From Few-Shot Segmentation to Clinician-in-the-Loop Medical Image Analysis:

Decidable Self-Assessment as the Precondition for Interaction and Adaptation

Yazhou Zhu
rangerzhuyazhou@gmail.com

Abstract

Few-shot medical image segmentation (FSMIS) seeks to delineate unseen structures from a small support set, but its standard formulation fixes task-defining evidence before inference. This assumption is fragile when query cases exhibit acquisition shift, atypical pathology, ambiguous boundaries, or poor image quality. A natural response is to add clinician interaction and rapid adaptation. We argue that this response, stated at that level, is incomplete: promptable interfaces and test-time adaptation are already available, so the binding constraint is not the ability to ask or the ability to change, but the ability to *decide* whether asking or changing is warranted for the case at hand. Both capabilities depend on confidence estimates that are calibrated on typical data and degrade precisely in the rare regime that motivates them. This Perspective therefore reframes FSMIS as a three-layer problem. The base layer is decidable self-assessment: a system must separate errors that a bounded intervention can repair from errors that no admissible intervention can reach. We formalize this through the *correctable set*, the family of segmentations reachable under a bounded update operator and the remaining interaction budget, and show that it supplies a decision rule for accept, query, and defer, rather than a cost comparison alone. The middle layer is selective interaction: with a static support budget K and a distinct interaction budget B, queries are chosen by response-conditioned net expected value of information, and the default action is not to ask. The upper layer is bounded adaptation, where the operative property is reversibility rather than speed. We further promote cross-case experience from an optional extension to an explicit second memory layer, and argue that what transfers across rare cases is not a mask prior but a correction prior over failure modes, gated by cross-reader reproducibility. We restate the research agenda as six hypotheses with an explicit dependency order, in which the value of querying is conditional on the validity of multi-level risk estimation, and specify a minimal pilot that can falsify the foundational claim before any clinician study. The central claim is not that interaction resolves domain shift, but that scarce expert attention should be spent only where a bounded intervention is expected to reach a clinically better outcome.

Keywords: few-shot learning; medical image segmentation; cross-domain generalization; clinician-in-the-loop; selective prediction; uncertainty calibration; correctable failure; foundation models

1 Introduction

Few-shot medical image segmentation is motivated by a simple constraint: expert masks are expensive, whereas new anatomical structures, pathologies, scanners, and imaging protocols appear continually. The canonical answer conditions a segmentation model on a small set of labeled support images and infers the mask for a query image. This framing has generated substantial progress, but it implicitly assumes that the initial support set contains enough evidence to resolve the query. That assumption is weakest precisely in the cases for which adaptable medical imaging systems

are most valuable: rare disease, atypical anatomy, postoperative change, severe artifact, unfamiliar acquisition, or a genuinely ambiguous boundary.

A widely shared intuition is that the next step is therefore twofold: let the model interact with the clinician, and let it adapt quickly from what the clinician says. We agree with the direction and disagree with the level at which it is usually stated. Interaction and adaptation are *execution* capabilities, and both are, in isolation, largely solved. Promptable segmentation supplies a general interaction interface [50, 19]; test-time adaptation, adapters, and prompt optimization supply mechanisms for fast change [52, 66, 31]. Section 4 documents that human correction, context accumulation, support selection, deferral, and preference alignment are all individually established. If the missing ingredient were the ability to ask or the ability to change, the problem would already be closed.

What is missing is the *judgment* that governs both: knowing when to ask, what to ask, and when not to change. These three judgments share a single dependency, namely an estimate of residual risk for the present case, and that estimate fails in exactly the regime under discussion. Modern networks are overconfident [44], calibration degrades under distribution shift [45], and atypicality scores measure how unusual a case is rather than how likely it is to fail [82]. An interaction policy built on such a signal will interrupt clinicians on unusual but easy cases and release confident errors on familiar but hard ones. An adaptation rule built on it will update on the wrong evidence and cannot detect that it has done so.

This Perspective consequently organizes clinician-in-the-loop few-shot analysis into three layers with a strict dependency order.

Layer 1: decidable self-assessment. The system must distinguish an error that a bounded intervention can repair, an error that no admissible intervention can reach, and an apparent uncertainty that does not correspond to error at all. Without this layer, interaction degenerates into asking the clinician to keep clicking, and adaptation degenerates into drift.

Layer 2: selective interaction. Given the first layer, the operative virtue of interaction is parsimony rather than speed. Clinician attention is the most expensive resource in the system, and an unnecessary interruption can cost more clinically than a minor boundary deviation. The default action is to *not* ask, and a query must justify itself.

Layer 3: bounded adaptation. Given the first two layers, the operative virtue of adaptation is reversibility rather than speed. One-sample updates from possibly erroneous feedback, whose effects can propagate beyond the edited region, are a structural hazard; "fast" is not the right objective.

Stated as a single sentence, the position of this paper is that the next step for few-shot medical image analysis is not to extract a stronger representation from fewer examples, but to give a system a decidable account of its own failure, so that it calls a human only at the few moments when a human can actually help, and changes itself only within the range that evidence supports.

Contributions. First, we formalize static task evidence and sequential clinician feedback as separate resources, K and B, which prevents ordinary interactive segmentation from being relabeled as few-shot learning. Second, we introduce the correctable set and an associated decidability condition, which converts accept/query/defer from a cost comparison into a partially decidable test and supplies a formal definition of correctable versus non-correctable failure. Third, we narrow the novelty claim explicitly: what remains untested is response-conditioned selection *among heterogeneous feedback channels* coupled to release decisions, not interaction, deferral, or adaptation

as such. Fourth, we promote cross-case experience to an explicit second memory layer and argue that the transferable object is a correction prior over failure modes rather than a mask prior over rare diseases, with cross-reader reproducibility as the write gate. Fifth, we address the cold-start problem of an unknown response model through a pessimistic value of information. Sixth, we restate the agenda as six hypotheses with an explicit dependency order and specify a minimal pilot that can falsify the foundational hypothesis before any clinician is recruited. Segmentation is the motivating and formally developed case; extension to reconstruction, detection, or diagnosis is outside the present claims. The synthesis is a selective Perspective rather than a systematic review, and literature claims are restricted to the cited design space.

2 Problem Foundations: Scarcity, Shift, and Ambiguity

2.1 Three forms of scarcity

The motivation for FSMIS is normally described as label scarcity. In clinical edge cases this is only one of three interacting constraints. *Data scarcity* reflects the specialist labor required for dense masks. *Coverage scarcity* arises because unusual protocols, rare pathology, atypical anatomy, and severe artifacts are intrinsically uncommon. *Attention scarcity* reflects the fact that a clinician cannot inspect and correct every model output. Improving average Dice under a fixed one-shot protocol addresses part of the first constraint and leaves the other two implicit. The three are not independent: coverage scarcity is what makes a case hard, and attention scarcity is what makes it costly to resolve, so a policy that optimizes one in isolation will typically spend the other.

Acquisition, representation, and interpretation are also coupled. MRI pulse sequences, field strengths, reconstruction pipelines, and vendor-specific choices alter contrast and texture; CT phase, dose, reconstruction kernel, and hardware alter noise and edge appearance. Pathology can simultaneously change target morphology and the clinical meaning of an uncertain boundary. Consequently the same apparent segmentation failure may arise from missing task evidence, a shifted feature geometry, an ambiguous reference standard, or a model limitation. Table 1 separates these sources; a clinically useful controller must not treat them as interchangeable, and in particular must not collapse them into a single out-of-distribution label.

