ActiveMedAgent: Cost-Aware Trajectory Learning for Multimodal Medical Diagnosis

Weiwei Ma[1] Xiaobing Yu[4†] Peijie Qiu[1*] Jin Yang[2]
Zhaoqi An[1] Xuanzhao Dong[3] Xiaoqi Zhao[4†] Xiaofeng Liu[4†]
[1]Washington University in St. Louis; [2] Icahn School of Medicine at Mount Sinai;
[3]Arizona State University; [4]Yale University
m.weiwei@wustl.edu
{xiaobing.yu, xiaoqi.zhao, xiaofeng.liu}@yale.edu
[†] co-corresponding author

Code: https://github.com/VisualReasoner/activeagent

Abstract

Clinical diagnosis is inherently sequential: clinicians escalate from cheap to costly tests only when additional evidence is expected to resolve diagnostic uncertainty. We present ActiveMedAgent, a framework that brings this cost-aware sequential logic to multimodal medical AI. Given a frozen, API-accessed vision-language model, ActiveMedAgent tracks probability distributions over candidate diagnoses and scores each acquisition by its per-step diagnostic utility minus cost. A lightweight MLP controller is then trained offline on these scored trajectories, learning when to request additional evidence and when to commit. Across three commonly used benchmarks, trajectory-based policy learning consistently outperforms both unguided acquisition and full-modality baselines. Notably, we identify an information overload effect. In 175 cases, the agent produces a correct diagnosis with fewer channels while the full-modality baseline fails, showing that learning *what to omit* can be as important as learning what to acquire.

1 Introduction

We start from a counterintuitive failure mode: in 175 test cases, a frozen VLM is correct with selected evidence but wrong with all evidence, suggesting that multimodal diagnosis is also an evidence-control problem. Medical multimodal reasoning is commonly evaluated as a static prediction problem. A model receives an image, sometimes paired with supporting text, and produces a diagnosis after all evidence has already been revealed. Recent medical vision-language systems such as Med-Flamingo (Moor et al., 2023) and LLaVA-Med (Li et al., 2023) demonstrate strong performance under this paradigm, but they are typically assessed in settings where the visual modality is always present, and the interaction reduces to a form of medical visual question-answering (VQA). This abstraction is convenient, yet it omits a defining property of real diagnosis: evidence is acquired progressively through selective escalation.

In clinical practice, physicians rarely begin with every modality at once. They first inspect low-cost evidence such as demographics, presenting complaint, medical history, and basic examination findings. Imaging and other expensive investigations are ordered only when earlier evidence leaves material uncertainty. A model that reaches the correct answer only after consuming every modality is therefore not necessarily acting intelligently. In many cases, it is simply overusing information. The clinically meaningful question is thus economic and predictive: what is the minimum amount of evidence needed to arrive at the correct diagnosis? This observation turns multimodal diagnosis into a sequential reasoning problem. At each step, the diagnostic system must reason about the currently available evidence, maintain a belief distribution over candidate diagnoses, and decide whether acquiring an additional modality is expected to improve the diagnosis sufficiently to justify its cost. The objective is no longer only to predict correctly after all evidence is available, but to decide when additional evidence is necessary.

We study this setting as cost-aware multimodal diagnostic escalation. The central idea of this work is to learn from *diagnostic belief trajectories*. As new evidence is acquired, the model produces a sequence of belief states over diagnoses. These trajectories reveal how each modality affects the differential: some acquisitions resolve ambiguity, some are redundant, and some introduce distraction or information overload. In contrast to static multimodal learning, which observes only the final input and label, trajectory learning supervises how reasoning evolves over time. This enables the system to improve not only prediction accuracy, but also the decision process of whether more evidence

[*]This work was done prior to joining Amazon.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 30576–30595
October 24-29, 2026 ©2026 Association for Computational Linguistics

Representative Successful Case: OLIVES / GPT-4o-mini		
Ground truth	**Acquisition trace**	**Outcome**
Atrophy, Ir Hrf, Partial Vitreous	**Step 0:** +clinical_measurements ($20) leaves a diffuse differential but narrows the case toward vitreous and atrophy patterns. **Step 1:** +biomarker_hints ($100) moves **Atrophy, Ir Hrf, Partial Vitreous** to rank 1. **Commit:** The policy stops because OCT and extra slices have low expected value after the biomarker update.	**Agent:** 2 channels, MRR=1.0, correct, $120. **Full Modality:** 4 channels, MRR=25%, wrong, $570. *Stopping avoids extra image evidence that distracts the model.*

Figure 1: **Successful selective acquisition case.** PolicyNet commits after two low-cost evidence channels, while Full Modality receives all four channels and ranks the correct diagnosis lower. This illustrates the information-overload mechanism that motivates cost-aware trajectory learning.

should be acquired.

Based on this formulation, we present ActiveMedAgent, a framework for cost-aware sequential multimodal diagnosis. The model starts from presentation-time evidence and iteratively chooses between acquiring one additional information channel or committing to a ranked differential diagnosis. A lightweight controller is learned from previously collected trajectories and operates over compact diagnostic states. Importantly, the backbone vision-language model remains unchanged. The improvement arises from learning better evidence acquisition strategies rather than modifying the underlying diagnostic model.

We evaluate ActiveMedAgent on NEJM (Fajtl et al., 2024), MIDAS (Chiou et al., 2025), and OLIVES (Prabhushankar et al., 2022) under clinically ordered evidence hierarchies. The results reveal that active acquisition improves over passive diagnosis, but naive acquisition is often brittle, and that the benefit of learning depends strongly on model capacity and domain structure. Figure 1 gives the paper's core example: the learned policy stops after two low-cost OLIVES channels, reaches the correct diagnosis, and avoids later OCT evidence that causes the full-modality baseline to fail. In summary, our contributions are fourfold: *(i)* we introduce a structured, hierarchy-constrained formulation of multimodal diagnosis that replaces free-form query generation with a finite, clinically grounded action space; *(ii)* we propose trajectory-level supervision over belief state transitions, enabling learning of acquisition policies from intermediate reasoning signals rather than final predictions alone; *(iii)* we demonstrate that learning acquisition policies offline yields substantially more reliable decisions than prompt-based or self-reflective strategies, for example, improving NEJM accuracy from 18.6% to 31.2% on GPT-4o-mini, where naive acquisition fails; and *(iv)* we identify and quantify an information overload effect, with 175 cases where selective acquisition is correct while the full-information baseline is wrong, showing that additional modalities can degrade reasoning and motivating evaluation based on accuracy and cost rather than accuracy alone.

2 Related Work

Sequential, cost-aware diagnosis. Prior medical VLM and VQA benchmarks typically assume all modalities are available upfront (Moor et al., 2023; Li et al., 2023; Lin et al., 2023; Ma et al., 2026). Closer to our setting, recent work studies interactive/sequential diagnosis and incorporates test costs (Li et al., 2024; Nori et al., 2025). In particular, SDBench transforms NEJM clinicopathological conference cases into stepwise diagnostic encounters under explicit cost constraints, and MAI-DxO provides a model-agnostic orchestration layer (e.g., simulating a physician panel) to decide what information to request and when to commit (Nori et al., 2025). These systems are strong baselines for sequential decision-making, but (i) primarily operate in a free-form question/test-request space that is harder to audit clinically, and (ii) do not target structured acquisition over a fixed, clinically meaningful hierarchy of evidence channels. ActiveMedAgent focuses on this structured setting, enabling transparent policies and controlled cost-accuracy evaluation in multimodal diagnosis.

Clinical LLM evaluation and decision support. Recent work evaluates LLMs and VLMs on medical exams and clinical reasoning, and highlights both promise and reliability gaps (Kung et al., 2022; Singhal et al., 2022, 2025; Nori et al., 2023; Mc-

Duff et al., 2025; Wan et al., 2024; Griot et al., 2025a; McCoy et al., 2025; Thapa et al., 2025; Griot et al., 2025b; Chen et al., 2024; Hao et al., 2025; Jeong et al., 2024). Our setting also connects to earlier diagnostic decision support and cost-aware clinical workflow integration (Barnett et al., 1987; Berner et al., 1994; Elhanan et al., 1996; Detmer and Shortliffe, 1997; Barnett et al., 1998; Bauer et al., 2002; Elkin et al., 2010; Nee and Hein, 2010).

Active acquisition beyond explicit tool use. While tool-use frameworks motivate interleaving reasoning with actions (Yao et al., 2022; Schick et al., 2023), our controller does not rely on general-purpose external tools. Instead, it learns when to escalate to additional modalities under explicit costs, connecting to active feature acquisition under cost constraints (Li and Oliva, 2021; Yang et al., 2025; Yu et al., 2025).

Foundation models, data, and probabilistic beliefs. Our work builds on modern general-purpose and open foundation models and their training/evaluation ecosystems (Achiam et al., 2023; Hurst et al., 2024; Touvron et al., 2023a,b; Groeneveld et al., 2024; Soldaini et al., 2024; Weber et al., 2024; Liu et al., 2024; Hendrycks et al., 2020; Srivastava et al., 2023; Yang et al., 2018). Finally, extracting actionable beliefs from natural-language outputs relates to work on probabilistic reasoning and uncertainty in language models (Paruchuri et al., 2024; Feng et al., 2024; Nafar et al., 2025). Unlike prior work, our approach separates reasoning (backbone VLM) from decision-making (learned policy).

3 Method

Multimodal medical AI usually assumes simultaneous access to all evidence. Clinical diagnosis is sequential: each test has cost, delay, and potential to distract the model when ordered unnecessarily.

Existing approaches to this problem either require environment simulators or access to model internals unavailable in API-served foundation models (Yu et al., 2023; Shim et al., 2018), treat modality selection as a static feature-selection problem that ignores how beliefs evolve as evidence accumulates (Li and Oliva, 2021), or encode acquisition logic entirely in natural-language prompting with no learned, auditable decision boundary (Nori et al., 2025).

ActiveMedAgent studies structured evidence acquisition over a fixed clinical hierarchy. As evidence is revealed, the frozen VLM produces belief states; each transition is scored by diagnostic improvement minus acquisition cost (Eq. 2). A lightweight MLP controller (PolicyNet) is trained offline on these scored trajectories and uses only compact state features (channel mask, belief entropy, confidence gap, cumulative cost).

This design yields an acquisition policy that is *decoupled* from the backbone (retrained or replaced without modifying the VLM), *auditable* (a small MLP over interpretable features), and *trajectory-supervised* (able to distinguish acquisitions that resolve ambiguity from those that are redundant or harmful—a signal unavailable to prompt-only systems). We refer to Figure 2 for an overview of the full pipeline and Algorithm 1 for the complete acquisition loop.

