropriateness, or safer deployment. It cannot establish clinical utility, patient benefit, or reduced hallucination in deployment.

6 Conclusion

ProActIS provides a lightweight way to rank examinations and STOP actions against evolving patient states, improving recovery of documented examination trajectories and evidence coverage for downstream agents in our benchmark. These findings support proactive examination proposal as a measurable workflow component, while physician-validated, counterfactual, prospective, and multi-institution studies remain necessary before making claims about clinical appropriateness or optimal policy.

Limitations

This work has several limitations. First, reference trajectories are recorded orders, not counterfactual or physician-validated optimal policies. They may reflect institutional practice, availability, billing, workflow, and clinician habits; alignment metrics are therefore observational proxies rather than measures of clinical appropriateness or patient benefit.

Second, residual retrospective leakage remains possible. HPI and physical-examination text is derived from discharge summaries and may contain post-index, result-revealing, or temporally ambiguous information. Our single-LLM audit and 100-case admission-only masking analysis provide exploratory sensitivity evidence, not proof that retrospective leakage is absent.

Third, physical-examination findings may be available as initial context, but physical-examination actions are not requestable in the benchmark. The resulting action space therefore omits an important part of the clinical information-seeking process and may understate both achievable performance and failure modes.

Fourth, rare and highly indication-specific examinations are underrepresented and substantially harder to recover. Aggregate results are consequently weighted toward common actions and should not be interpreted as uniform reliability across the action bank, particularly for rare or indication-specific requests.

Fifth, neither ProActIS nor the automated trajectory audits have physician, prospective, or counterfactual validation. They cannot establish examination appropriateness, necessity, sufficiency, safety, or cost-effectiveness; the LLM-based analyses provide complementary failure characterization, not clinician adjudication.

Sixth, ProActIS ranks candidate examinations and STOP actions; it does not itself execute requests or determine when the workflow terminates. ProActIS can under-rank an observed examination or rank STOP before the end of the recorded trajectory. In dynamic workflows, the downstream agent retains final control of examination requests and stopping, so observed failures cannot be attributed to the proposer alone.

Finally, end-to-end outcomes depend on the downstream LLM, simulator or feedback mechanism, prompt, request budget, and stopping interface. The cross-agent results therefore provide evidence of limited transfer under this fixed eval-

uation interface, not robustness to arbitrary agent policies or deployment settings.

Acknowledgments

This work was supported in part by the Research Grants Council of Hong Kong (27206123, 17200125, C5055-24G, and T45-401/22-N), and the Hong Kong Innovation and Technology Fund (ITS/544/24FP).

References

- Anthropic. 2026. [Introducing claude sonnet 5](#). Published: 2026-06-30; accessed: 2026-07-10.
- Baichuan-M2 Team, Chengfeng Dou, Chong Liu, et al. 2025. [Baichuan-m2: Scaling medical capability with large verifier system](#). *Preprint*, arXiv:2509.02208.
- David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Nassir Navab, and Matthias Keicher. 2025. Language agents for hypothesis-driven clinical decision making with reinforcement learning. *arXiv preprint arXiv:2506.13474*.
- DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. [Deepseek-v4: Towards highly efficient million-token context intelligence](#). *Preprint*, arXiv:2606.19348.
- Arthur S. Elstein and Alan Schwarz. 2002. [Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature](#). *BMJ*, 324(7339):729–732.
- Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. [AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.
- Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo P. Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju S. Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott J. Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, and 11 others. 2023. [Medalign: A clinician-generated dataset for instruction following with electronic medical records](#). *Preprint*, arXiv:2308.14089.
- Google DeepMind. 2026. [Gemini 3.1 Pro Model Card](#). Published: 2026-02-19; accessed: 2026-05-26.