2.2 Two independent supervision budgets

Let K denote the number of static support image-mask pairs that initially specify an unseen class or task. Let $B \in \mathbb{R}^m_+$ denote a vector of hard interaction limits available *after* observing the query, such as maximum elapsed time, number of interruptions, or corrected area; $b(a) \in \mathbb{R}^m_+$ measures the same resources for action a, and inequalities are componentwise. Each experiment should prespecify the active components and units of B. A separate soft cost $c(a)$ represents burden not already hard-constrained, with accounting rules that prevent double counting.

"Few-shot" refers to small K; the extension proposed here asks how B should be spent. This separation does two kinds of work. It prevents ordinary interactive segmentation from being relabeled as few-shot learning, and it prevents a gradually accumulated target-domain dataset from being mistaken for a bounded per-case episode. Because the distinction is load-bearing for every subsequent claim, it is stated here rather than deferred: a method that consumes unbounded B has not solved a few-shot problem, whatever its K.

Table 1: Distinct sources of difficulty and their consequences for a clinician-in-the-loop policy. The categories can co-occur but should not be collapsed into a single OOD label. The final column anticipates Section 5.2: only the first two rows are generally correctable by bounded intervention.

Source	Operational meaning	Typical evidence	Policy implication
Label-space novelty	The target structure or lesion was not represented among meta-training classes.	Small labeled support set; semantic prompt; anatomical relation.	Improve task specification; query semantic identity when the support set is insufficient.
Domain shift	Meta-training and deployment differ in scanner, protocol, site, modality, or appearance.	Domain-robust features; acquisition metadata; target support examples.	Reweight or adapt representation while monitoring transfer and retention.
Case-level OOD	The individual case lies outside the validated deployment envelope.	Atypicality score, predicted mask quality, provenance, clinical consequence.	Use as a routing signal; do not equate atypicality with failure probability.
Clinical ambiguity	More than one contour can remain plausible even under matched acquisition.	Multiple readers, provenance, task intent, downstream tolerance.	Represent disagreement, ask a task-specific question, or defer to adjudication.

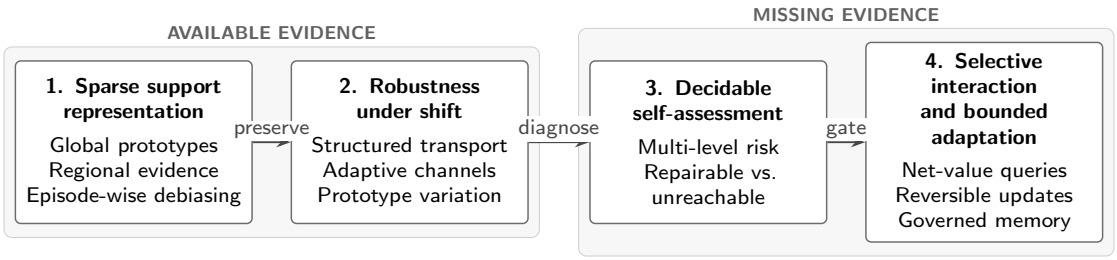

Figure 1: Conceptual trajectory. The first two stages improve what a small support set retains; the third is the precondition introduced here, and the fourth is what it licenses. The ordering is logical rather than a claim of strict historical succession.

2.3 Central thesis

Reliable segmentation for rare, ambiguous, and shifted cases should be formulated as a sequential decision problem in which the first decision is epistemic. A model uses a small support set as initial task evidence, estimates not only what remains uncertain but whether that uncertainty is *reachable* by an admissible intervention, and spends limited clinician attention only when the expected reduction in clinical risk justifies the cost. Reliability must emerge from externally validated decision rules, response-conditioned querying, bounded and reversible adaptation, and principled deferral, not from model confidence or model scale alone.

3 From Fixed Support to Cross-Domain Robustness

To motivate a decision-theoretic framework, limitations in how sparse evidence is *represented* must first be distinguished from limitations caused by evidence that is *absent*.

3.1 The canonical few-shot assumption

One-shot semantic segmentation established that a single annotated example could specify a new dense-prediction task at inference time [1]. Prototypical Networks supplied a metric-learning principle in which each class is represented by the center of its support embeddings [2]; PANet translated this principle to dense prediction through masked prototypes and support-query alignment [3]. In medical imaging, volumetric task conditioning, self-supervised representation learning, and anomaly-inspired foreground-background modeling have relaxed this idealized geometry [4, 5, 6]. The common pipeline nevertheless remains fixed: encode the support set and query, compress the labeled foreground into a task representation, and transfer that representation to an unseen class.

Medical images expose the statistical weakness of this compression. Foreground and background are highly imbalanced, labeled base structures can be scarce, and a structure rarely forms one compact cluster across slices, patients, modalities, and disease states. The support mask is therefore not merely small; it is an incomplete and potentially biased description of a heterogeneous target.

3.2 Representing heterogeneous support evidence

A sequence of within-domain studies can be read as complementary relaxations of the single-prototype assumption rather than as a list of architectures. Three moves recur. *Regional decomposition* replaces one global prototype with several representative subregions, on the premise that a structure is a mixture rather than a point [7, 8]. *Relation-aware refinement* lets support and query features rectify one another instead of treating the support summary as fixed [7]. *Episode-conditioned filtering* challenges masked average pooling directly by removing episode-specific foreground features before prototype construction [9]. Comparable strategies appear outside this line in the broader domain-generalization literature, where feature-level regularization and semantic-alignment objectives play the same role of preventing a single summary statistic from absorbing nuisance variation [83, 84].

The theoretical continuity is more important than the individual method: what must a few labeled images retain in order to represent a heterogeneous anatomical structure? All three moves imply that elements of a support mask have unequal relevance and that relevance depends on the target case. None of them can create relevance where the support set contains none.

3.3 Cross-domain FSMIS

CD-FSMIS introduces a second generalization axis: source-trained meta-knowledge must transfer to both unseen classes and an unfamiliar target domain [10]. Even when test support and query images originate from the same target domain, the encoder and similarity function learned on source data may provide a poorly aligned coordinate system; an acquisition mismatch between test support and query compounds the problem.

Recent work addresses complementary failure points. RobustEMD replaces point-wise prototype comparison with structured transport between decomposed support and query features, with texture- and structure-aware weighting suppressing domain-sensitive nodes [11]. Dynamic Semantic Matching performs support-query reweighting, dynamically selects domain-robust channels, and estimates semantic centers from dual perspectives [12]. Adversarial Prototypical Perturbation constructs gradient-derived perturbations from intra- and inter-class variation objectives and combines them with local-imbalance-aware whitening [13]. MAUP uses DINOv2 features and a frozen Segment Anything Model with multi-center prompt generation, uncertainty-aware prompt selection, and adaptive prompt optimization without parameter training [14]. These branches are complementary: one improves the representation of sparse task evidence, another improves matching under shift, and the foundation-model branch supplies a promptable interface.

3.4 What better representation cannot recover

Better support representations and domain-robust matching preserve available evidence. They cannot recover case-specific anatomy or clinical intent that the support set never contained, and none of them establishes whether a missing item of evidence is worth requesting from a clinician. This is the boundary at which representation research ends and decision research begins, and it is the boundary this Perspective is about.

4 Adjacent Paradigms and the Integration Gap

4.1 Interaction is necessary but not sufficient

Interactive medical segmentation has established that points, scribbles, boxes, and image-specific optimization can correct a current output [15, 16, 17, 18]. ScribblePrompt demonstrated generalization to unseen biomedical tasks and improved annotation efficiency relative to SAM ViT-B in a study involving 16 academic-hospital imaging researchers who were shown target masks [19]. This supports known-target annotation efficiency, not independent patient-level boundary judgment.