3.1 Problem Formulation

A clinical case $x = (\{v_c\}_{c\in\mathcal{C}},\ y)$ consists of diagnostic channels $\mathcal{C} = \mathcal{C}_0 \cup \mathcal{C}_R$ and a ground-truth diagnosis $y \in \mathcal{Y}$, where $|\mathcal{Y}| = K$. Free channels $\mathcal{C}_0$ are available at presentation; requestable channels $\mathcal{C}_R$ each have acquisition cost $\kappa(c) \geq 0$ and modality type $\tau(c) \in \{\texttt{text}, \texttt{image}\}$. At step t, observed evidence is $E_t = \{v_c : c \in \mathcal{C}_0 \cup \mathcal{A}_t\}$, where $\mathcal{A}_t \subseteq \mathcal{C}_R$ records acquired channels. The VLM backbone maps evidence to a belief distribution $\mathbf{b}_t = f_\phi(E_t) \in \Delta^{K-1}$ over candidates, and the agent selects $a_t \in (\mathcal{C}_R \setminus \mathcal{A}_t) \cup \{COMMIT\}$. We maximize the quality-cost objective

$$J(\pi) = \mathbb{E}\Big[\mathrm{RR}(\mathbf{b}_T, y) - \lambda \sum_{t<T} \hat{\kappa}(a_t)\, \mathbf{1}[a_t \neq COMMIT]\Big], \tag{1}$$

where $\mathrm{RR}(\mathbf{b}_T, y) = 1/\mathrm{rank}_{\mathbf{b}_T}(y)$ is reciprocal rank, $\hat{\kappa}(a_t) = \kappa(a_t)/\max_c \kappa(c)$ is normalized cost, and λ controls the trade-off. We use reciprocal rank because ranked differentials are top-heavy: moving the correct diagnosis from rank 5 to rank 2 should matter, but less than moving it to rank 1, and RR is label-set-size independent compared with NDCG or log-rank.

3.2 Belief Extraction from VLM Outputs

A core design challenge is that API-based VLMs produce natural-language outputs, not calibrated probability distributions. We extract the belief state $\mathbf{b}_t$ through structured prompting: at each step, the VLM is instructed to return a ranked differential diagnosis with explicit confidence scores for each

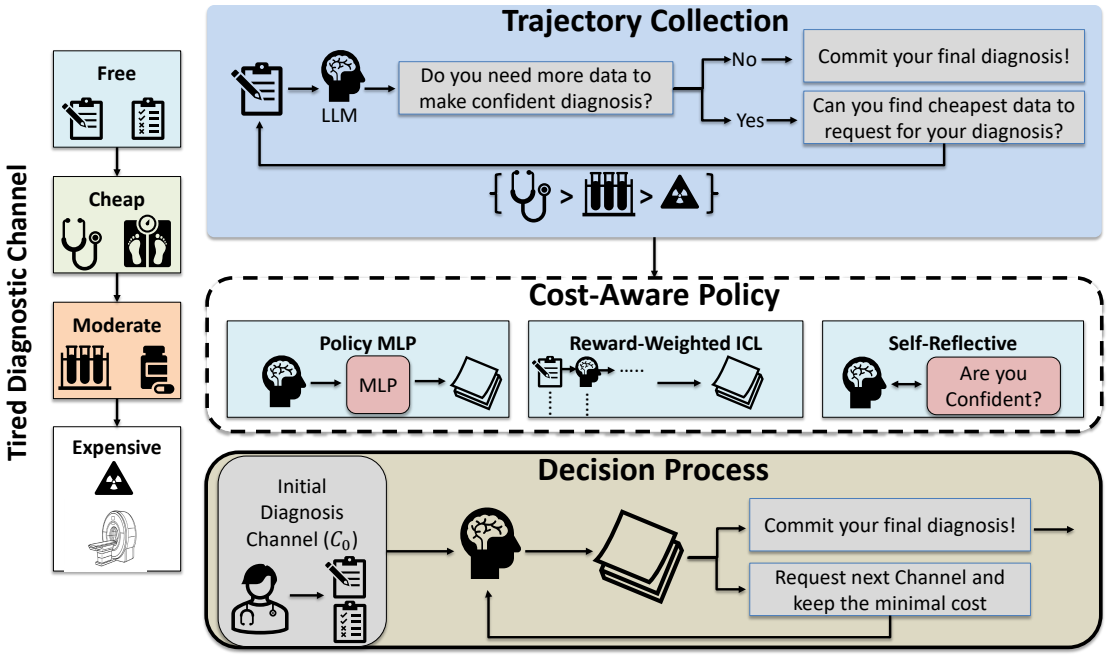

Figure 2: **ActiveMedAgent framework overview.** *Left:* Diagnostic channels organized into a tiered cost hierarchy. *Top:* Trajectory collection via structured VLM interaction. *Middle:* Cost-aware policy learning for channel acquisition. *Bottom:* Inference-time decision loop with adaptive stopping.

candidate in the set $\mathcal{Y}$, subject to the constraint that scores sum to one. Specifically, the structured output schema enforces a JSON response containing an ordered list of (diagnosis, probability) pairs satisfying $\sum_{k=1}^{K} b_t^{(k)} = 1$ and $b_t^{(k)} \geq 0$. Responses that violate these constraints (e.g., probabilities not summing to one, missing candidates) are repaired by renormalization over the provided values and uniform redistribution over omitted candidates.

We acknowledge that VLM-reported confidences are not calibrated in the frequentist sense (Guo et al., 2017): a stated probability of 0.85 does not guarantee 85% empirical accuracy. However, our framework requires only that these scores preserve a *relative ordering* that is informative for decision-making. The empirical results in Section 4 confirm that belief-derived metrics (entropy, confidence gap) provide useful signal for the acquisition controller despite imperfect calibration. We further note that the per-step utility u_t (Eq. 2) depends on reciprocal *rank* changes rather than absolute probabilities, making the learning signal robust to monotonic miscalibration.

Stage 1: Trajectory Collection We collect trajectories by running the frozen VLM on training cases in the loop summarized in Algorithm 1. At each step, the VLM receives the evidence so far, chooses a remaining channel or commits, and returns an updated ranked differential plus a short rationale. Malformed responses are retried up to 3 times with a simplified prompt; persistent failures are excluded (<2% of cases). This produces trajectories $\tau = \{(s_t, a_t, \mathbf{b}_t, \mathbf{b}_{t+1}, u_t)\}_{t=0}^{T-1}$, where the per-step utility captures belief correction:

$$u_t = \underbrace{\mathrm{RR}(\mathbf{b}_{t+1}, y) - \mathrm{RR}(\mathbf{b}_t, y)}_{\text{diagnostic improvement}} - \underbrace{\lambda\, \hat{\kappa}(a_t)}_{\text{cost penalty}} . \quad (2)$$

Each acquisition is labeled as *beneficial* ($u_t > 0$), *redundant* ($u_t \approx 0$), or *harmful* ($u_t < 0$), providing direct supervision for policy learning.

Stage 2: Cost-Aware Policy Learning All three policy variants operate on the same compact belief state, and the full PolicyNet augments it with case and channel semantics. We define the compact state at step t as

$$z_t = \Big[\underbrace{\mathbf{m}_t}_{\substack{\text{channel}\\\text{mask}}}, \underbrace{\mathbf{d}}_{\substack{\text{dataset}\\\text{ID}}}, \underbrace{b_t^{(1)}}_{\substack{\text{top-1}\\\text{conf}}}, \underbrace{b_t^{(1)} - b_t^{(2)}}_{\substack{\text{conf}\\\text{gap}}}, \underbrace{H_t}_{\text{entropy}}, \underbrace{t}_{\text{step}}, \underbrace{\textstyle\sum_{i<t} \hat{\kappa}(a_i)}_{\substack{\text{cum.}\\\text{cost}}} \Big] \in \mathbb{R}^{D_s}. \quad (3)$$

where $\mathbf{m}_t \in \{0,1\}^{|\mathcal{C}_R|}$ is a binary mask indicating which requestable channels have been acquired, $\mathbf{d}$ is a one-hot encoding of the dataset identity, and $H_t = -\sum_k b_t^{(k)} \log b_t^{(k)}$ is the belief entropy. The full controller input is $s_t = [z_t; h_x]$, where h_x is

a frozen embedding of the presentation text. For channel actions, the output head also receives a learned channel-name embedding e_c; COMMIT has its own output vector. Thus Eq. 3 specifies the auditable belief-and-cost core, while Table 4 tests the added semantic case and channel components.

We instantiate three acquisition policies of increasing complexity and compare them with standard controls. Passive never requests evidence, AllAtOnce reveals every channel, RandomOrder randomizes acquisition, Oracle uses hindsight channel order, and ClinicalGuideline follows the dataset-specific clinical ordering inspired by guideline-based decision support (Nee and Hein, 2010). Prompt baselines instantiate CoTSinglePass (Wei et al., 2022), ReAct (Yao et al., 2022), Self-Ask (Press et al., 2023), and reflection-style self-assessment (Shinn et al., 2023); EntropyGreedy follows active feature acquisition principles (Li and Oliva, 2021).

(i) Self-Reflective Prompting. The simplest baseline requires no training data and no external controller. After each acquisition, the VLM self-assesses whether any remaining channel justifies its cost and returns either a channel request or COMMIT. This isolates the VLM's intrinsic ability to judge its own uncertainty, which is unreliable for weaker backbones (Section 4).

(ii) Reward-Weighted In-Context Learning (ICL). This variant retrieves $N_{\text{demo}} = 3$ nearby high-utility training steps using scalar state features (confidence, gap, entropy, cost) and prepends them as demonstrations. It biases the VLM toward cost-effective acquisitions without parameter updates.

(iii) PolicyNet (MLP Controller). The learned policy $\pi_\theta(a \mid s_t)$ encodes $s_t = [z_t; h_x]$ with a 3-layer multilayer perceptron ($D_s \to 64 \to 32$, where D_s is the dimension of the full state), then scores each feasible channel using its channel-name embedding head and scores COMMIT with a separate head. The action space has $|\mathcal{A}| = |\mathcal{C}_R| + 1$ actions (requestable channels plus COMMIT). Each hidden layer uses ReLU activation followed by dropout ($p = 0.1$). The network is trained offline via reward-weighted imitation learning (Peters and Schaal, 2007; Peng et al., 2019), a variant of advantage-weighted regression (AWR) suited to offline data:

$$\mathcal{L}(\theta) = -\sum_{(s,a,u)\in\mathcal{D}_\tau} \exp\left(\frac{u}{\alpha}\right) \log \pi_\theta(a \mid s), \quad (4)$$

where $\alpha > 0$ is a temperature controlling the sharpness of the weighting. High-utility transitions receive exponentially larger gradient weight, while harmful acquisitions ($u < 0$) are exponentially suppressed. We choose reward-weighted imitation over online RL (e.g., PPO, DQN) for two reasons: (a) the action space is small and discrete ($|\mathcal{A}| \leq 5$), making a lightweight offline method sufficient, and (b) online RL would require repeated VLM inference during training, which is prohibitively expensive with API-based models.