- Arya Hadizadeh Moghaddam, Mohsen Nayeibi Kerdabadi, Mei Liu, and Zijun Yao. 2024. [Contrastive learning on medical intents for sequential prescription recommendation](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, page 748–757. ACM.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.
- Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. [MIMIC-IV](#). *PhysioNet*. Version 3.1.
- Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. [MIMIC-IV-Note: Deidentified free-text clinical notes](#). *PhysioNet*. Version 2.2.
- Jerome P. Kassirer. 1983. [Teaching clinical medicine by iterative hypothesis testing: Let's preach what we practice](#). *New England Journal of Medicine*, 309(15):921–923.
- Xingwang Li, Yijia Zhang, Xiaobo Li, Hao Wei, and Mingyu Lu. 2023. Dgcl: distance-wise and graph contrastive learning for medication recommendation. *Journal of Biomedical Informatics*, 139:104301.
- National Academies of Sciences, Engineering, and Medicine. 2015. [Improving Diagnosis in Health Care](#). The National Academies Press, Washington, DC.
- OpenAI. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- Nassim Oufattole, Hyewon Jeong, Matthew B. A. McDermott, Aparna Balagopalan, Bryan Jangeesingh, Marzyeh Ghassemi, and Collin Stultz. 2024. [Event-based contrastive learning for medical time series](#). In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. PMLR.
- Pengcheng Qiu, Chaoyi Wu, Junwei Liu, Qiaoyu Zheng, Yusheng Liao, Haowen Wang, Yun Yue, Qianrui Fan, Shuai Zhen, Jian Wang, Jinjie Gu, Yanfeng Wang, Ya Zhang, and Weidi Xie. 2025. [Evolving diagnostic agents in a virtual clinical environment](#). *Preprint*, arXiv:2510.24654.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Regenstrief Institute, Inc. 2025. [LOINC \(Logical Observation Identifiers Names and Codes\)](#). Accessed: 2026-05-26.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. 2026. [Agentclinic: A multimodal benchmark for tool-using clinical AI agents](#). *npj Digital Medicine*, 9:499.
- Andrew SELLERGRÉN, Sahar Kazemzadeh, Tiam Jaroensri, et al. 2025. [Medgemma technical report](#). *arXiv preprint arXiv:2507.05201*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450.
- Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. [LinkBERT: Pretraining language models with document links](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.
- Qing Yin, Linda Zhong, Yunya Song, Liang Bai, Zhihua Wang, Chen Li, Yida Xu, and Xian Yang. 2025. A decision support system in precision medicine: contrastive multimodal learning for patient stratification. *Annals of Operations Research*, 348(1):579–607.

A Declaration of Generative AI Usage

LLMs were used in several limited, documented roles: language polishing; proposing candidate mappings and action-card text during action-bank

construction; assisting sentence-level retrospective-leakage screening; and performing the complementary automated trajectory audit reported in this appendix. The authors designed the study, selected the evaluation protocols, verified the outputs, and made all final scientific decisions. LLM assistance did not constitute physician adjudication or prospective clinical validation. LLM-generated action cards are semantic representations for the benchmark, not clinical guidelines. For visual communication, AI tools were used only to generate individual flowchart icons; the architecture, layout, and scientific data visualizations were designed or generated from raw data by the authors. All AI-assisted outputs were reviewed and curated by the authors before inclusion.

B Dataset Construction Details

Leakage Control We retained only admission-facing sections from each discharge summary as model-visible patient context. Discharge diagnosis, hospital course, discharge medications, and sections containing reported laboratory or imaging results were removed because they may reveal downstream diagnostic conclusions or examination outcomes. The Diagnosis field was retained only for cohort filtering and reference analysis. Cases were further screened for explicit diagnosis mentions in Past Medical History and History of Present Illness using string matching, clinical synonym matching, and DeepSeek-assisted binary leakage review before training.

B.1 Retrospective Leakage Audit

We conducted an exploratory sentence-level audit on 100 untouched test cases. A single LLM reviewed all 5,998 HPI and physical-examination sentences and assigned mutually non-exclusive risk flags for explicit diagnosis leakage, result-revealing content, post-index information, and temporal ambiguity. No sentence was marked as explicit diagnosis leakage; 151 (2.52%) were result-revealing, 89 (1.48%) were post-index, and 335 (5.59%) were temporally ambiguous. A case was flagged when it contained at least one post-index or result-revealing sentence; 14 of 100 cases were flagged.

As a conservative sensitivity analysis, we masked all sentences not classified as admission-available in the same 100 cases and reran the dynamic workflow without retraining. The full and masked conditions therefore differ in visible con-

Table 9: Sentence-level retrospective-leakage audit over 5,998 HPI and physical-examination sentences from 100 test cases. Flags are exploratory and may overlap.