Interaction is also not universally reactive. PseudoClick predicts candidate next clicks automatically [20]; MECCA combines action confidence with reinforcement learning to guide the next interaction region [21]; sequential-memory models encode the order of user corrections [22]. Interactive few-shot learning uses simulated point corrections for regularized test-time optimization [23], while IFSS-Net propagates expert-seeded information in volumetric ultrasound [24]. Correction-driven adaptation across images and domain change is likewise established [25].

More recent systems occupy additional parts of the design space: clinician-corrected predictions for cross-center test-time adaptation [26]; multiple plausible masks for iterative refinement [27]; image ordering as context grows [28]; support selection and in-context failure detection [29]; pixel- or region-level deferral to multiple experts [30]; online adaptation from user-refined masks under distribution shift [31]; and uncertainty-guided candidate selection for preference alignment [32]. Their evidence levels differ: several are preprints, one is a workshop contribution, and others are peer-reviewed conference or journal studies.

We draw the conclusion explicitly, because it constrains what this Perspective may claim. Human correction, adaptation, context accumulation, support selection, deferral, and preference alignment are *not novel in isolation*. Any position paper whose thesis is "few-shot systems should interact and adapt" is therefore restating existing capability. The open problem must lie elsewhere.

4.2 Active learning, in-context learning, and deferral

Pool-based active learning asks which samples or regions should be labeled to improve a future global model [77]. Information-theoretic and Bayesian acquisition formalize informativeness through expected information gain or epistemic uncertainty [33, 34]; medical systems combine these signals with representativeness, consistency, diversity, or simplified interactive labels [35, 36]. Interactive segmentation usually targets the current mask, whereas in-context segmentation uses completed examples to specify a task without gradient-based retraining: UniverSeg demonstrates broad in-context medical segmentation [37], MultiverSeg accumulates completed examples as context [38], and the ordering of these examples can alter both accuracy and effort [28]. Deferral has an established decision-theoretic basis in learning-to-defer [78, 62], and classical information value theory asks whether an observation changes a decision [39].

Table 2: Primary objectives of related paradigms. The rows are not mutually exclusive, and existing methods already bridge several categories; the final row states what the present framework adds rather than what it invents.

Paradigm	**When supervision arrives**	**What supervision changes**	**Central limitation or advance**
Few-shot segmentation	Fixed support before inference	Task representation or prototypes	No mechanism to acquire missing case evidence.
Interactive segmentation	Corrections during inference	Current mask; sometimes current image representation	Optimizes refinement rather than joint risk and release.
Pool-based active learning	Selected labels during development	Future global model	Does not manage current-case release risk.
In-context segmentation	Examples accumulate as context	Task specification without weight updates	Selection, calibration, and deferral remain incompletely coupled.
Learning to defer	At routing time	Who decides	Routes on predicted error, not on predicted repairability.
Proposed framework	Model-initiated multimodal feedback under budget, gated by reachability	Case, task, and domain representation, plus governed memory	Intended to jointly optimize and evaluate error, effort, escalation, and retention.

4.3 The narrowed claim

Given the preceding two subsections, we state the residual gap in the narrowest form that survives the literature. What remains insufficiently tested is not another interaction primitive, and not the general proposition that interaction helps. It is a controlled integration of three specific elements:

1. a *decidable* account of which current-case errors a bounded intervention can reach, evaluated under explicit few-shot, cross-domain, and case-level OOD conditions;
2. response-conditioned selection *among heterogeneous feedback channels* whose bandwidth and semantics differ, coupled to accept/query/defer release decisions rather than to mask refinement alone;
3. joint evaluation of risk, effort, adaptation stability, and escalation on the same trajectory.

Net expected value of information is itself classical [39] and standard in active learning; we claim no novelty for it. The claim is the coupling of channel-level value estimation to a release decision that is gated by a reachability test.

5 Decision-Theoretic Formulation

5.1 Episode state, actions, and bounded update

An episode begins with a query image or volume x, K support image-mask pairs S_K, an initial model f_0, and interaction budget B. At step t the state

$$s_t = (x, S_K, h_t, f_t, B_t, u_t) \tag{1}$$

contains the interaction history h_t, current model or episode representation f_t, remaining budget B_t, and a hierarchy of local, structural, and case-level risk evidence u_t. The controller chooses $d_t \in \{A, Q, D\}$: accept the current mask, query the clinician through action $a_t \in \mathcal{A}_Q$, or defer to full review. A query action jointly specifies a *location*, a *modality*, and a *question*. The response z_t may include disagreement, correction error, no response, or latency. A bounded transition operator $\mathcal{T}(s_t, a_t, z_t)$ can update prompts, prototypes, memory, latent features, or a restricted parameter subset. The episode terminates at stopping time τ.

Acceptance and deferral have different consequences. Acceptance incurs clinically weighted loss from a released mask. Deferral incurs full-review time and the residual loss of the expert or joint result; it is neither free nor perfect. For feasible policies $\Pi(B) = \{\pi : \sum_{t<\tau} b(a_t) \preceq B\}$, where $\preceq$ is componentwise, consider

$$\pi^\star = \arg\min_{\pi \in \Pi(B)} \mathbb{E}_\pi \Big[\mathbb{I}(d_\tau = A)\, L_{\mathrm{clin}}(\hat{y}_\tau, y) + \mathbb{I}(d_\tau = D)\{C_D + L_{\mathrm{clin}}(y_\tau^D, y)\} + \lambda_C \sum_{t<\tau} c(a_t) + \lambda_R \big[L_{\mathcal{R}}(f_\tau) - L_{\mathcal{R}}(f_0) - \varepsilon_R\big]_+ \Big]. \tag{2}$$

Here $b(a)$ records prespecified hard resource use, whereas $c(a)$ grades residual burden not already represented in the active components of $b(a)$; the cost ledger and normalization are fixed before evaluation to prevent double counting. The final term penalizes degradation beyond tolerance ε_R on an immutable protected reference set $\mathcal{R}$. Clinical losses must be elicited for a defined downstream decision: a missed lesion, a critical boundary violation, and a harmless surface discrepancy cannot share an arbitrary common weight.

5.2 The correctable set and the decidability condition

Equation (2) is a cost comparison. On its own it does not say whether a query *can* help, only whether it would be cheap. This matters because of an information-theoretic asymmetry that the interaction literature rarely states: a click carries a few bits, whereas domain shift induces a high-dimensional mismatch in the representation. Low-bandwidth input cannot inject high-dimensional information. What it can do is *collapse ambiguity* among hypotheses the model already entertains. Interaction therefore helps only when an acceptable solution lies within the range the system can reach.

We make this precise. Let $\mathcal{A}_Q(B_t) = \{a \in \mathcal{A}_Q : b(a) \preceq B_t\}$ be the admissible query actions under the remaining budget, and let $\mathcal{Z}(a)$ be the support of plausible responses to a.

Definition 1 (Reachable set)**.** *The one-step reachable set at state s_t is*

$$\mathcal{M}_1(s_t) = \{\hat{y}(\mathcal{T}(s_t, a, z)) : a \in \mathcal{A}_Q(B_t),\ z \in \mathcal{Z}(a)\},$$

and the budget-limited reachable set $\mathcal{M}(s_t)$ is its closure under admissible action sequences whose cumulative cost satisfies $\sum b(a) \preceq B_t$.