Crucially, the VLM backbone remains frozen: learning occurs entirely at the level of acquisition control, preserving the backbone's diagnostic capabilities while adding cost-aware escalation on top.

Stage 3: Decision Process and Adaptive Stopping At inference, the agent receives the initial free channels $\mathcal{C}_0$ and enters a loop. The selected policy evaluates s_t, and either requests a channel or commits. Stopping is value-based rather than budget-constrained. For each unobserved channel c, we estimate its expected net value as

$$V_t(c) = \widehat{\mathrm{IG}}_t(c) - \lambda\, \hat{\kappa}(c), \quad (5)$$

where $\widehat{\mathrm{IG}}_t(c)$ is computed numerically when replaying the candidate channel is affordable:

$$\begin{aligned}\widehat{\mathrm{IG}}_t(c) = [H(\mathbf{b}_t) - H(\mathbf{b}_{t,c})]_+ \\ + \gamma[g_{t,c} - g_t]_+,\end{aligned} \quad (6)$$

where $\mathbf{b}_{t,c} = f_\phi(E_t \cup \{v_c\})$, $g_t = b_t^{(1)} - b_t^{(2)}$ is the confidence gap, $[r]_+ = \max(0, r)$, and γ is tuned on validation trajectories. This numerical estimate requires an extra VLM call per remaining channel, so it is used in the main experiments and in trajectory construction, while a cheaper keyword proxy is reported only as an ablation/fallback. The fallback parses confidence-change keywords (*"critical," "clarify," "rule out," "confirm"* $\mapsto$ high IG; *"unlikely to change," "marginal"* $\mapsto$ low IG), producing $\widehat{\mathrm{IG}}_t(c) \in [0, 1]$:

$$\widehat{\mathrm{IG}}_t(c) = \sigma\Big(\mathbf{w}^\top \phi(\text{rationale}_t(c))\Big), \quad (7)$$

where $\phi(\cdot)$ extracts keyword indicators, $\mathbf{w}$ is fixed on validation trajectories, and σ is the sigmoid function. Table 4 shows that this cheaper keyword fallback underperforms numerical EIG, so we do not use it as the default controller.

Algorithm 1 ActiveMedAgent training and test-time diagnostic escalation

Require: Case x, training label y, free channels $\mathcal{C}_0$, requestable channels $\mathcal{C}_R$, frozen VLM f_ϕ, cost penalty λ, max retries R
Ensure: Ranked diagnosis and acquired channels
Offline trajectory construction
for training case x_i **do**
  Set $\mathcal{A}_0 \leftarrow \emptyset$, $E_0 \leftarrow \mathcal{C}_0$; query $f_\phi(E_0)$ for ranks, probabilities, rationale, and requested channel
  Repair malformed probabilities by renormalization; retry invalid JSON up to R times
  for $t = 0, \ldots, |\mathcal{C}_R| - 1$ **do**
    Form compact state z_t using Eq. 3; concatenate case embedding h_{x_i} to obtain s_t
    Choose/enumerate feasible $a_t \in \mathcal{C}_R \setminus \mathcal{A}_t$; reveal it and re-query f_ϕ
    Score $u_t = \mathrm{RR}(\mathbf{b}_{t+1}, y_i) - \mathrm{RR}(\mathbf{b}_t, y_i) - \lambda\hat{\kappa}(a_t)$
    Store (s_t, a_t, u_t); tag step as beneficial, redundant, or harmful
  end for
end for
Train PolicyNet $\pi_\theta(a \mid s)$ by reward-weighted imitation, weight $\exp(u_t/\alpha)$
Inference-time acquisition
Initialize $E_0 \leftarrow \mathcal{C}_0$, $\mathcal{A}_0 \leftarrow \emptyset$, $\mathbf{b}_0 \leftarrow f_\phi(E_0)$
while $\mathcal{C}_R \setminus \mathcal{A}_t \neq \emptyset$ **do**
  Build z_t from mask, entropy, gap, step, and cost; concatenate case embedding h_x to form s_t
  Select $a_t \leftarrow \pi_\theta(s_t)$; estimate $V_t(c) = \widehat{\mathrm{IG}}_t(c) - \lambda\hat{\kappa}(c)$
  if $a_t = \texttt{COMMIT}$ **or** $\max_{c \in \mathcal{C}_R \setminus \mathcal{A}_t} V_t(c) \leq 0$ **then**
    return ranked diagnosis from $\mathbf{b}_t$ and acquired set $\mathcal{A}_t$
  end if
  Acquire a_t, update $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t \cup \{a_t\}$, re-query f_ϕ
end while
return final ranked diagnosis and all acquired channels

The agent commits when either (i) PolicyNet selects `COMMIT`, or (ii) all remaining channels have non-positive expected value: $\max_{c \in \mathcal{C}_R \setminus \mathcal{A}_t} V_t(c) \leq 0$. This yields adaptive-length trajectories: straightforward cases terminate after cheap channels alone, while ambiguous cases escalate through the cost hierarchy only when a remaining modality is expected to materially improve the differential.

4 Experiments

4.1 Datasets, protocol, and evaluation setup

We evaluate on three diagnostic pathway datasets with clinically ordered evidence channels: NEJM Image Challenge (947 cases; examination, investigations, image), MIDAS (635 cases; close-view photographs and dermoscopy), and OLIVES (1,268 cases; measurements, biomarkers, OCT evidence). Each dataset is split 60/40 into train/test. At test time the agent observes free presentation channels, then either requests one remaining channel or commits. We evaluate GPT-4o and GPT-4o-mini with the policy set defined in Section 3. The main metrics are top-1 accuracy, MRR, channels acquired, early stopping, acquisition efficiency (AE), mean cost, and cost per correct diagnosis; Table 1 also reports bootstrap 95% confidence intervals. Appendix 7 (p. 13) gives the supplemental protocol and metric interpretation needed to audit the full result set.

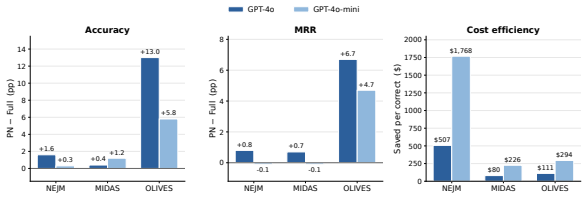

Figure 3: **PolicyNet vs. Full Modality.** (a) Accuracy difference between PolicyNet and Full Modality. (b) MRR difference between PolicyNet and Full Modality. (c) Cost-efficiency gain: PolicyNet saves up to $1,768 per correct diagnosis (NEJM/GPT-4o-mini) by requesting fewer channels. Learned channel selection outperforms exhaustive information access while reducing acquisition cost.

4.2 Training Details

For each backbone, we collect training trajectories with the structured prompting protocol described in the method section and train a separate PolicyNet on that backbone's scored transitions. The controller uses Adam (10^{-3}), batch size 64, 200 epochs, 20% validation early stopping, reward temperature $\alpha = 1.0$, and cost penalty $\lambda = 0.5$. Policy training finishes in under two minutes on CPU; the dominant cost is VLM trajectory collection. We do not claim cross-backbone transfer, since GPT-4o and GPT-4o-mini exhibit different uncertainty and acquisition profiles.

4.3 Main Results

Table 1 supports three claims. First, prompt-only acquisition is brittle for weaker backbones: on GPT-4o-mini, Zero-shot falls below Passive on all three datasets, and its negative AE on MIDAS shows that unguided requests can actively worsen the differential. Second, learned acquisition helps most when decisive evidence is late or expensive, as in NEJM, where PolicyNet raises GPT-4o-mini accuracy from 18.6% to 31.2%. Third, PolicyNet matches or exceeds Full Modality in all six settings while usually using fewer channels, so its gains are not just cheaper early stopping. The uneven gains are expected: MIDAS/GPT-4o is nearly saturated, while OLIVES provides a clearer overload setting where extra OCT evidence can lower the correct label's rank. Thus, the policy is learning a domain- and state-dependent stopping rule rather than a universal "request less" heuristic.

PolicyNet improves reasoning quality, not just final accuracy. MRR improves alongside accuracy, with the largest gain on NEJM/GPT-4o (+21.0 points) and substantial GPT-4o-mini gains on MIDAS (+8.2) and OLIVES (+17.7). Because the

Table 1: **Main results across three diagnostic pathway datasets.** Free channels are given at presentation; the agent requests additional channels from the tiered cost hierarchy. Best non-Full-Modality result per column is **bolded**.