Risk flag	Sentences	Share
Explicit diagnosis leakage	0	0.00%
Result-revealing	151	2.52%
Post-index	89	1.48%
Temporally ambiguous	335	5.59%

text and may also introduce a train–test distribution shift. The masking comparison is a robustness stress test, not evidence that leakage has been removed.

Masking reduces Top-3 Micro Hit by 5.60 points and downstream diagnostic accuracy by 2.21 points. Top-3 still exceeds Top-1 by 7.68 Micro Hit points and 2.61 diagnostic-accuracy points under masking, compared with +5.84 and +2.36 points under full context. Because a single-LLM audit cannot establish temporal provenance and masking can remove valid but ambiguous admission information, these results quantify residual sensitivity rather than demonstrate leakage-free evaluation.

Action Bank Construction Raw laboratory tests, microbiology tests, and radiology records were first collected within a 24-hour window centered on the case index time and ordered chronologically. Because MIMIC-IV records are often item-level rather than order-level, raw items and studies were normalized into order-level examination actions. We used LOINC codes, examination metadata, and clinical semantics to propose mappings from raw records to standardized actions. Candidate mappings were reviewed with GPT-5 (OpenAI, 2026) and Gemini 3.1 Pro (Google DeepMind, 2026), manually inspected, and then frozen. The frozen mapping was applied deterministically across all splits.

The final action bank contains 129 standardized examination actions, grouped into 31 laboratory, 64 microbiology, and 34 radiology actions. A terminal STOP action was added to represent the end of evidence acquisition. DeepSeek was used to assist action-bank and action-card generation. Each action card contains the examination name, bucket, parent ontology, modality, body system, clinical purpose, information type, and typical indications. These cards provide examination semantics to the action encoder rather than serving as one-hot labels. They are not clinically validated guidelines and contain no patient-specific information or case-

Table 10: Admission-available-only sensitivity analysis on the same 100 cases. The masked condition retains only sentences classified as admission-available; no retraining is performed. Percent-valued columns are shown with percent signs.

Guidance	Initial context	Avg. requests	Micro Hit	Macro Hit	Jaccard	Diag. accuracy	Judge
Top-1	Full	5.71	61.01%	59.53%	35.85%	54.64%	0.585
Top-1	Admission-available only	5.93	53.57%	53.27%	32.59%	52.18%	0.506
Top-3	Full	5.34	66.85%	69.64%	36.47%	57.00%	0.588
Top-3	Admission-available only	5.91	61.25%	63.78%	37.22%	54.79%	0.562

specific labels.

Physical-examination findings may remain in the initial patient context, but physical-examination actions are not requestable in the benchmark because the source data do not provide sufficiently standardized and temporally resolved bedside action-result pairs. The action bank therefore covers laboratory, microbiology, radiology, and STOP actions, not the full clinical information-seeking process.

Turn-Level Example Construction Each raw examination record was mapped to one standardized action. Admissions without mapped examinations were removed. Repeated standardized actions within the same case were collapsed to their first occurrence to avoid over-weighting repeated item-level measurements. Each case trajectory was then converted into turn-level examples: at each turn, the model state includes the initial patient context and the ordered history of previously selected examinations with returned results, while the target is the next reference action. STOP is appended as the final target for each case.

C Additional Evaluation Details

C.1 Dynamic Workflow Protocol

The cross-agent experiments use the same cases, examination environment, acquisition prompt, request budget, and stopping interface as the original DeepSeek workflow. The full-information condition provides the complete patient record before diagnosis; dynamic conditions begin with initial context and reveal examination feedback only after a request. Simulated Acquisition supplies the same interaction feedback without proposer guidance, Hint-Guided Acquisition supplies an unmatched-request hint, and ProActIS conditions provide ranked state-conditioned proposals. The downstream agent retains final control over examination requests, result interpretation, diagnosis, and termination. The core outcomes for all three agents are reported in the main-text dynamic workflow table (Table 2); the remaining Appendix sections

provide audit, calibration, counting, model, and subgroup details.