Definition 2 (Correctable set)**.** *Given a clinically elicited acceptability criterion $\mathcal{C}$ with tolerance η, the correctable set at s_t is $\mathcal{M}^{\mathcal{C}}(s_t) = \{\hat{y} \in \mathcal{M}(s_t) : \mathcal{C}(\hat{y}; \eta) = 1\}$. The current case is* correctable *at s_t if $\mathcal{M}^{\mathcal{C}}(s_t) \neq \emptyset$.*

Three consequences follow. First, correctability is defined relative to the update operator and the remaining budget, not as a property of the image; enlarging $\mathcal{T}$ or B can make a case correctable, which is precisely what a safety argument must bound. Second, the qualitative distinction between repairable and unreachable failure acquires a formal meaning rather than resting on descriptive language. Third, the accept/query/defer rule gains a structural component:

Proposition 1 (Reachability gate)**.** *If $\mathcal{M}^{\mathcal{C}}(s_t) = \emptyset$, then for every $a \in \mathcal{A}_Q(B_t)$ the post-response state cannot attain an acceptable mask, and any strictly positive soft cost $c(a) > 0$ implies* $\mathrm{nEVI}(a \mid s_t) \leq 0$*. The correct action is therefore D (or A, if the current mask is already acceptable), not Q.*

The proposition is deliberately weak, and its usefulness lies in what it excludes rather than in what it proves. In practice $\mathcal{M}^{\mathcal{C}}$ is not computable, because $\mathcal{C}$ depends on the unobserved reference standard. It must therefore be replaced by an estimated reachability score $\hat{\rho}(s_t) \approx \Pr[\mathcal{M}^{\mathcal{C}}(s_t) \neq \emptyset]$, trained on simulated and recorded interaction traces where the post-hoc reachable outcome is known. Two properties make $\hat{\rho}$ a more tractable estimation target than case-level accuracy itself. It is a function of the model's own output geometry and the admissible operator, both of which are observable at inference time, and it does not require identifying the correct contour, only whether an acceptable one is within range. We regard the empirical behavior of $\hat{\rho}$ under shift as the single most informative experiment this framework proposes; if $\hat{\rho}$ cannot be estimated better than chance at an external site, the three-layer architecture fails at its base and the rest of the agenda is moot.

Remark 1. *The reachability gate also clarifies the relation between this framework and deferral research. Learning to defer typically routes on predicted model error [78, 62]. The gate routes on predicted* irreparability*, which is a different quantity: a case may be badly wrong and trivially correctable, or nearly right and unreachable because the residual disagreement is irreducible ambiguity rather than error.*

5.3 Response-conditioned value of clinical feedback

Uncertainty alone is not a reason to interrupt a clinician. A query is useful only if a plausible response is expected to change the mask, lower residual risk, or determine that the case requires deferral. Define the retention penalty $P_{\mathcal{R}}(s) = \lambda_R[L_{\mathcal{R}}(f_s) - L_{\mathcal{R}}(f_0) - \varepsilon_R]_+$. Let $J_A(s)$ and $J_D(s)$ denote the expected costs of accepting and deferring, each including $P_{\mathcal{R}}(s)$, and let $J_{\mathrm{stop}}(s) = \min\{J_A(s), J_D(s)\}$. The net one-step expected value of information is

$$\mathrm{nEVI}(a \mid s_t) = J_{\mathrm{stop}}(s_t) - \mathbb{E}_{z \sim p(z \mid s_t, a)}\big[J_{\mathrm{stop}}(\mathcal{T}(s_t, a, z))\big] - \lambda_C c(a). \tag{3}$$

A greedy controller selects

$$a_t^g = \arg\max_{a \in \mathcal{A}_Q(B_t)} \mathrm{nEVI}(a \mid s_t) \quad \text{subject to} \quad \hat{\rho}(s_t) \geq \rho_{\min}, \tag{4}$$

and queries only when the maximum is positive. The reachability constraint is stated explicitly in (4) because it is not implied by positivity of nEVI under a misspecified response model: a response model that is optimistic about what a click achieves can manufacture positive value for an unreachable case. The non-myopic form of the controller is given in Appendix A; it matters because one answer can alter the value of later questions, but it is not required to state the position of this paper.

Channel semantics differ and should not be conflated. A click is inexpensive but low bandwidth; a boundary correction conveys geometry; a reference case or short text may convey domain context or task intent. The last two are hypotheses to be tested, not capabilities implied by click- or box-prompt models. Because J_A and J_D contain $P_{\mathcal{R}}$, a query cannot acquire positive value merely by sacrificing protected capabilities. Candidate updates must also be reversible: when a separately frozen monitor rejects a post-response state, $\mathcal{T}$ returns the pre-update state and routes the case to another query or to deferral. Equations (2)–(4) are normative until their risk, response, cost, and update models have been estimated for a defined clinical use case and separately safeguarded.

5.4 Cold start: acting under an unknown response model

Equation (3) requires $p(z \mid s, a)$, the distribution of clinician responses. At deployment onset this distribution is unknown, and it is exactly then that a miscalibrated policy does the most damage to trust. We therefore treat the response model as an explicit staged object rather than an assumption.

Let $\mathcal{P}_a$ be an ambiguity set of response distributions consistent with prior evidence for action a: for instance, all distributions within a divergence ball of a simulated-user model, or all mixtures of an expert-agreement model and a no-response model with bounded weight. Define the pessimistic net value

$$\underline{\text{nEVI}}(a \mid s_t) = J_{\text{stop}}(s_t) - \sup_{p \in \mathcal{P}_a} \mathbb{E}_{z \sim p}\big[J_{\text{stop}}(\mathcal{T}(s_t, a, z))\big] - \lambda_C c(a), \tag{5}$$

and require $\underline{\text{nEVI}}(a \mid s_t) > 0$ during the cold-start phase. Because $\underline{\text{nEVI}} \leq \text{nEVI}$, this is conservative by construction: it suppresses queries whose value depends on an optimistic reading of how clinicians respond, at the cost of some missed opportunities. The phase transition should be prespecified rather than discretionary. A workable rule is to shrink $\mathcal{P}_a$ toward the empirical response distribution once a minimum number of observed responses per channel and per stratum is reached, with the minimum fixed in advance and the shrinkage schedule reported; strata that never reach the minimum remain permanently pessimistic. Sensitivity of the resulting policy to $\mathcal{P}_a$ should be reported alongside its performance, since a policy whose behavior is invariant to the ambiguity set is not actually using response information.

5.5 Accept, query, or defer

Querying is appropriate when risk is elevated, reachability is plausible, and a compact intervention is likely to resolve the case. Deferral is appropriate when the case lies outside the validated adaptation envelope, support evidence is inadequate, ambiguity is irreducible, reachability is low, or another interaction has low expected value. Acceptance is permitted only by an empirical decision rule whose clinically defined threshold is frozen after stratified external validation. Selective prediction provides the risk-coverage perspective [40]; selective medical segmentation demonstrates why performance on a confident subset differs from average accuracy [41]. Distribution-free risk control supplies a language for stating what a frozen threshold does and does not guarantee [79], and its exchangeability assumptions are precisely what external-site deployment strains.

The clinician is not assumed to be an infallible oracle. Boundaries may be ambiguous, experts may disagree, and a hurried correction may be wrong. Feedback should retain provenance, modality, timing, and confidence. STAPLE models a probabilistic consensus and rater-specific performance under conditional-independence assumptions [42]; probabilistic segmentation, hierarchical probabilistic models, and MedUHIP illustrate alternative ways to represent several plausible masks [43, 80, 27]. The framework must distinguish annotation error, systematic reader preference, irreducible ambiguity, and legitimate use-case-specific contouring rather than collapsing them into one latent truth. Section 6 makes this distinction operational, because it determines what may be written to memory.