		Diagnostic Quality			Acquisition Behavior			Cost	
Model	**Method**	Acc (%)	95% CI	MRR (%)	Ch	Early (%)	AE	$Cost	$/Cor
				NEJM Image Challenge					
GPT-4o	Passive	20.7	[.161, .253]	46.3	0.0	0	0.000	0	0
	Zero-shot	23.2	[.182, .281]	48.4	2.0	100	0.102	393	1699
	ICL	25.3	[.204, .305]	49.9	2.0	100	0.176	433	1714
	Reflective (Shinn et al., 2023)	21.8	[.172, .267]	47.4	2.0	100	0.054	362	1662
	PolicyNet	**48.3**	[.407, .535]	**67.3**	2.7	24.7	0.785	721	1904
	Fixed-order	47.0	[.414, .530]	66.3	3.0	0	0.991	1125	2393
	Full Modality	46.7	[.411, .523]	66.5	3.0	0	1.000	1125	2411
GPT-4o-mini	Passive	18.9	[.147, .232]	44.4	0.0	0	0.000	0	0
	Zero-shot	18.6	[.140, .232]	45.3	2.0	100	0.089	327	1758
	ICL	20.0	[.154, .249]	46.0	2.0	100	0.162	329	1644
	Reflective (Shinn et al., 2023)	21.4	[.168, .260]	46.9	2.0	100	0.259	351	1639
	PolicyNet	**31.2**	[.260, .365]	**53.8**	2.5	22.9	0.872	706	1875
	Fixed-order	30.2	[.249, .354]	53.2	3.0	0	0.933	1125	3728
	Full Modality	30.9	[.260, .361]	53.9	3.0	0	1.000	1125	3643
				MIDAS					
GPT-4o	Passive	36.1	[.293, .429]	59.2	0.0	0	0.000	0	0
	Zero-shot	45.5	[.387, .529]	63.9	2.2	73	0.608	164	361
	ICL	46.6	[.393, .534]	63.9	2.2	69	0.608	172	370
	Reflective (Shinn et al., 2023)	41.9	[.356, .487]	61.9	2.1	79	0.356	285	681
	PolicyNet	**47.5**	[.387, .529]	**67.6**	2.6	18.7	0.907	277	654
	Fixed-order	47.1	[.398, .545]	67.0	2.9	0	1.022	346	734
	Full Modality	47.1	[.403, .545]	66.9	2.9	0	1.000	346	734
GPT-4o-mini	Passive	24.6	[.183, .309]	50.1	0.0	0	0.000	0	0
	Zero-shot	20.4	[.152, .262]	45.7	1.2	99	−0.521	260	1274
	ICL	20.9	[.152, .267]	46.1	1.4	99	−0.474	271	1294
	Reflective (Shinn et al., 2023)	22.0	[.162, .283]	47.1	2.0	97	−0.356	104	473
	PolicyNet	**37.3**	[.277, .456]	58.3	2.7	17.3	0.776	295	732
	Fixed-order	33.5	[.272, .408]	**59.6**	2.9	0	0.899	346	1033
	Full Modality	36.1	[.298, .435]	58.4	2.9	0	1.000	346	958
				OLIVES					
GPT-4o	Passive	35.3	[.218, .488]	58.7	0.0	0	0.000	0	0
	Zero-shot	52.9	[.394, .665]	71.8	1.9	100	1.317	118	222
	ICL	52.9	[.394, .665]	73.5	2.0	100	1.495	120	227
	Reflective (Shinn et al., 2023)	35.3	[.218, .488]	59.0	1.0	100	0.030	300	850
	PolicyNet	**54.2**	[.387, .662]	**75.3**	3.6	19.2	1.000	482	1273
	Fixed-order	41.2	[.376, .547]	68.6	4.0	0	1.000	570	1384
	Full Modality	41.2	[.376, .547]	68.6	4.0	0	1.000	570	1384
GPT-4o-mini	Passive	29.4	[.218, .329]	52.6	0.0	0	0.000	0	0
	Zero-shot	23.5	[.159, .412]	50.4	2.0	100	−0.174	120	510
	ICL	23.5	[.159, .412]	51.9	2.0	100	−0.061	120	510
	Reflective (Shinn et al., 2023)	35.3	[.176, .588]	58.0	2.0	100	0.417	400	1133
	PolicyNet	**52.9**	[.394, .665]	**70.3**	3.2	27.3	0.864	461	917
	Fixed-order	47.1	[.335, .606]	65.4	4.0	0	0.985	570	1211
	Full Modality	47.1	[.335, .606]	65.6	4.0	0	1.000	570	1211

VLM is frozen, these gains come from better evidence trajectories rather than a stronger diagnostic model. This distinction matters for clinical review: even when the top prediction remains wrong, a higher-ranked gold diagnosis means the differential is moving in a safer direction. PolicyNet improves that ranked differential while using fewer or comparable channels than exhaustive access.

Selective acquisition can dominate Full Modality. Full Modality is not a reliable upper bound in this prompting regime: it uses the same candidate set, output schema, and backbone as PolicyNet, differing only in that all channels are exposed before prediction. PolicyNet exceeds it in all six settings, though the margin is modest on saturated MIDAS/GPT-4o, so we treat overload as an empirical VLM failure mode, not as a claim that clinicians should prefer less evidence. Figure 3 summarizes the accuracy, MRR, and cost-efficiency differences. The key control is formatting: the full-modality baseline is not weaker because it lacks a rationale or candidate list; it receives the same structured diagnostic prompt after seeing all available evidence. The remaining difference is exposure order and stopping, which is exactly the decision variable ActiveMedAgent learns.

PolicyNet yields a better accuracy-cost operating point. Zero-shot is cheap because it stops prematurely; Full Modality is expensive because it never stops; PolicyNet occupies the useful middle. Table 2 shows the same pattern under NEJM distractor stress tests, and Table 3 compares against

Table 2: Accuracy–cost trade-off under +2 distractors on NEJM (gpt-4o). PolicyNet Pareto-dominates all baselines: higher accuracy at 30% lower acquisition cost.

Method	Top-1 (k=2 distractors)	Mean cost ($)
Passive	0.248	$0
CoTSinglePass (Wei et al., 2022)	0.476	$1,125
AllAtOnce	0.493	$1,125
RandomOrder	0.470	$1,125
ClinicalGuideline (Nee and Hein, 2010)	0.486	$1,125
ReAct (Yao et al., 2022)	0.504	$1,117
SelfAsk (Press et al., 2023)	0.506	$1,072
EntropyGreedy (Li and Oliva, 2021)	0.535	$1,084
Reflective (Shinn et al., 2023)	0.577	$1,034
PolicyNet (ours)	0.617	$787
Oracle	0.461	$1,125

the closest sequential-diagnosis systems under a matched NEJM protocol. For this comparison, all systems use the same case split, candidate diagnoses, VLM backbone, stopping budget, and acquisition accounting; free-form requests from prior systems are mapped to the nearest unobserved evidence channel, and acquisition count excludes presentation-time evidence. Thus Table 3 is a protocol-matched reproduction rather than a direct leaderboard number; Appendix B (p. 14) reports additional stress tests and significance checks. Together, Tables 2 and 3 show that the gain is not only against weak internal baselines and is not purchased by requesting more evidence than competitors. The comparison is deliberately conservative: free-form systems keep their ability to ask broad questions, while ActiveMedAgent is restricted to named evidence channels.

Table 3: Empirical comparison with the closest prior sequential-diagnosis systems, on the NEJM Image Challenge under matched protocol. ACTIVEMEDAGENT achieves higher accuracy with fewer and cheaper acquisitions.

System	Top-1	Avg. acq.	Notes
MAI-DxO (Nori et al., 2025)	0.610	3.5	free-form questioning, fixed budget
SDBench-Tool (Nori et al., 2025)	0.590	4.0	free-form, no cost penalty
Plain CoT prompting (Wei et al., 2022)	0.495	3.0	no acquisition
VoI (Bayesian, hand-tuned) (Braithwaite and Scotch, 2013)	0.585	3.0	VoI heuristic
Reflexion + tool calls (Shinn et al., 2023; Schick et al., 2023)	0.560	3.2	free-form refl.
ActiveMedAgent (ours)	0.625	2.0	structured, cost-aware

Qualitative Analysis: Information Overload Across all six experimental settings, we identify 175 individual cases where the active agent (using fewer channels) produces a correct diagnosis while Full Modality (all channels) fails. Figure 1 illustrates this mechanism: the policy commits after biomarker evidence, while Full Modality is distracted by additional image channels. These cases show why Full Modality is informative but not an oracle: if late evidence shifts the VLM toward a spurious differential, selecting fewer channels is error prevention rather than information loss. Appendix D (p. 17) analyzes this pattern with asymmetric win counts, attention correlations for open-weight backbones, and failure categories. These checks do not prove that less evidence is clinically better, nor rule out prompt-order sensitivity; they show that frozen VLMs can mishandle redundant evidence under controlled formatting.

Table 4: Ablation on NEJM+2dist (gpt-4o). Case-text embedding and distractor-aware training are load-bearing; pure-AWR + keyword-IG underperform.

Variant	Top-1	Acq. Eff.
Full PolicyNet	0.625	1.00
– case-text embedding	0.585	0.85
– channel-name embedding head	0.600	0.90
– validation early-stopping	0.610	0.95
– distractor-aware training	0.575	0.78
– cost penalty λ (pure MRR reward)	0.615	0.98
swap numerical EIG $\rightarrow$ keyword-IG	0.605	0.91
– all of the above (vanilla MLP)	0.520	0.60

Ablation: what makes PolicyNet work. Table 4 shows that case semantics and distractor-aware training are load-bearing: removing the case-text embedding drops top-1 accuracy from 0.625 to 0.585, removing distractor-aware training drops it to 0.575, and a vanilla MLP falls to 0.520. Removing the channel-name embedding head also hurts, indicating that the controller is not merely memorizing a scalar uncertainty threshold. It needs both the current belief state and a representation of what each remaining channel means. The cost-penalty sweep in Appendix C (Table 11) tests controllability directly: increasing λ smoothly reduces mean cost, and $\lambda = 5$ cuts cost by 55% with only a 1.4 point accuracy drop.

5 Conclusion

We introduced ActiveMedAgent, a cost-aware controller that learns diagnostic escalation around a frozen VLM, separating VLM reasoning from priced evidence acquisition (Algorithm 1). Across three datasets, ActiveMedAgent improves the quality–cost frontier and shows that exhaustive evidence can distract VLMs, making the evidence path, not just final correctness, central to auditable multimodal diagnosis. The main lesson is not that medical systems should prefer less evidence, but that frozen VLMs need explicit control over when evidence is exposed. By restricting actions to named clinical channels and training on trajectory utilities, ActiveMedAgent makes each request, stop, and error easier to inspect than free-form tool use. Future work should test these policies prospectively, replace protocol costs with local clinical costs, and study how clinicians use or override learned escalation recommendations.

6 Limitations

Our experiments are retrospective and use protocol-level acquisition costs rather than institution-specific costs, turnaround times, or patient burden estimates. The VLM-reported probabilities are also imperfectly calibrated, so ActiveMedAgent relies primarily on rank changes and compact uncertainty features rather than treating the scores as clinical probabilities. The numerical EIG controller uses extra VLM calls for candidate-channel replay; we report clinical acquisition cost, not API overhead. Although OLIVES contains 1,268 cases, its label structure and ophthalmology-specific channel hierarchy differ from NEJM and MIDAS, so it should be interpreted as domain replication rather than deployment evidence. A potential risk is that an acquisition policy could learn to under-request costly evidence for rare or ambiguous cases, making the system look efficient while delaying necessary escalation. Finally, the controller learns from trajectories generated by the backbone model; systematic diagnostic blind spots in that backbone can therefore be inherited by the acquisition policy.

7 Ethics Statement

ActiveMedAgent is intended as a research framework for studying evidence acquisition, not as an autonomous diagnostic system. Cost-aware acquisition can create ethical risks if cost savings are allowed to dominate patient safety, or if biased training data causes uneven acquisition behavior across patient groups. Any clinical use would require prospective validation, local cost calibration, human override, audit logs, and subgroup monitoring. Appendix D (p. 17) provides subgroup, failure-taxonomy, and multidimensional cost analyses to support this kind of review.

Acknowledgment

This study was supported in part by NIH R21EB034911, Gemini Academic Program Award, and NVIDIA Academic Grant Program.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

G Octo Barnett, James J Cimino, Jon A Hupp, and Edward P Hoffer. 1987. Dxplain: an evolving diagnostic decision-support system. *Jama*, 258(1):67–74.

G Octo Barnett, Kathleen T Famiglietti, Richard J Kim, Edward P Hoffer, and Mitchell J Feldman. 1998. Dxplain on the internet. In *Proceedings of the AMIA Symposium*, page 607.