C.2 Calibration Protocol Details

The temperature is fitted only on a held-out validation calibration split and then frozen for test evaluation; the test split is not used to select the temperature. Non-terminal test states are partitioned into equal-sized groups by calibrated maximum softmax probability (MSP). E/F gain is computed as the Top-3 E/F rate minus the Top-1 E/F rate within each group. The interval for the Q5 gain is obtained by patient-level bootstrap resampling. For adaptive- K , validation-only MSP thresholds map states to $K \in \{1, 2, 3, 4, 5\}$ and are frozen before test evaluation; the resulting candidate count is averaged over test states. ECE and these stratified metrics measure alignment with the observed next action, not clinical necessity, correctness, or safety.

C.3 Automated Whole-Trajectory Audit

As a complementary non-clinical audit, Claude Sonnet 5, GPT-5, and Gemini 3.1 Pro independently evaluated the same 100 cases under three trajectory conditions: the recorded observed trajectory, a DeepSeek V4 Pro rollout with ProActIS Top-1 support, and a rollout with ProActIS Top-3 support. In the truth-blind setting, the reference final diagnosis was omitted; in the outcome-aware setting, it was provided. Continuous criteria use five-point scales. “Acceptable” and “Outcome Adequate” require agreement from at least two of the three evaluators. The observed trajectory is treated as another candidate trajectory, not as a clinically optimal policy.

These audits show that Top-3 improves several metrics over Top-1, but remains below the observed trajectory on outcome-aware measures and does not establish clinical appropriateness or safety. The evaluators may share systematic biases, and automated audit cannot replace blinded physician adjudication.

Table 11: Truth-blind automated whole-trajectory audit on 100 cases. The reference final diagnosis was omitted. Higher is better for Acceptable, Overall, Coherence, and Diagnosis Supported; lower is better for Major Omission and Premature STOP.

Setting	Acceptable	Overall	Coherence	Diagnosis Supported	Major Omission	Premature STOP
Observed	77.0%	3.81	3.90	98.0%	11.0%	12.0%
ProActIS Top-1	65.0%	3.47	3.60	85.0%	22.0%	24.0%
ProActIS Top-3	69.0%	3.73	3.99	90.0%	19.0%	19.0%

Table 12: Outcome-aware automated whole-trajectory audit on 100 cases. The reference final diagnosis was provided. Higher is better for Outcome Adequate, Reference Supported, Evidence Sufficiency, and Critical Coverage; lower is better for Major Omission and Premature STOP.

Setting	Outcome Adequate	Reference Supported	Evidence Sufficiency	Critical Coverage	Major Omission	Premature STOP
Observed	63.0%	87.0%	3.96	3.66	39.0%	38.0%
ProActIS Top-1	52.0%	80.0%	3.25	2.66	51.0%	48.0%
ProActIS Top-3	53.0%	82.0%	3.39	2.73	43.0%	44.0%

C.4 Error and Recommendation-Bias Counting Rules

A temporal-order error is counted when the Top-1 action differs from the immediate observed action but appears later in the same trajectory. A Top-1 error is trajectory-unsupported when the observed action appears nowhere later in that trajectory. Common-panel turn counts include every predicted or reference common-panel action at a non-terminal decision; case coverage counts unique cases with at least one such action; unique common actions per case counts distinct action identities within each case. Advanced-imaging counts use the same definitions. These rules yield the turn-level and case-level ratios reported in the main text. The qualitative examples (Urine Panel versus Urine Culture, alternative CT protocols, XR versus DXA, and premature STOP in multimorbid cases) are failure-characterization examples rather than clinical adjudications.

D Method Detail

This appendix specifies implementation details omitted from the high-level model description in Section 4; notation follows that section.

Model and Training Details. We initialize the text encoder with BioLinkBERT (Yasunaga et al., 2022) to encode patient states, examination-result histories, and action-card descriptions. To reduce trainable parameters and stabilize optimization, we apply LoRA-based (Hu et al., 2021) parameter-efficient fine-tuning to the backbone encoder while training the task-specific projection layers, channel calibration parameters, action bias, STOP head, and ranking objectives. The model is trained with

full-bank candidate scoring over the standardized action set, including laboratory, microbiology, radiology, and STOP actions. All main training runs are conducted on 8 NVIDIA H800 GPUs with distributed data-parallel training.