6 A Two-Layer Memory Architecture

The second capability commonly proposed for rare-case few-shot systems is learning from clinician experience accumulated over edge cases. We treat this as a first-class part of the architecture rather than an optional extension, but with a change of learning target, because the naive version is statistically self-defeating.

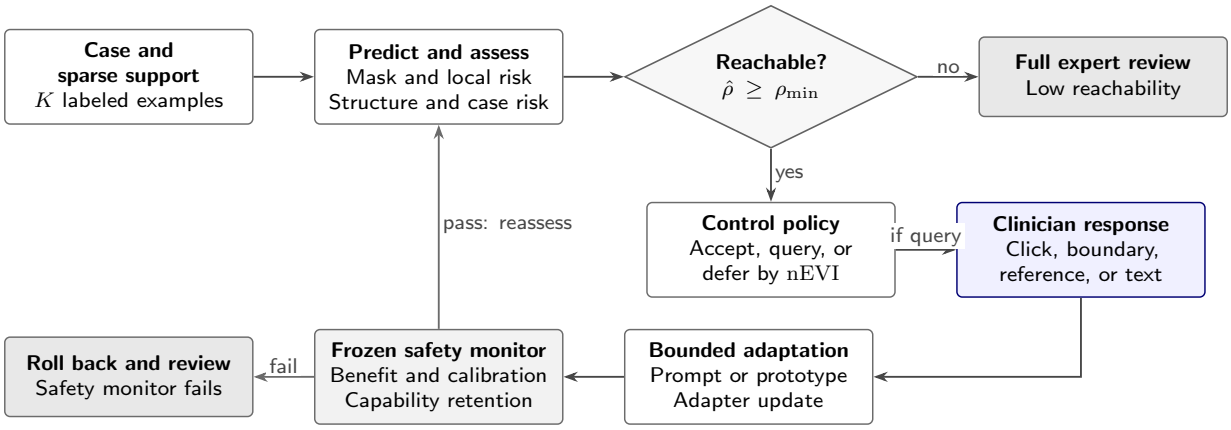

Figure 2: Selective interactive adaptation loop with an explicit reachability gate. A foundation model can supply the representation and prompt interface, but a separate controller governs interaction and stopping, and adaptation cannot certify its own safety: each update is independently reassessed against clinical risk and protected capabilities.

6.1 Why the naive target is unlearnable

Rare cases are rare. If the object to be learned is $p(\text{mask} \mid \text{rare presentation})$, then each class of presentation may be encountered once or twice in an institution's history, and no amount of governance repairs a sample size of one. Worse, the presentations are heterogeneous in ways that do not pool: an unusual postoperative anatomy and an unfamiliar reconstruction kernel share nothing at the level of appearance.

The resolution is to change what is being estimated. Failure *modes* are far fewer than diseases. Boundary overshoot into an adjacent bright structure, topological fragmentation of a thin structure, missed small lesions near a high-gradient interface, and spurious attachment to an artifact recur across anatomies, modalities, and pathologies. What generalizes across rare cases is therefore not clinical knowledge but a *correction prior* over geometric and semantic failure modes: an estimate of $p(\text{correction direction} \mid \text{failure mode}, \text{context})$. This target has a workable effective sample size because every case that exhibits a mode contributes, regardless of its diagnosis. It also has a natural interface to Section 5.2, since the modes that admit a compact correction direction are largely the modes that populate the correctable set.

6.2 Layer A: the per-case episode

The episode layer is the one formalized in Equations (2)–(5). Updates are episode-local and reversible: prototype revision, prompt optimization, adapters, or small normalization modules, combined with source anchors, trust regions, and snapshots. Per-case test-time adaptation can recover performance across scanner and protocol shifts [52], and source-free domain adaptation addresses a distinct target-cohort setting in which source images cannot be retained [53]. Continual test-time adaptation and entropy minimization can accumulate pseudo-label error or forget source capability in changing streams [54, 55], although much of this evidence comes from non-medical benchmarks. Correction-driven adaptation is established [25], and OAIMS directly studies online adaptation from refined masks under medical distribution shift [31]. These are baselines, not novelty claims. Within this layer, one intervention should correct the present mask and, where justified, clarify the task for later slices in the same volume; nothing written here survives the episode.

Table 3: The two memory layers differ in unit, horizon, objective, and governance. Conflating them is what allows an accumulating target-domain dataset to be reported as few-shot performance.

	Layer A: episode	**Layer B: experience**
Unit of storage	Prompts, prototypes, adapters for the current case	Failure-mode → correction-direction priors, de-identified
Horizon	Single case; discarded at τ	Multi-case; explicit retention period
Objective	Equation (2)	Expected reduction in future episode cost
Update timing	Online within the episode	Periodic, batched, audited
Write gate	Frozen monitor; rollback	Cross-reader reproducibility (Section 6.4)
Governance	Ephemeral; no consent beyond care	Consent, provenance, deletion, protected-task tests
Failure mode	Within-episode drift	Institutional bias entrenchment

6.3 Layer B: the cross-case experience layer

The experience layer holds correction priors, not cases. Its unit of storage is a tuple of failure-mode descriptor, context descriptor, correction direction, and provenance, deliberately stripped of case identity. Its objective is not the per-episode cost in (2) but the expected reduction in future episode cost, evaluated over a separately governed horizon with its own resource budget. Its updates are periodic and audited rather than online. Table 3 contrasts the two layers; the separation exists so that a claim about long-run learning cannot be smuggled into an evaluation of per-case behavior, which is the most common ambiguity in this literature.

6.4 The write gate: cross-reader reproducibility

Section 5.5 required distinguishing annotation error, systematic reader preference, irreducible ambiguity, and legitimate use-case-specific contouring, but a principle is not a mechanism. We propose reproducibility across readers as the operational criterion, because it is measurable and because it aligns the destination of a correction with the scope of its validity.

Shared write. A correction pattern that is independently reproduced by multiple readers under comparable context is admitted to the shared experience layer. Reproducibility is assessed on the failure-mode descriptor and correction direction, not on pixel agreement, and requires a prespecified minimum number of distinct readers and cases.

Personal write. A pattern reproduced by a single reader across cases, but not by others, is admitted only to that reader's personalization profile and never influences others' outputs. This is the correct destination for legitimate use-case-specific contouring and for systematic reader preference.

No write. A pattern observed once remains in the episode and expires with it. This is the default, and it is where genuine annotation error is expected to land.

Explicit ambiguity. A context in which readers reproducibly *disagree* is recorded as such and routes future similar cases toward deferral or adjudication rather than toward a single prior [42, 43, 80].

The rule has a cost, which should be acknowledged: it is conservative for institutions with few readers, and it will systematically delay learning from genuine expertise held by a single specialist. We regard this as the correct default in a safety-critical setting, but the threshold is a parameter to be reported, and its effect on learning rate is itself an experimental quantity.

6.5 Governance of the experience layer

Persistent adaptation raises privacy and governance questions because site- or clinician-specific memory may encode identifiable or institution-specific information. Retained representations require access control, audit trails, and deletion mechanisms. Storing correction priors rather than cases mitigates but does not eliminate this: a sufficiently specific failure-mode descriptor can be re-identifying. The experience layer therefore requires its own consent basis, protected-task tests before any interaction may influence later patients, and a demonstrated transfer benefit; absent measurable transfer, forgetting is the safer default. The scientific question is not whether a model can keep learning, but under what evidence and governance conditions retaining an interaction is safer than discarding it.