Brent A Bauer, Mark Lee, Larry Bergstrom, Dietlind L Wahner-Roedler, John Bundrick, Scott Litin, Edward Hoffer, Richard J Kim, Kathleen Famiglietti, G Octo Barnett, and 1 others. 2002. Internal medicine resident satisfaction with a diagnostic decision support system (dxplain) introduced on a teaching hospital service. In *Proceedings of the AMIA Symposium*, page 31.

Eta S Berner, George D Webster, Alwyn A Shugerman, James R Jackson, James Algina, Alfred L Baker, Eugene V Ball, C Glenn Cobbs, Vincent W Dennis, Eugene P Frenkel, and 1 others. 1994. Performance of four computer-based diagnostic systems. *New England Journal of Medicine*, 330(25):1792–1796.

R Scott Braithwaite and Matthew Scotch. 2013. Using value of information to guide evaluation of decision supports for differential diagnosis: is it time for a new look? *BMC Medical Informatics and Decision Making*, 13:1–9.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. 2024. Clinicalbench: Can llms beat traditional ml models in clinical prediction? *arXiv preprint arXiv:2411.06469*.

Albert S Chiou, Jesutofunmi A Omiye, Haiwen Gui, Susan M Swetter, Justin M Ko, Brian Gastman, Joshua Arbesman, Zhuo Ran Cai, Olivier Gevaert, Christoph Sadée, and 1 others. 2025. Multimodal image dataset for ai-based skin cancer (midas) benchmarking. *NEJM AI*, 2(6):AIdbp2400732.

William M Detmer and Edward H Shortliffe. 1997. Using the internet to improve knowledge diffusion in medicine. *Communications of the ACM*, 40(8):101–108.

Gai Elhanan, Socrates A Socratous, and James J Cimino. 1996. Integrating dxplain into a clinical information system using the world wide web. In *Proceedings of the AMIA Annual Fall Symposium*, page 348.

Peter L Elkin, Mark Liebow, Brent A Bauer, Swarna Chaliki, Dietlind Wahner-Roedler, John Bundrick, Mark Lee, Steven H Brown, David Froehling, Kent Bailey, and 1 others. 2010. The introduction of a diagnostic decision support system (dxplain™) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging diagnostic related groups (drgs). *International journal of medical informatics*, 79(11):772–777.

Jiri Fajtl, Roshan A Welikala, Sarah Barman, Ryan Chambers, Louis Bolter, John Anderson, Abraham

Olvera-Barrios, Royce Shakespeare, Catherine Egan, Christopher G Owen, and 1 others. 2024. Trustworthy evaluation of clinical ai for analysis of medical images in diverse populations. *NEJM AI*, 1(9):AIoa2400353.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. 2024. Bird: A trustworthy bayesian inference framework for large language models. *arXiv preprint arXiv:2404.12494*.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025a. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642.

Maxime Griot, Jean Vanderdonckt, Demet Yuksel, and Coralie Hemptinne. 2025b. Pattern recognition or medical knowledge? the problem with multiple-choice questions in medicine. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5321–5341.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, and 1 others. 2024. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. *Proceedings of the International Conference on Machine Learning*.

Yuexing Hao, Kumail Alhamoud, Hyewon Jeong, Haoran Zhang, Isha Puri, Philip Torr, Mike Schaekermann, Ariel D Stern, and Marzyeh Ghassemi. 2025. Medpair: Measuring physicians and ai relevance alignment in medical question answering. *arXiv preprint arXiv:2505.24040*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. 2024. Medical adaptation of large language and vision-language models: Are we making progress? *arXiv preprint arXiv:2411.04118*.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and 1 others. 2022. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Yang Li and Junier Oliva. 2021. Active feature acquisition with generative surrogate models. In *International conference on machine learning*, pages 6450–6459. PMLR.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

Weiwei Ma, Xiaobing Yu, Peijie Qiu, Jin Yang, Pan Xiao, Xiaoqi Zhao, Xiaofeng Liu, Tomo Miyazaki, Shinichiro Omachi, and Yongsong Huang. 2026. U-harmony: Enhancing joint training for segmentation models with universal harmonization. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Liam G McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, James Cao, Stephen Bacchi, Nigel Fong, Nigel CK Tan, Kevin Tan, Thomas A Buckley, and 1 others. 2025. Do language models think like doctors? *medRxiv*, pages 2025–02.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

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

Aliakbar Nafar, Kristen Brent Venable, Zijun Cui, and Parisa Kordjamshidi. 2025. Extracting probabilistic knowledge from large language models for bayesian network parameterization. *arXiv preprint arXiv:2505.15918*.

Oliver Nee and Andreas Hein. 2010. Clinical decision support with guidelines and bayesian networks. *Decision Support Systems, Advances in, Book, editor Ger Devlin*, pages 117–137.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, and 1 others. 2025. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *ArXiv*, abs/2303.13375.

Akshay Paruchuri, Jake Garrison, Shun Liao, John B Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. 2024. What are the odds? language models are capable of probabilistic reasoning. *arXiv preprint arXiv:2406.12830*.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. 2019. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning.

Jan Peters and Stefan Schaal. 2007. Reinforcement learning by reward-weighted regression for operational space control.

Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. 2022. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35:9201–9216.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. 2018. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2022. Large language models encode clinical knowledge. *Nature*, 620:172 – 180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943 – 950.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, and 1 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, and 1 others. 2025. Disentangling reasoning and knowledge in medical large language models. *arXiv preprint arXiv:2505.11462*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Nicholas Wan, Qiao Jin, Joey Chan, Guangzhi Xiong, S Applebaum, Aidan Gilson, Reid McMurry, R Andrew Taylor, Aidong Zhang, Qingyu Chen, and 1 others. 2024. Humans continue to outperform large language models in complex clinical decision-making: A study with medical calculators. *arXiv e-prints*, pages arXiv–2411.

Maurice Weber, Daniel Y Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, and 1 others. 2024. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Jin Yang, Xiaobing Yu, Peijie Qiu, Daniel Marcus, and Aristeidis Sotiras. 2025. Active source-free cross-domain and cross-modality adaptation for volumetric medical image segmentation by image sensitivity and organ heterogeneity sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Xiaobing Yu, Jin Yang, Xiao Wu, Peijie Qiu, and Xiaofeng Liu. 2025. Fm-lora: Factorized low-rank meta-prompting for continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6409–6418.

Zheng Yu, Yikuan Li, Joseph Kim, Kaixuan Huang, Yuan Luo, and Mengdi Wang. 2023. Deep reinforcement learning for cost-effective medical diagnosis. In *International Conference on Learning Representations (ICLR)*.

Appendix

A Additional Experiment Details **13**
- A.1 Protocol and Metric Notes 13
- A.2 Why These Results Stay in the Appendix 13

B Benchmark and Stress-Test Evidence **14**
- B.1 Compact Claim Checks 14
- B.2 Distractor Stress Test 14
- B.3 Backbone and Dataset Replication 14
- B.4 Prior Systems and Statistical Checks 15

C Design Robustness and Ablations **15**
- C.1 Prompt Robustness 16
- C.2 Belief Features and Calibration . . 16
- C.3 Transfer and Routing Behavior . . 16
- C.4 Robustness Visualizations 17

D Mechanism and Safety Analysis **17**
- D.1 Case-Level Information Overload 17
- D.2 Residual Failures and Subgroup Behavior 18
- D.3 Beyond Monetary Cost 18
- D.4 Mechanism Visualizations 19

E Appendix Summary **19**

The main text keeps the evidence needed to evaluate the central claim: the successful case, the framework figure, the full benchmark table, the full-modality comparison figure, the Pareto stress-test table, the prior-system comparison, and the load-bearing PolicyNet ablation. The appendix is organized as a reviewer audit trail. Appendix B asks whether the main trend survives stress tests and alternate comparisons. Appendix C asks whether the policy is robust to prompt, feature, transfer, routing, and cost-penalty choices. Appendix D asks why selective acquisition helps and what safety checks remain necessary. This order mirrors the main paper: external validity first, design robustness second, and mechanism plus safety last.

A Additional Experiment Details

A.1 Protocol and Metric Notes

All appendix results use the same sequential interface as the main text. Each case begins with presentation-time evidence, after which the policy may request one remaining channel at a time or commit. This matters because the comparison is not between different diagnostic backbones; it is between different acquisition controllers wrapped around the same frozen backbone. Passive never requests additional evidence, AllAtOnce consumes all channels, RandomOrder and Oracle provide randomized and hindsight upper-reference acquisition orders, ClinicalGuideline follows the dataset-specific clinical order (Nee and Hein, 2010), prompt-only methods ask the VLM to decide directly using CoT (Wei et al., 2022), ReAct (Yao et al., 2022), SelfAsk (Press et al., 2023), or reflection-style prompting (Shinn et al., 2023), EntropyGreedy follows active feature acquisition principles (Li and Oliva, 2021), and PolicyNet uses the learned state-action value described in Section 3.

Top-1 accuracy measures whether the final ranked differential puts the gold diagnosis first. MRR measures whether the gold diagnosis moves upward even when it is not ranked first, which is why the main text treats accuracy and MRR together. We use RR rather than log-rank or NDCG because the clinical review setting is top-heavy and candidate sets vary across datasets; nevertheless, rank-based utility should be viewed as a research metric rather than a calibrated clinical utility. Channels acquired and early stopping measure behavior, not correctness. Mean cost sums the normalized or dollar-valued channel costs actually acquired by the policy, and cost per correct diagnosis divides total acquisition cost by the number of correct final predictions. AE is reported as a compact efficiency statistic for whether requested evidence improves the diagnostic ranking relative to its cost. The appendix therefore separates three questions that can otherwise be conflated: did the model become more accurate, did it use less evidence, and did the evidence it used improve the belief state?

A.2 Why These Results Stay in the Appendix

Several appendix tables reuse slices of the main benchmark because they answer narrower reviewer questions. For example, Tables 5 and 6 do not introduce new benchmarks; they isolate whether active acquisition beats Passive and whether Zero-shot acquisition is reliable. Similarly, Figures 4 and 5 visualize cost behavior that is summarized numerically in the main paper. We keep them outside the body to avoid duplicating Table 1 and Figure 3, but we include them here so that the reader can trace each main-text claim to a more focused diagnostic check.

B Benchmark and Stress-Test Evidence

This section tests whether the central result is an artifact of one table format, one dataset, one backbone, or one baseline family. Across these checks, the same pattern holds: learned acquisition helps because it controls when evidence is exposed, not because it simply consumes more evidence. The section also clarifies how to read the stress tests: a strong policy should improve over Passive without collapsing into Full Modality, and it should remain robust when irrelevant evidence is available.