Compute Cost and Hyperparameters. Training for 10 epochs required approximately 60 GPU-hours on 8 GPUs; we used batch size 14 per GPU, bfloat16 with FlashAttention-2, embedding dimension 256, dropout 0.1, AdamW with base learning rate 2×10^{-5} , LoRA rank 8 with $\alpha = 16$ and dropout 0.05, temperature 1.0, max 8 static chunks of length 512 with overlap 64, max 32 events, max event/action length 512/256, hard-negative $K = 5$, rank margin 0.3, rank loss weight 0.05, STOP sufficiency loss weight 0.05, future loss weight 0.05 with denominator discount 0.25, class-balance strength 0.5 with weights clipped to $[0.5, 3.0]$, structured channel calibration and global action bias enabled, and seed 13.

Matched baseline configurations. The BioLinkBERT classifier uses the same BioLinkBERT backbone, attention pooling, and a linear 130-way classification head, trained for four epochs with multiclass cross-entropy. SFT Qwen3.5-4B is instruction-tuned on the same turn-level state/action pairs with causal language modeling and token-level cross-entropy. It uses BF16 LoRA, per-device batch size 2, gradient accumulation of 4, and a reported global batch size of 32; training stops after 7,680 steps, with a final training loss of 0.0345. Both baselines use the standardized action bank and legal-action masking at evaluation time, and the generative baseline is constrained to return actions from that bank.

Structured Ranker Details. The main text defines the four state channels and the shared state-action score. The implementation details below specify how the static, dynamic, residual, and VOI-inspired representations are instantiated.

The static branch encodes the initial patient information S_0 into a global static representation h_{static} and static problem slots Z_{static} . The prior channel is obtained by projecting the global representation:

$$u_t^{\text{prior}} = W_{\text{prior}} h_{\text{static}}.$$

The dynamic branch encodes the ordered examination-result prefix p_t . Each examination-result pair (e_i, r_i) is encoded as an event representation, positional information is added to preserve order, and a sequence encoder produces an event memory M_t and pooled trajectory representation h_{traj} . The trajectory channel is

$$u_t^{\text{traj}} = W_{\text{traj}} h_{\text{traj}}.$$

For the first decision step, where $t = 0$, no examination-result event has been observed. We use a learned empty-prefix representation so that trajectory-dependent components remain well-defined.

The residual channel is produced by static-dynamic cross-attention. ProActIS uses static problem slots as queries and trajectory event memory as keys and values:

$$A_t = \text{MHA}(Q = Z_{\text{static}}, K = M_t, V = M_t).$$

The attended evidence is combined with the original static slots through a gated update:

$$G_t = \sigma \left(W_g [Z_{\text{static}}; A_t; Z_{\text{static}} \odot A_t; Z_{\text{static}} - A_t] \right).$$

$$\Delta_t = W_{\Delta} [Z_{\text{static}}; A_t; Z_{\text{static}} \odot A_t],$$

$$R_t = \text{LayerNorm} (Z_{\text{static}} + G_t \odot \Delta_t).$$

Pooling the updated slots gives the residual representation:

$$h_{\text{res}} = \text{Pool}(R_t), \quad u_t^{\text{res}} = W_{\text{res}} h_{\text{res}}.$$

The VOI-inspired channel uses an orthogonal residual between the static representation and the evidence-conditioned residual representation. The projection of h_{static} onto h_{res} is

$$\text{Proj}_{h_{\text{res}}}(h_{\text{static}}) = \frac{h_{\text{static}}^\top h_{\text{res}}}{\|h_{\text{res}}\|_2^2 + \epsilon} h_{\text{res}}.$$

The orthogonal residual is

$$d_t = h_{\text{static}} - \text{Proj}_{h_{\text{res}}}(h_{\text{static}}).$$

The VOI-inspired channel is then computed as

$$u_t^{\text{voi}} = \text{LayerNorm} (W_{\text{voi}} d_t + e_{\ell(t)}),$$

where $e_{\ell(t)}$ is a learned prefix-length embedding.

The action-card schema and its split-independent construction are described in the Dataset Construction Details appendix. For each candidate examination e , the action encoder produces a shared representation $h_{\text{act}}(e)$. Channel-specific projections then map this shared representation into four action vectors:

$$q^{\text{prior}}(e), \quad q^{\text{traj}}(e), \quad q^{\text{res}}(e), \quad q^{\text{voi}}(e).$$

Inference and STOP Details. At inference, the score defined in Section 4 is applied to every legal candidate:

$$\pi_t = \text{argsort}_{e \in \mathcal{C}_t} s(t, e), \quad \hat{e}_t = \pi_t(1).$$

Examination actions and STOP therefore share one ranking space. Future trajectory labels, final diagnoses, and a separate stopping threshold are not used at inference.