7 Design Principles and Research Directions

7.1 Calibrated recognition of correctable and non-correctable failure

The first direction estimates residual risk at three linked levels. Voxel- or patch-level uncertainty should localize candidate errors and support spatial querying. Structure-level uncertainty should reveal missing components, topology violations, boundary failures, and small lesions that an average heat map can hide. Case-level risk should govern release or deferral. Accuracy and confidence must remain separate: modern networks are often overconfident [44], calibration degrades under distribution shift [45], aleatoric and epistemic components behave differently [81], and medical segmentation studies demonstrate both the promise and the limits of uncertainty for quality control [46, 47, 48].

To this we add the reachability estimate $\hat{\rho}$ of Section 5.2 as a fourth, orthogonal quantity. An OOD score measures atypicality, not the probability of segmentation failure [82]; predicted mask quality measures failure, not repairability; $\hat{\rho}$ measures repairability, not correctness. An unusual case may be easy, a serious error may remain in-distribution, and a small error may be unreachable. Conditional calibration by site, scanner, modality, structure size, pathology, and interaction step is more informative than pooled voxel calibration. A key negative result is scientifically meaningful: if confidence increases after feedback without improved correctness, the signal is unsuitable for query control.

7.2 Cost-sensitive multimodal feedback as task specification

The second direction treats interaction modalities as semantically distinct supervision channels. Existing systems establish points, boxes, scribbles, masks, and constrained semantic prompts [19, 49, 50]; in-context systems show that labeled image-mask pairs can specify unseen segmentation tasks [37, 38]. These findings do not establish that free clinical text can explain an exception, that a reference case reliably communicates acquisition context, or that accepting a suggested region is valid weak supervision. Those channels should be experimental variables. Earlier work predicts sufficient annotation strength by jointly considering accuracy and human cost [51]; the unresolved problem in CD-FSMIS is response-conditioned selection *among* channels, coupled to accept/query/defer and measured in real effort.

The policy must learn where to ask, what to ask, and how to ask, and it must be evaluated against the strongest available alternative to asking at all. A foreground point may resolve semantic identity; a short correction may resolve geometry; a matched reference case may reveal unfamiliar appearance; and full review may dominate when no compact answer is reliable.

7.3 Bounded, reversible adaptation

The third direction uses clinician feedback without converting deployment into uncontrolled online training, and its design objective is reversibility rather than rate. A separately frozen monitor should control update gates and rollback; operational separation requires fixed monitor parameters, calibration on data disjoint from episode adaptation, and no updating from the same clinician response that triggered the update. Reporting should include rollback frequency, because a system that never rolls back has either a perfect update rule or an inert monitor, and the two are distinguishable only by measurement.

7.4 Selective and personalized clinician-model teaming

Clinical collaboration depends on both partners. Evidence transferred from diagnostic classification shows that decision support can improve performance, while inaccurate advice, including advice presented as AI-generated, can mislead clinicians [56, 57]. A large radiology study found heterogeneous effects across 140 radiologists and 15 chest-radiograph tasks, with AI error a major determinant of harm [58]. Segmentation-specific studies report gains in contouring accuracy or time in some settings [59], but heterogeneous savings across structures and institutions [60] and possible automation bias in prostate radiotherapy contouring [61]. Classification evidence is not proof for interactive segmentation; together, these studies demand direct workflow evaluation. Uncertainty-guided selection for preference alignment [32] is a comparator, not a precedent.

Clinician behavior can enter the system state, but response time or correction style must not be interpreted directly as competence, because both are confounded by case difficulty, interface familiarity, fatigue, and prior allocation. Learning-to-defer provides a basis for routing to humans [78, 62] and for adapting to an unseen expert from a population [63]; it does not validate dense segmentation or transfer clinical responsibility. Personalization should require consent, interpretable features, minimum evidence, bounded exploration, subgroup workload audits, reversibility, and an unconditional clinician override, and its write destination is the personal layer of Section 6.4.

8 Foundation Models as an Enabling Substrate

The decision framework does not prescribe a backbone. Foundation models warrant separate treatment because they can unify representation and interaction interfaces while leaving calibration, stopping, and adaptation safety unresolved.

SAM established a broadly promptable segmentation interface [50]. Points, boxes, and masks can alter an output without task-specific retraining, which makes this interface attractive for clinician-in-the-loop inference. Yet direct retrospective medical evaluation found large variation across tasks and strong dependence on protocol-generated prompt type [64]. MedSAM expands promptable segmentation through large-scale medical adaptation but primarily uses box prompting [65]; Medical SAM Adapter still requires task-specific parameter optimization [66]; and FM-ABS uses active expert cross-labeling in semi-supervised model development rather than real-time clinical inference [67]. Strong specialized pipelines remain competitive baselines and should be reported as such [85]. These studies establish technical feasibility, not calibrated abstention, model-initiated queries, or prospective safety.

A foundation model should therefore be treated as a reusable representation and interaction substrate. MAUP provides an immediate bridge: multi-center representations and uncertainty-aware prompt selection condition a frozen SAM at inference without parameter updates [14]. Its uncertainty score ranks prompts; it has not established patient-level risk calibration or justified

abstention. Notably, a promptable interface also makes the reachable set of Section 5.2 unusually tractable, because the admissible operator is enumerable over prompt perturbations, which is why we regard frozen-SAM pipelines as the natural testbed for the pilot in Section 9.1.

The preferred architecture is modular: a medical foundation encoder provides reusable features; episode memory combines support examples, gated clinician-confirmed masks, and interaction history; a task head proposes the segmentation; separate modules estimate local risk, case risk, and reachability; a controller selects accept, query, or defer; and a constrained adapter modifies only approved components. This separation allows each safety claim to be tested independently. A stronger backbone can improve the initial mask, but cannot conceal a poorly calibrated controller or a drifting update rule.

9 Experimental and Clinical Validation

Because each interaction changes both the prediction and the remaining information state, evaluation must address the complete clinician-model trajectory rather than only its final mask.

9.1 A minimal pilot that can falsify the foundational claim

A purely normative framework invites the objection that it cannot be wrong. We therefore specify the smallest study that could refute its base layer, before any clinician is recruited and without new data collection. It uses an existing CD-FSMIS setting and a frozen promptable backbone [14, 50], for which the admissible operator is enumerable over prompt perturbations and the reachable set can be sampled directly.

1. **Construct reachability labels.** For each held-out case, sample admissible interaction trajectories under a fixed budget using several independent simulated-user policies, and record whether any trajectory attains the acceptability criterion. This yields a binary reachability label without any new annotation.
2. **Estimate $\hat{\rho}$ and test it out of domain.** Train the reachability estimator on source domains only and evaluate discrimination and calibration at held-out target domains, stratified by shift type and difficulty. *This is the falsification point.* If $\hat{\rho}$ does not exceed chance at external targets, or is not better than predicted mask quality alone, the base layer fails.
3. **Compare controllers at matched effort.** Evaluate the gated nEVI controller of (4) against voxel-entropy querying, random querying, a fixed-click heuristic, and passive correction, holding total simulated effort equal. Report risk-coverage curves and quality-effort curves as the two primary displays.
4. **Ablate the gate.** Run the same controller with the reachability constraint removed. If performance is unchanged, the gate is decorative and the central formal contribution of this paper is unsupported.
5. **Probe the response model.** Repeat under deliberately misspecified simulated responders, including a disagreeing responder and a non-responding one, and compare the standard controller against the pessimistic form (5).