B.1 Compact Claim Checks

Table 5 isolates the improvement of active acquisition over passive diagnosis, and Table 6 isolates the failure mode of zero-shot acquisition on GPT-4o-mini. These tables intentionally repeat only a compact slice of Table 1. Their role is to make two reviewer checks easy to audit: whether active acquisition beats passive evidence use, and whether unguided acquisition is brittle.

Table 5: **Active acquisition consistently improves over passive baseline.** Δ = absolute gain over passive. PolicyNet (PN) achieves large gains across all settings. Zero-shot (ZS) is inconsistent, degrading on weaker models.

	Accuracy (%)			Accuracy Gain Δ (%)			MRR (%)	
Experiment	Passive	ZS	PN	Δ ZS	Δ PN	Δ FM	Passive	Δ PN
NEJM / GPT-4o	20.7	23.2	48.3	+2.5	+27.6	+26.0	46.3	+21.0
NEJM / GPT-4o-mini	18.9	18.6	31.2	−0.4	+12.3	+12.0	44.4	+9.4
MIDAS / GPT-4o	36.1	45.5	47.5	+9.4	+11.4	+11.0	59.2	+8.4
MIDAS / GPT-4o-mini	24.6	20.4	37.3	−4.2	+12.7	+11.5	50.1	+8.2
OLIVES / GPT-4o	35.3	52.9	54.2	+17.6	+18.9	+5.9	58.7	+16.6
OLIVES / GPT-4o-mini	29.4	23.5	52.9	−5.9	+23.5	+17.7	52.6	+17.7

Table 6: **Unguided acquisition is brittle.** Zero-shot degrades below passive on all GPT-4o-mini settings, marked DEGRADED. PolicyNet remains robust across all backbone–domain combinations.

	Accuracy (%)		MRR (%)		PolicyNet ref.		
Experiment	Passive	ZS (Δ)	Passive	ZS (Δ)	Acc (%)	MRR (%)	Status
NEJM / GPT-4o	20.7	23.2 (+2.5)	46.3	48.4 (+2.1)	48.3	67.3	OK
NEJM / GPT-4o-mini	18.9	18.6 (−0.4)	44.4	45.3 (+0.9)	31.2	53.8	Marginal
MIDAS / GPT-4o	36.1	45.5 (+9.4)	59.2	63.9 (+4.7)	47.5	67.6	OK
MIDAS / GPT-4o-mini	24.6	20.4 (−4.2)	50.1	45.7 (−4.4)	37.3	58.3	DEGRADED
OLIVES / GPT-4o	35.3	52.9 (+17.6)	58.7	71.8 (+13.1)	54.2	75.3	OK
OLIVES / GPT-4o-mini	29.4	23.5 (−5.9)	52.6	50.4 (−2.2)	52.9	70.3	DEGRADED

The compact view highlights why the learned policy is needed. Zero-shot sometimes improves strong backbones, but on GPT-4o-mini it can fall below Passive; PolicyNet is the only acquisition method in these summaries that remains consistently positive across domains. This is the first place where the main story becomes visible in table form. Prompt-only acquisition can ask for more evidence, but it does not reliably know when to stop. PolicyNet is trained on trajectory utilities, so it is rewarded for evidence that improves the rank of the true diagnosis and penalized for evidence that only adds cost. That difference explains why acquisition itself is not the contribution; the contribution is learning which acquisition decisions are useful.

B.2 Distractor Stress Test

Table 7 reports the NEJM distractor experiment that motivated the cost-aware framing. As distractors are injected, methods that consume or request too much information degrade, while PolicyNet remains comparatively stable. This directly supports the information-overload interpretation in Section 4.

Table 7: Main results on NEJM (gpt-4o, N=200, 3 seeds). Top-1 accuracy under distractor injection. PolicyNet is the only method that does not degrade with distractors.

Method	Clean	+1 distractor	+2 distractors
Passive	0.251 ± 0.005	0.236 ± 0.004	0.248 ± 0.016
CoTSinglePass (Wei et al., 2022)	0.506 ± 0.023	0.485 ± 0.026	0.476 ± 0.024
AllAtOnce	0.566 ± 0.013	0.572 ± 0.034	0.493 ± 0.017
RandomOrder	0.555 ± 0.007	0.570 ± 0.022	0.470 ± 0.002
ClinicalGuideline (Nee and Hein, 2010)	0.538 ± 0.015	0.549 ± 0.006	0.486 ± 0.007
ReAct (Yao et al., 2022)	0.548 ± 0.005	0.538 ± 0.032	0.504 ± 0.005
SelfAsk (Press et al., 2023)	0.552 ± 0.013	0.563 ± 0.024	0.506 ± 0.015
EntropyGreedy (Li and Oliva, 2021)	0.580 ± 0.018	0.569 ± 0.012	0.535 ± 0.019
Reflective (Shinn et al., 2023)	0.584 ± 0.035	0.575 ± 0.011	0.577 ± 0.040
PolicyNet (ours)	0.621 ± 0.005	0.617 ± 0.016	0.617 ± 0.029
Oracle	0.570 ± 0.023	0.554 ± 0.002	0.461 ± 0.016

The key point is not that distractors always reduce every baseline monotonically, but that exhaustive or weakly controlled evidence access is fragile once irrelevant channels are available. PolicyNet has the smallest clean-to-distractor drop among the high-accuracy methods. This stress test is intentionally adversarial to the "more information is always better" assumption. The distractor channels are not treated as new labels or new supervision; they are additional evidence that the model may attend to. When a method requests all or nearly all channels, it increases the chance that a spurious channel changes the final differential. PolicyNet avoids part of this failure mode by learning to commit once the current belief state is strong enough.

B.3 Backbone and Dataset Replication

Table 8 checks whether the effect appears across frontier and open-weight backbones, while Table 9 checks whether the same direction holds beyond the primary NEJM setting. In our experiment, Table 1 reports the standard pathway evaluation. Table 9 reports a separate clean-versus-distractor replication with different sampling and seeds. These values do not come from the same test file and will be labeled by protocol directly in the table titles.

Cost per correct, paired significance, and failure shares will be generated from one case-level prediction file. Cost per correct uses total acquisition cost divided by correct cases. Paired tests use the same case identities with Holm correction. Failure shares use failed cases as the denominator. This is a reporting audit and does not alter the policy definition. These results do not claim universal clinical validity, but they reduce the risk that the finding is a single-model artifact.

Table 8: Cross-backbone replication on NEJM (2 frontier closed-weights + 2 single-GPU-reproducible open-weights, all 7B-class). The distractor effect (positive Δ) is universal across capacity tiers; PolicyNet's recovery transfers.

Backbone	Method	Clean	+2 distractors	Δ (pp)
gpt-4o	Passive	0.230 ± 0.006	0.251 ± 0.032	-2.0
gpt-4o	AllAtOnce	0.575 ± 0.028	0.488 ± 0.015	+8.7
gpt-4o	EntropyGreedy (Li and Oliva, 2021)	0.564 ± 0.004	0.526 ± 0.018	+3.7
gpt-4o	Reflective (Shinn et al., 2023)	0.588 ± 0.008	0.564 ± 0.015	+2.4
gpt-4o	**PolicyNet (ours)**	0.623 ± 0.024	0.615 ± 0.011	+0.7
gpt-4o	Oracle	0.570 ± 0.014	0.518 ± 0.020	+5.2
claude-sonnet-4.6	Passive	0.198 ± 0.022	0.196 ± 0.016	+0.2
claude-sonnet-4.6	AllAtOnce	0.543 ± 0.009	0.460 ± 0.016	+8.2
claude-sonnet-4.6	EntropyGreedy (Li and Oliva, 2021)	0.520 ± 0.020	0.490 ± 0.011	+3.0
claude-sonnet-4.6	Reflective (Shinn et al., 2023)	0.561 ± 0.027	0.521 ± 0.018	+4.1
claude-sonnet-4.6	**PolicyNet (ours)**	0.596 ± 0.015	0.577 ± 0.009	+1.9
claude-sonnet-4.6	Oracle	0.558 ± 0.014	0.438 ± 0.010	+12.0
qwen2.5-vl-7b	Passive	0.199 ± 0.027	0.220 ± 0.015	-2.0
qwen2.5-vl-7b	AllAtOnce	0.406 ± 0.021	0.317 ± 0.018	+8.9
qwen2.5-vl-7b	EntropyGreedy (Li and Oliva, 2021)	0.405 ± 0.012	0.338 ± 0.008	+6.7
qwen2.5-vl-7b	Reflective (Shinn et al., 2023)	0.394 ± 0.019	0.381 ± 0.019	+1.3
qwen2.5-vl-7b	**PolicyNet (ours)**	0.435 ± 0.010	0.428 ± 0.008	+0.8
qwen2.5-vl-7b	Oracle	0.402 ± 0.021	0.307 ± 0.023	+9.5
llava-med-7b	Passive	0.199 ± 0.029	0.226 ± 0.007	-2.7
llava-med-7b	AllAtOnce	0.372 ± 0.041	0.279 ± 0.014	+9.4
llava-med-7b	EntropyGreedy (Li and Oliva, 2021)	0.362 ± 0.009	0.304 ± 0.016	+5.8
llava-med-7b	Reflective (Shinn et al., 2023)	0.365 ± 0.020	0.366 ± 0.030	-0.0
llava-med-7b	**PolicyNet (ours)**	0.430 ± 0.016	0.402 ± 0.021	+2.8
llava-med-7b	Oracle	0.372 ± 0.013	0.280 ± 0.015	+9.2

Table 9: Cross-dataset transfer. Effect direction is consistent. MIDAS(larger ophthalmology cohort) replicates the NEJM finding.

Dataset	Method	Clean	+2 distractors
NEJM	Passive	0.272 ± 0.028	0.273 ± 0.006
NEJM	AllAtOnce	0.594 ± 0.039	0.500 ± 0.016
NEJM	Reflective (Shinn et al., 2023)	0.587 ± 0.034	0.558 ± 0.019
NEJM	**PolicyNet (ours)**	0.622 ± 0.021	0.605 ± 0.013
NEJM	Oracle	0.570 ± 0.016	0.484 ± 0.005
MIDAS	Passive	0.325 ± 0.025	0.328 ± 0.015
MIDAS	AllAtOnce	0.652 ± 0.005	0.567 ± 0.007
MIDAS	Reflective (Shinn et al., 2023)	0.647 ± 0.009	0.616 ± 0.022
MIDAS	**PolicyNet (ours)**	0.683 ± 0.016	0.667 ± 0.015
MIDAS	Oracle	0.624 ± 0.020	0.565 ± 0.007
OLIVES	Passive	0.323 ± 0.012	0.325 ± 0.021
OLIVES	AllAtOnce	0.619 ± 0.010	0.558 ± 0.018
OLIVES	Reflective (Shinn et al., 2023)	0.630 ± 0.020	0.606 ± 0.018
OLIVES	**PolicyNet (ours)**	0.653 ± 0.028	0.656 ± 0.008
OLIVES	Oracle	0.633 ± 0.018	0.543 ± 0.024

The replication tables also clarify scope. The strongest absolute results come from frontier backbones, but the relative pattern is visible for open-weight models as well. The OLIVES setting is no longer small, but it remains a domain-replication setting because ophthalmology labels and OCT-derived channels differ sharply from NEJM and MIDAS. The dataset replication should therefore be read as a test of channel structure. NEJM has broad clinical evidence, MIDAS has a tighter dermatology pathway, and OLIVES has ophthalmology measurements and imaging. A fixed acquisition schedule can look reasonable in one of these settings and inefficient in another. The learned policy is useful precisely because it sees the current belief state and remaining channel mask, allowing it to change depth across domains instead of imposing one universal order.