STOP terminates evidence acquisition and is represented in the same action space, while also receiving an auxiliary supervision signal for the observed terminal decision. The auxiliary STOP head receives the four state channels and the prefix-length embedding:

$$g_t = [u_t^{\text{prior}}; u_t^{\text{traj}}; u_t^{\text{res}}; u_t^{\text{voi}}; e_{\ell(t)}].$$

Future-Aware Training Details. For each turn t , the exact positive is the immediate next reference action e_{t+1}^* . We define the future-valid set as legal actions that appear later in the same trajectory:

$$\mathcal{F}_t = \{e_j : j > t + 1\} \cap \mathcal{C}_t.$$

Future-valid actions are not treated as additional positives; the model is still trained to rank e_{t+1}^* highest. However, they are protected from strong negative pressure because they remain consistent with the reference trajectory.

For hard-negative learning, ProActIS mines negatives only from invalid legal candidates:

$$\mathcal{I}_t = \mathcal{C}_t \setminus (\{e_{t+1}^*\} \cup \mathcal{F}_t).$$

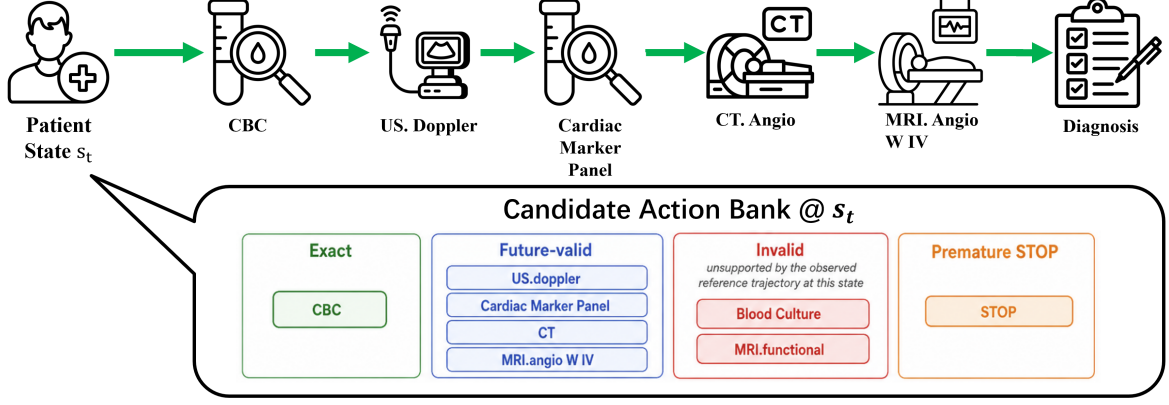


Figure 4: Future-aware candidate partitioning at a partially observed state s_t . The immediate next examination in the reference trajectory is the exact target; later examinations in that trajectory are future-valid, unsupported actions are invalid, and STOP is shown as a separate terminal candidate.

Given the invalid candidate set $\mathcal{I}_t$, the future-filtered hard-negative set is

$$\mathcal{H}_t^K = \text{TopK}_{e \in \mathcal{I}_t} s(t, e).$$

The exact full-bank cross-entropy loss is

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp s(t, e_{t+1}^*)}{\sum_{e \in \mathcal{C}_t} \exp s(t, e)}.$$

Let

$$\mathcal{F}_t^{\text{exam}} = \mathcal{F}_t \setminus \{\text{STOP}\}.$$

Define the non-future-valid legal candidates as

$$\mathcal{U}_t = \mathcal{C}_t \setminus (\{e_{t+1}^*\} \cup \mathcal{F}_t^{\text{exam}}).$$

For discount coefficient $\mu \in [0, 1]$, the future-relaxed denominator is

$$Z_t^{\text{future}} = \exp s(t, e_{t+1}^*) + \sum_{e \in \mathcal{U}_t} \exp s(t, e) + \mu \sum_{e \in \mathcal{F}_t^{\text{exam}}} \exp s(t, e).$$

The future-relaxed cross entropy is

$$\mathcal{L}_{\text{FCE}} = -s(t, e_{t+1}^*) + \log Z_t^{\text{future}}.$$

When $\mu = 1$, this recovers ordinary cross entropy; when $\mu = 0$, non-STOP future-valid examinations are removed from the relaxed denominator. STOP is not discounted, preventing later terminal actions from becoming easier to rank before the reference stopping point.