This pilot cannot establish clinical benefit, and we do not claim otherwise; simulated users omit search time, navigation, windowing, and cognitive switching. Its value is that four of its five steps can return a negative result that would end the program, which is the property a normative framework most needs.

9.2 Study domains and edge-case construction

Full evaluation should begin with multi-institutional MRI and CT tasks that expose both controlled and naturally occurring shifts: MRI sequence, field strength, vendor, and reconstruction; CT phase, dose, kernel, and hardware; institution-specific protocols; and temporal drift. Task novelty should include unseen structures and lesions. Edge-case strata should include small targets, weak boundaries, atypical morphology, postoperative anatomy, artifact, and poor image quality. Factorial designs should separate acquisition shift from clinical ambiguity wherever possible, so that a query policy is not rewarded for treating every unfamiliar texture as a request for annotation.

Average external performance is insufficient. Cross-hospital studies show variable generalization and exploitation of institution-specific signals [68]; causality-inspired domain generalization demonstrates the need to address acquisition appearance and spurious correlation [69, 84]. Results should be stratified by site, scanner, sequence, target size, pathology, and difficulty, with worst-group performance and low-performing quantiles reported. Held-out institutions and prospective temporal cohorts should remain untouched until the method and thresholds are frozen.

9.3 Evaluation of the coupled system

Overlap measures such as Dice remain useful for continuity but cannot define clinical success. Candidate geometric proxies include normalized surface Dice at a clinically elicited tolerance, average symmetric surface distance, and HD95; their relationship to editing time, inter-reader acceptability, and downstream outcomes must be tested rather than assumed. Metrics Reloaded shows why metric choice should follow target geometry, data hierarchy, imbalance, and decision purpose [70]; clinically applicable contouring studies illustrate tolerance-aware surface evaluation and independent expert comparison [71].

The primary calibration target should be a patient- or structure-level probability that the complete mask fails a prespecified, blinded acceptability criterion. On an independent calibration cohort, reliability curves, Brier and log scores, and uncertainty intervals should be reported by deployment stratum and interaction step. Pooled-voxel expected calibration error is secondary because it can conceal case-level miscalibration. AUROC and AUPRC assess ranking, not threshold safety, so selective-safety reporting should include false-accept rate, sensitivity, positive predictive value, and empirical coverage at a frozen threshold with confidence intervals. A claim of distribution-free risk control must state its exchangeability assumptions explicitly [79]; otherwise "maximum risk" remains an empirical operating target.

A false query has no directly observed label on a single trajectory, because the safe no-query outcome is counterfactual. Operationally it is a query whose adjudicated net expected value is non-positive: the pre-query result was acceptable and realized or estimated benefit falls below a prespecified minimum after measured effort. Randomized query/no-query assignment or validated offline policy evaluation is needed. Interaction outcomes should be expressed against elapsed clinician time, corrected area or slices, interface actions, and latency, with summaries such as time to acceptable mask, area under the quality-effort curve, improvement per minute, and failure within a fixed budget. Decision-curve analysis offers a framework for explicit net benefit rather than inferring clinical utility from geometry alone [72].

Adaptation should be evaluated after every response: immediate gain, calibration change, monotonicity, protected-task retention, within-volume transfer, error propagation, and rollback frequency. The experience layer of Section 6 should be evaluated separately and should additionally report cross-case transfer, write-gate admission rates, and performance over changing domain sequences. An update that increases confidence while increasing clinically weighted error is a safety

Table 4: Evaluation follows the causal chain from local prediction to interaction, adaptation, and workflow. The reachability row is new and is where the framework is most exposed to falsification.

Level	Representative measures	Primary safety question
Voxel or region	Calibration, error localization, boundary distance, Dice.	Does uncertainty identify a clinically meaningful correction?
Structure or case	NSD, ASSD, HD95, missed small targets, failure AUROC/AUPRC.	Can the system recognize a clinically unacceptable complete result?
Reachability	Discrimination and calibration of $\hat{\rho}$; gate ablation; unreachable-case deferral rate.	Can the system tell a repairable error from an unreachable one?
Selective policy	Risk-coverage, false accepts, counterfactual query value, frozen-threshold coverage.	Are release and escalation decisions accountable?
Interaction	Time to acceptable mask, quality-effort AUC, improvement per minute, latency.	Does assistance save expert effort outside simulation?
Adaptation	Immediate gain, calibration change, retention, rollback rate, within-volume transfer.	Does learning help without drift or propagated error?
Experience layer	Cross-case transfer, write-gate admission, subgroup entrenchment audits.	Does retained memory earn its governance cost?
Workflow	Editing time, workload, override, downstream effect, subgroup results.	Does the joint system outperform both components safely?

failure. Downstream endpoints such as volumetry stability, radiotherapy contour acceptability, or derived-measurement sensitivity should be used when the application permits.

9.4 From robot users to clinicians

Simulated users are indispensable for scalable ablation but do not establish clinical interaction efficiency. Many protocols use the ground-truth mask to place the next click and count heterogeneous modalities as one interaction step; they therefore omit search time, navigation, windowing, and cognitive switching, and cannot measure anchoring, fatigue, frustration, or recovery from a misleading output. Interactive segmentation research shows that user protocols and interface design can alter conclusions [73, 74]. Radiotherapy workflow studies support measuring operational and cognitive time, although small-reader studies cannot establish broad effectiveness [75].

Validation should proceed as a ladder: the pilot of Section 9.1; manually collected traces disjoint from the final human test; a reader study on representative edge cases; prospective silent evaluation; and a limited workflow study with stopping and rollback criteria. Reader studies should use a randomized crossover with washout or a balanced incomplete-block design. Manual delineation, review of a raw automatic mask, a conventional interactive tool, and the proposed system should be matched where possible for backbone, initial mask, display, training, latency, and interface. Analyses should model crossed clinician and case effects, prespecify power, use blinded adjudication, and record quality-effort intervals, workload, usability, override, and recovery from deliberately imperfect suggestions. A silent phase can estimate case flow, shift prevalence, trigger rates, and potential workload; without responses it cannot validate query utility, adaptation, anchoring, or time savings. Early reporting should follow DECIDE-AI [76], recognizing that a reporting framework is not itself evidence of safety.

10 Falsifiable Hypotheses and Their Dependency Order

The research program is useful only if its central assumptions can fail, and only if their failure has consequences for one another. The hypotheses below are therefore ordered by dependency rather than by importance, and Figure 3 states the order explicitly. The practical implication is that the program is sequentially abandonable: a negative result at a lower node makes the higher nodes untestable rather than merely unsupported.

H1 (foundation): multi-level risk supports action selection. Combining local error localization, structure quality, and case atypicality improves accept/query/defer decisions over voxel entropy alone. It is falsified if selective risk does not improve at held-out institutions, or if patient-level calibration fails after thresholds are frozen.

H2 (foundation): repairability is estimable, and distinct from error. The reachability score $\hat{\rho}$ discriminates correctable from unreachable failure at external sites, beyond what predicted mask quality and atypicality already provide, and removing the gate degrades the policy. It is falsified by chance-level external discrimination, by subsumption under predicted mask quality, or by a null gate ablation.

H3: dynamic queries improve value per unit effort. *Conditional on H1 and H2.* At matched clinician time, a value-guided policy reaches an acceptable mask more often than random queries, fixed-click heuristics, or passive correction. It is falsified if the gain disappears under elapsed-time accounting or with real users. If H1 or H2 fails, H3 is not tested, because a value estimate built on an invalid risk or reachability signal is uninterpretable rather than merely weak.