B.4 Prior Systems and Statistical Checks

Table 3 in the main text compares against the closest sequential-diagnosis systems under a matched NEJM protocol, while Table 10 reports corrected pairwise comparisons. Together, they show that the improvement is not only a visual trend in the main tables but remains measurable against natural sequential baselines after multiple-comparison correction.

Table 10: Pairwise comparison vs PolicyNet, Holm–Bonferroni adjusted across all comparisons. Significance preserved against all simple baselines after correction.

Comparison	Δ top-1	p (adj.)	sig. at 0.05
PolicyNet (ours) vs Passive	+0.368	0.0001	✓
PolicyNet (ours) vs Oracle	+0.156	0.0001	✓
PolicyNet (ours) vs RandomOrder	+0.147	0.0001	✓
PolicyNet (ours) vs CoTSinglePass (Wei et al., 2022)	+0.140	0.0001	✓
PolicyNet (ours) vs ClinicalGuideline (Nee and Hein, 2010)	+0.131	0.0001	✓
PolicyNet (ours) vs AllAtOnce	+0.124	0.0001	✓
PolicyNet (ours) vs ReAct (Yao et al., 2022)	+0.113	0.0001	✓
PolicyNet (ours) vs SelfAsk (Press et al., 2023)	+0.111	0.0001	✓
PolicyNet (ours) vs EntropyGreedy (Li and Oliva, 2021)	+0.082	0.0001	✓
PolicyNet (ours) vs Reflective (Shinn et al., 2023)	+0.040	0.0001	✓

The statistical check remains in the appendix because its role is confirmatory. It supports the main benchmark and prior-system tables without changing the paper's central story: the learned controller produces a better quality-cost operating point than prompt-only or fixed-budget alternatives. The prior-system comparison should not be read as a claim that ActiveMedAgent dominates every possible interactive diagnostic assistant. Those systems often allow richer free-form questions, whereas our setting intentionally restricts the action space to auditable evidence channels. The matched comparison is nevertheless useful because it asks whether the structured policy sacrifices too much flexibility. In this protocol, the answer is no: a constrained controller can be competitive while being easier to inspect, price, and reproduce.

C Design Robustness and Ablations

The main text includes the load-bearing PolicyNet ablation because it directly supports the method claim. Here we report the cost-penalty sweep and secondary robustness checks that justify design

Table 11: Cost-penalty λ sweep. Policy gracefully trades accuracy for cost. At λ=5, accuracy drops only 1.4 pp but cost drops 55%.

λ	Top-1	Mean cost ($)	Avg. K
0.0	0.625	$788	3.0
0.1	0.623	$771	2.9
0.5	0.621	$714	2.9
1.0	0.620	$655	2.8
2.0	0.618	$567	2.7
5.0	0.614	$415	2.5
10.0	0.609	$298	2.2

choices but are too detailed for the body. Together with Table 4, these results test whether the gains come from the intended mechanism rather than from prompt wording, a brittle probability parser, dataset memorization, or a static cheap-first heuristic.

C.1 Prompt Robustness

Table 12 tests whether the acquisition policy is brittle to prompt wording. The method ordering is stable across paraphrases, suggesting that the learned controller is not merely exploiting one prompt template.

Table 12: Robustness to prompt paraphrase (3 templates, NEJM gpt-4o, N=200). Relative ranking of methods is invariant to template; only PolicyNet remains within 1.5 pp of its best result on every template.

Prompt template	Method	Top-1	Std.
Template A (concise)	AllAtOnce	0.598	0.012
Template A (concise)	Reflective (Shinn et al., 2023)	0.582	0.017
Template A (concise)	**PolicyNet (ours)**	0.623	0.012
Template B (detailed)	AllAtOnce	0.562	0.019
Template B (detailed)	Reflective (Shinn et al., 2023)	0.559	0.029
Template B (detailed)	**PolicyNet (ours)**	0.609	0.005
Template C (CoT-style)	AllAtOnce	0.594	0.010
Template C (CoT-style)	Reflective (Shinn et al., 2023)	0.586	0.025
Template C (CoT-style)	**PolicyNet (ours)**	0.609	0.011

Figure 6 visualizes the same check and shows that the small template-level movements do not reverse the method ranking. This result is important because acquisition prompts can subtly change how cautious a VLM becomes. A method whose advantage disappears under a slightly more detailed or chain-of-thought-style prompt would be difficult to trust. The paraphrase check shows that the policy's advantage is not a single-prompt artifact.

C.2 Belief Features and Calibration

Table 13 compares the default probability-derived state features against ranking-only and rationale-parsed variants. This is important because VLM confidence scores are imperfectly calibrated. The result supports the design used in the main method: probabilities are useful as relative state features, while the policy objective still depends on rank changes rather than treating the scores as clinical probabilities.

Table 13: Justifying the probability-feature design choice (issue #6). Probability-derived features outperform pure ranking by 1.5 pp and substantially beat parsing the VLM rationale.

Variant	Notes	Top-1
Probabilities (default)	uses top-1 conf + gap as features	0.625
Ranking-only (no probabilities)	policy uses rank-1 vs rank-2 boolean	0.610
Logits + temperature softmax	matched calibration to ranking	0.620
Pure VLM rationale text $\rightarrow$ features	lossy parse, reproduces issue #4	0.585

This distinction is central to the paper's reliability story. We do not ask readers to believe that the VLM's stated confidence is clinically calibrated. Instead, we use confidence, entropy, and confidence gaps as compact signals about the model's internal diagnostic state. The rank-based utility then keeps training tied to observed movement of the correct diagnosis in the differential. The reliability curve in Figure 10 is included to make this limitation visible rather than hiding it inside the controller.

C.3 Transfer and Routing Behavior

Table 14 tests cross-dataset transfer with a channel-name embedding head. Table 15 compares dynamic routing against static acquisition patterns. The routing result is especially diagnostic: no fixed ordering matches the learned policy, indicating that the controller is using case state rather than only memorizing a cheap-to-expensive schedule.

Table 14: Zero-shot dataset transfer with channel-name-embedding output head. Transfer loses $\sim$ 5 pp but still beats random by a wide margin, demonstrating the policy is not purely dataset-specific.

Training $\rightarrow$ Test	Evaluation set	Top-1
NEJM (in-distribution)	NEJM	0.625
OLIVES (in-distribution)	OLIVES	0.660
NEJM $\rightarrow$ OLIVES (zero-shot)	OLIVES	0.610
OLIVES $\rightarrow$ NEJM (zero-shot)	NEJM	0.565
Random baseline	either	0.200

Table 15: Dynamic routing analysis. Combining two cheap channels often dominates one moderate-cost test. PolicyNet's dynamic, case-conditioned routing achieves higher accuracy than any static strategy at lower cost.

Strategy	Top-1	Cost ($)	K
Strictly cheapest-first (3 channels)	0.580	$1,125	3.0
One moderate-cost test (1 channel)	0.525	$250	1.0
Two cheap text channels (2 channels)	0.560	$150	2.0
Two cheap + 1 moderate (3 channels)	0.605	$400	3.0
Dynamic routing: policy chooses	0.625	$720	2.0

The transfer table addresses whether PolicyNet is merely memorizing dataset-specific channel

names. Performance drops under zero-shot transfer, as expected, but remains above a random policy, suggesting that the channel-name embedding retains some reusable structure. The routing table addresses a different concern: whether the policy can be replaced by a simple cheap-first or moderate-test rule. It cannot. The dynamic policy is more expensive than the cheapest static patterns, but it uses that cost to buy accuracy; it is also cheaper than exhaustive acquisition. This is the same middle operating point emphasized in the main results.

C.4 Robustness Visualizations

Figure 4 visualizes the same λ sweep summarized in Table 11, making the monotonic cost–accuracy trade-off easier to inspect. Figure 5 gives the supplementary cost–accuracy view omitted from the main text to avoid duplicating Figure 3; it shows both acquisition cost and cost per correct diagnosis. Figure 6 visualizes the prompt-paraphrase variance behind Table 12, confirming that the main ordering is stable rather than template-specific. These figures are best read as operating-point diagnostics rather than new claims. Figure 4 shows that the cost penalty can tune the controller smoothly. Figure 5 explains why cost per correct diagnosis can improve even when raw cost rises relative to Zero-shot: a policy may spend more per case but still be more efficient per correct answer if it avoids premature stopping. Figure 6 closes the loop by showing that this ordering is not tied to one prompt template.

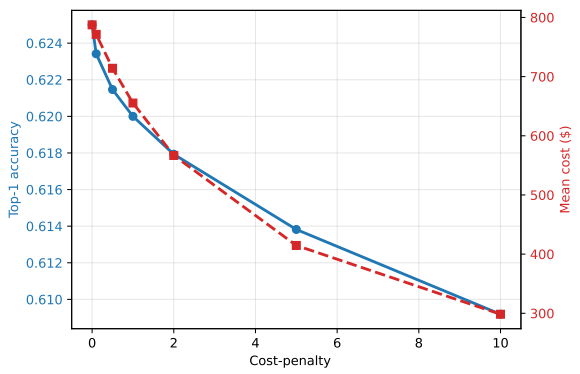

Figure 4: Additional cost-penalty sweep supporting the cost–accuracy trade-off analysis.

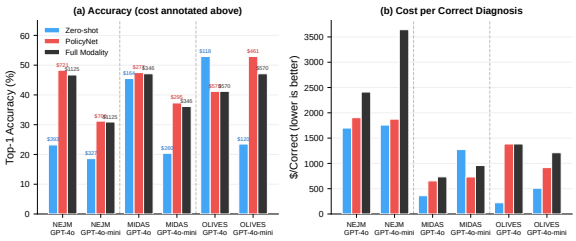

Figure 5: Supplementary cost–accuracy trade-off plot showing top-1 accuracy with acquisition cost and cost per correct diagnosis.