The ranker objective combines the exact and future-relaxed terms:

$$\mathcal{L}_{\text{ranker}} = (1 - \lambda_F) \mathcal{L}_{\text{CE}} + \lambda_F \mathcal{L}_{\text{FCE}}.$$

Given the invalid hard-negative set $\mathcal{H}_t^K$, ProAc-tIS uses a smooth margin objective:

$$\mathcal{L}_{\text{HN}} = \log \left(1 + \sum_{e^- \in \mathcal{H}_t^K} \exp(m - s(t, e_{t+1}^*) + s(t, e^-)) \right).$$

where m is the ranking margin.

The full objective combines the ranker loss with the future-filtered hard-negative loss and auxiliary STOP sufficiency loss:

$$\mathcal{L} = \mathcal{L}_{\text{ranker}} + \lambda_{\text{HN}} \mathcal{L}_{\text{HN}} + \lambda_{\text{STOP}} \mathcal{L}_{\text{STOP}}.$$

E Subgroup Evaluation Details

The following tables give the complete subgroup breakdown for the untouched test set after excluding terminal STOP decisions. Exact metrics measure the immediate observed action, while E/F metrics measure whether a trajectory-consistent action is recovered; neither is a physician-validated measure of clinical utility. Discharge service is used only as a specialty proxy, and the target-frequency threshold for rare examinations is $\leq 5\%$.

Table 13: Admission-type subgroup performance on non-terminal examination decisions.

Admission group	Cases	Exam turns	Exact@1	E/F@3
Emergency	2,516	25,521	55.69%	92.86%
Non-emergency	1,484	12,575	58.74%	93.42%

Table 14: Discharge-service subgroup performance on non-terminal examination decisions.

Discharge service	Cases	Exam turns	Exact@1	Exact@3	E/F@1	E/F@3
Cardiothoracic	110	1,008	59.82%	83.83%	83.43%	94.84%
Medicine	2,186	23,076	56.42%	79.45%	81.13%	93.02%
Neurology	346	4,040	52.30%	77.77%	76.56%	92.25%
Neurosurgery	140	1,103	55.21%	80.60%	81.60%	93.20%
Obstetrics/gynecology	154	1,028	56.52%	77.92%	77.82%	91.83%
Orthopaedics	236	1,214	64.00%	87.64%	84.27%	95.22%
Psychiatry	76	450	62.89%	82.44%	84.89%	94.44%
Surgery	684	5,713	58.22%	82.25%	80.01%	92.93%
Other/low-count	68	464	61.85%	83.84%	82.97%	93.97%

Table 15: Mapped-diagnosis subgroup performance on non-terminal examination decisions.

Diagnosis category	Cases	Exam turns	Exact@1	Exact@3	E/F@1	E/F@3
Cardiovascular	319	2,786	60.66%	83.81%	83.31%	93.65%
Digestive/hepatobiliary	579	5,269	57.30%	80.66%	79.60%	92.52%
Infectious	401	4,510	55.08%	78.56%	79.71%	92.99%
Injury/toxic	246	1,959	62.38%	87.39%	85.04%	95.61%
Musculoskeletal/spine	110	529	63.33%	84.12%	81.85%	93.95%
Neoplasm	164	1,464	55.40%	75.89%	78.48%	90.64%
Neurologic	273	3,147	54.46%	79.98%	79.00%	93.77%
Obstetric/gynecologic	64	391	58.06%	80.31%	80.05%	92.33%
Psychiatric/substance	64	411	62.04%	82.73%	84.91%	93.43%
Renal/genitourinary	122	1,173	53.62%	78.86%	80.56%	93.09%
Respiratory	266	3,104	56.77%	79.64%	81.51%	92.85%
Symptom-based	125	1,122	58.47%	80.12%	79.32%	91.71%
Other/low-count	87	851	55.58%	79.67%	80.02%	94.01%
Other/unspecified	1,180	11,380	55.55%	79.02%	80.54%	92.90%