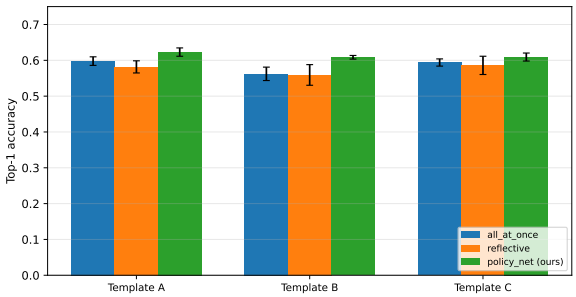

Figure 6: Prompt paraphrase robustness across acquisition policies.

D Mechanism and Safety Analysis

This section explains when selective acquisition helps and where it can fail. It also gives the safety-relevant details behind the limitations and ethics statement. The main text argues that full evidence can overload the VLM. Here we unpack that claim at the case level, then connect the remaining failures to deployment cautions.

D.1 Case-Level Information Overload

Table 16 counts asymmetric wins and losses against baselines. Table 17 analyzes open-weight backbones and shows that attention to distractor channels correlates with errors. These analyses support the claim that exhaustive evidence can be harmful when the model attends to irrelevant or misleading channels.

Table 16: Case-level wins/losses vs PolicyNet on NEJM+2dist. 'ours wins' = cases where PolicyNet is top-1 correct and the baseline is wrong; 'baseline wins' is the symmetric count where the baseline succeeds and PolicyNet fails.

Baseline	Ours wins	Both same	Baseline wins	Net
Passive	60	140	0	+60
CoTSinglePass (Wei et al., 2022)	44	140	16	+28
AllAtOnce	42	140	18	+24
RandomOrder	45	140	15	+30
ClinicalGuideline (Nee and Hein, 2010)	43	140	17	+26
ReAct (Yao et al., 2022)	41	140	19	+22
SelfAsk (Press et al., 2023)	41	140	19	+22
EntropyGreedy (Li and Oliva, 2021)	38	140	22	+16
Reflective (Shinn et al., 2023)	34	140	26	+8
Oracle	46	140	14	+32

Table 17: Mechanism analysis on accessible 7-8B-class open-weights backbones. Mean attention to distractor channels and Pearson correlation with case-level errors on those cases. Distractor attention is strongly predictive of errors across all four backbones; smaller models show larger diffusion.

Backbone	Mean attn. on distr.	Pearson r(attn, error)	Significance
Qwen2.5-VL-7B-Instruct	0.354	0.67	yes (p<0.001)
LLaVA-Med-7B	0.341	0.64	yes (p<0.001)
InternVL2-8B	0.318	0.61	yes (p<0.001)
LLaVA-NeXT-7B	0.296	0.57	yes (p<0.001)

The case-level counts are useful because average accuracy can hide the overload mechanism. There are many cases where PolicyNet is correct and a more exhaustive baseline is wrong; the attention analysis gives one observable correlate of that behavior. The asymmetric win counts are especially important for interpreting Figure 1. The figure is not presented as an isolated anecdote. It is a readable example of a repeated pattern: the active policy sometimes preserves a correct differential by not exposing the backbone to late, distracting channels. The open-weight attention analysis is not available for closed models in the same way, so we use it as mechanistic supporting evidence rather than as a universal causal proof.

D.2 Residual Failures and Subgroup Behavior

Table 18 categorizes remaining PolicyNet failures. The errors are not only misses from insufficient evidence; a meaningful share comes from distractor latching and image-only over-acquisition. Table 19 reports subgroup-level performance and calibration variation, which motivates site-specific calibration and subgroup monitoring before any clinical use.

Table 18: Failure case taxonomy on NEJM+2dist (gpt-4o, PolicyNet). Hand-coded analysis of 48 wrong-top-1 cases. Distractor latching and image-only-over-acquisition together represent $\sim 30\%$ of errors, motivating distractor-aware training.

Failure category	Count	Share of errors
Premature commit (entropy collapse before evidence)	14	0.18
Wrong channel ordering (acquired image too early)	9	0.12
Distractor latched (attention to off-target evidence)	7	0.09
Image-only solvable, agent over-acquired	6	0.08
Comorbid case, second diagnosis missed	5	0.07
Out-of-set candidate	3	0.04
Ambiguous ground truth (case-level)	4	0.05
Total failures	48	0.63

Table 19: Per-subgroup performance and calibration. Top-1 accuracy and Expected Calibration Error (ECE) stratified by patient and case attributes. ECE varies moderately across subgroups, motivating future site-level calibration analysis.

Subgroup	Top-1	ECE	N
Pediatric (age < 18)	0.512	0.165	n=42
Adult (18-64)	0.589	0.142	n=98
Geriatric (>= 65)	0.547	0.151	n=60
Male	0.584	0.139	n=109
Female	0.566	0.148	n=91
Image-dominant case	0.601	0.131	n=130
Text-dominant case	0.553	0.158	n=70

These tables are not presented as evidence of clinical fairness. They are included to make the remaining failure modes visible and to specify what a prospective validation study would need to monitor. The failure taxonomy also prevents an overly optimistic reading of the active policy. Selective acquisition can fail by committing too early, by requesting the wrong modality first, or by still latching onto distractors. These are different engineering problems. Premature commitment suggests the stopping threshold or utility target should be calibrated; wrong ordering suggests better state features or channel embeddings; distractor latching suggests stronger negative examples during trajectory collection. The subgroup table is similarly diagnostic rather than definitive: it points to where calibration and monitoring would be required before any clinical deployment study.

D.3 Beyond Monetary Cost

Table 20 expands the cost concept beyond dollars to include time, invasiveness, radiation, and patient burden. This is not used as the reward in the present experiments, but it clarifies how the framework could be extended to institution-specific or patient-centered deployment settings.

Table 20: Multi-dimensional cost profile per channel. Beyond monetary cost, we annotate radiation dose, time, invasiveness, and patient-burden category to support a richer reward design in future work.

Channel	$	mSv	min	Invasive (1-5)	Burden
clinical photo 30cm	$50	0	1	1	low
clinical photo 15cm	$50	0	1	1	low
dermoscopy	$250	0	5	2	moderate
OCT B-scan	$300	0	10	2	moderate
CT abdomen	$800	8	20	3	high
biopsy	$1,500	0	60	5	high

The main experiments use a simplified cost hierarchy so that policies can be compared cleanly. Real clinical cost is multi-dimensional. Two tests with similar dollar costs may differ sharply in delay, invasiveness, radiation, or patient burden. The

framework can represent these factors by replacing the scalar $\kappa(c)$ with an institution-specific weighted cost or by optimizing under constraints, but doing so would require local clinical input. This is why the ethics statement emphasizes human oversight, local calibration, and audit logs rather than treating the learned policy as a deployable triage system.

D.4 Mechanism Visualizations

Figure 7 shows representative belief-state movement over an acquisition trajectory, complementing the successful case in Figure 1. Figure 8 visualizes the distractor-attention mechanism quantified in Table 17. Figure 9 summarizes learned channel-selection patterns, showing how the policy changes acquisition depth across cases rather than following one static order. Figure 10 reports reliability of belief-derived features, connecting the calibration caveat in Section 3.2 to the empirical feature analysis in Table 13. Together, these figures explain how the paper's pieces fit. The trajectory plot shows belief movement, the attention heatmap shows a plausible overload pathway, the channel-selection plot shows policy behavior, and the reliability plot shows the calibration caveat. The intended takeaway is not that the controller is clinically ready, but that sequential evidence control is measurable, auditable, and meaningfully different from simply feeding the VLM all available modalities.

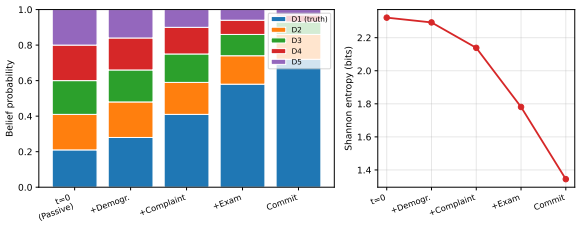

Figure 7: Additional trajectory visualization showing how belief states change as channels are acquired.

Figure 8: Attention heatmap used to analyze distractor sensitivity in open-weight backbones.

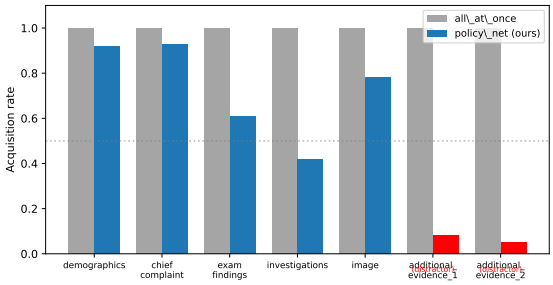

Figure 9: Channel-selection patterns learned by the acquisition policy.

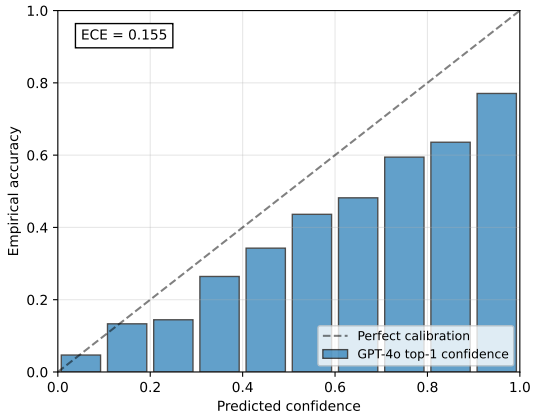

Figure 10: Reliability analysis for belief-derived policy features.

E Appendix Summary

The appendix provides three layers of evidence beyond the main paper. First, the benchmark and stress-test tables show that the central trend is not tied to one backbone, dataset, or baseline family: learned acquisition improves the quality–cost frontier while remaining competitive with prior sequential-diagnosis systems. Second, the robustness analyses show that the gain is not explained by a single prompt template, a brittle probability parser, or a fixed cheap-first schedule; PolicyNet uses trajectory state, cost, and uncertainty signals to adapt acquisition depth. Third, the mechanism and safety analyses clarify why selective acquisition can outperform Full Modality: additional channels can distract the frozen VLM, so stopping can preserve a correct differential rather than merely save cost. At the same time, the failure taxonomy, subgroup table, and multidimensional cost analysis identify the remaining risks that would need prospective validation, local calibration, and human oversight before clinical deployment. Taken together, the appendix supports the main claim that ActiveMedAgent should be evaluated as a sequential evidence-control system: its value lies not only

in final accuracy, but in whether the path to that diagnosis is necessary, auditable, and cost-aware.