H4: feedback should update task representation. *Conditional on H3.* At equal risk, a correction that updates support or prompt representation improves unedited slices in the same volume more than a local overwrite. It is falsified by negligible within-volume transfer, propagated error, or excessive protected-task loss.

H5: correction priors transfer where mask priors cannot. *Conditional on H4.* A cross-case experience layer keyed on failure modes reduces future episode cost on *unseen* rare presentations, whereas a case-similarity memory does not. It is falsified if transfer requires presentation-level similarity, if benefit vanishes under the cross-reader write gate, or if subgroup audits show entrenchment of institutional idiosyncrasy.

H6: bounded adaptation and governed personalization. *Conditional on H4 and H5.* Trust regions, anchors, frozen gates, and rollback retain most target-domain gain while reducing catastrophic accumulation and forgetting; with consent and sufficient observations, clinician-conditioned policies reduce time and unnecessary queries without increasing error or over-reliance. It is falsified if constraints only suppress useful adaptation, if benefit is limited to robot users, or if task allocation becomes inequitable or unexplainable.

11 Levels of Evidence Required for Translation

Level I: benchmark and decision evidence. A unified protocol should connect within-domain FSMIS, CD-FSMIS, promptable foundation models, interactive refinement, and selective prediction. Evidence at this level must establish calibration, reachability estimation, and accept/query/defer

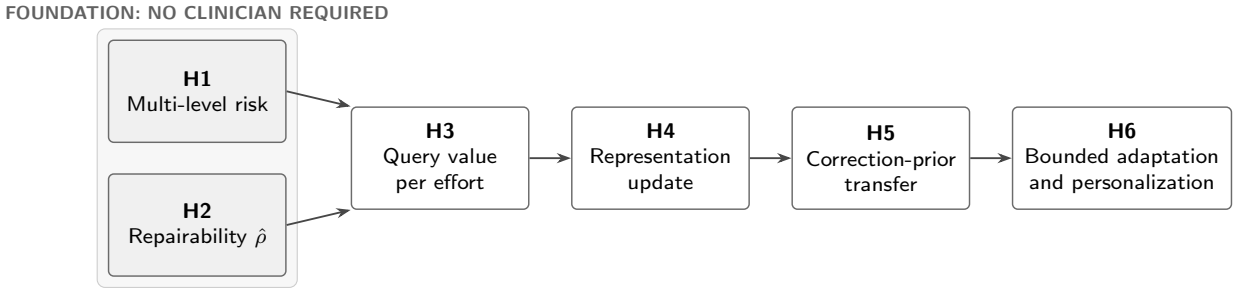

Figure 3: Dependency order of the hypotheses. H1 and H2 are testable with simulated users and existing data (Section 9.1); everything downstream presupposes them. A negative result at the foundation layer terminates the program rather than motivating a larger study.

baselines, define clinically weighted edge-case strata, and derive interaction-cost models from multiple robot policies and initial clinician traces. An auditable controller can be evaluated before any parameter adaptation is allowed. The pilot of Section 9.1 sits at the entry of this level.

Level II: interaction and adaptation evidence. Comparative studies should evaluate points, scribbles, boundaries, references, and text together with episode-local prompt, prototype, or adapter updates. Evidence must show whether feedback lowers risk for the intended reason, whether benefit transfers beyond the edited region, and whether rollback prevents non-monotonic failure. Clinical collaboration is essential because cost and acceptability cannot be inferred from benchmark masks.

Level III: prospective clinician-model evidence. Silent studies should first estimate shift, failure prevalence, trigger rates, and potential workload without claiming interaction efficiency. Limited workflow studies should then evaluate clinician-specific policies, inter-reader disagreement, and query utility. Any cross-case learning through the experience layer must be a separately consented and governed protocol. Thresholds should be frozen; model, clinician, and joint-system performance should be reported separately; and site and subgroup analyses should be prespecified.

12 Limitations, Governance, and Scope

The central technical risk is that uncertainty can be unreliable precisely under severe shift. Independent signals and conditional calibration can mitigate this, but detected shift can only block adaptation or trigger review; it does not prove error. The reachability estimate inherits this problem in a specific form: it is trained on simulated trajectories whose response model may not match clinical reality, so a systematic optimism in the simulator becomes a systematic overestimate of repairability. The pessimistic formulation of Section 5.4 is a partial remedy, not a solution, and reporting sensitivity to the simulator is mandatory rather than optional.

Feedback is another source of uncertainty. A correction can reflect ambiguity, fatigue, or idiosyncratic preference, which is why the write gate of Section 6.4 is conservative by design. That conservatism has a real cost in settings with few readers, and we do not claim to have resolved the resulting trade-off between safety and learning rate.

Human factors create a distinct failure mode. A plausible mask may anchor a reader, while repeated low-value queries create fatigue. Controlled studies must therefore measure override behavior, workload, and recovery from deliberately imperfect suggestions, not only accuracy or usability. Evidence that erroneous advice can degrade physician decisions [57, 58] makes this a primary safety endpoint rather than a secondary one.

Finally, segmentation quality is not identical to clinical benefit. A boundary deviation may be

harmless in one application and consequential in another. Each study must define the use case, tolerance, failure cost, and downstream endpoint before optimization. Retrospective benchmark gains, a single aggregate metric, or the presence of a large foundation model cannot establish general clinical safety. This framework remains normative and does not itself constitute evidence of prospective clinical effectiveness.

13 Conclusion

Conventional FSMIS asks how much can be learned from a few labeled images. Cross-domain FSMIS asks whether that evidence transfers across unfamiliar acquisition domains. The clinician-in-the-loop reformulation is often stated as a further step toward interaction and rapid adaptation, and we have argued that this statement is true but insufficiently demanding. Interaction interfaces exist, adaptation mechanisms exist, and neither is the binding constraint. What does not exist is a system that can say, for the case in front of it, whether its own error is the kind a bounded intervention can repair.

The framework proposed here therefore places decidable self-assessment beneath interaction and adaptation rather than beside them. It changes the unit of analysis from a final mask to the trajectory of a joint clinician-model system, gates that trajectory on an explicit reachability test, treats parsimony rather than speed as the virtue of interaction and reversibility rather than speed as the virtue of adaptation, and separates a per-case episode from a governed experience layer whose learning target is a correction prior over failure modes rather than a mask prior over rare diseases. Foundation models supply a reusable interface, not a safety argument. The desired endpoint is not a model that appears universally confident, but a system that knows when to act, when to ask, when asking cannot help, how to learn safely from an answer, and when to stop.

Declaration of competing interest

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No new data were created or analyzed for this Perspective.

A Non-myopic controller

The greedy rule (4) ignores that one answer can alter the value of later questions. A non-myopic controller instead satisfies the constrained Bellman relation

$$V(s, B) = \min\Big\{ J_A(s),\ J_D(s),\ \min_{a \in \mathcal{A}_Q :\, b(a) \preceq B} \big[\lambda_C c(a) + \mathbb{E}_z V(\mathcal{T}(s, a, z),\ B - b(a))\big] \Big\}, \tag{6}$$

with the reachability gate applied to the third branch as in (4). The relation is stated for completeness; in practice the state is high-dimensional and the expectation is taken under an estimated response model, so approximate solution by offline policy evaluation on recorded trajectories is the realistic route. The distinction between the greedy and non-myopic forms is empirically meaningful mainly when channels differ sharply in bandwidth, for example when a cheap click is worth issuing only because it determines whether an expensive reference-case query is warranted.

B Notation

Symbol	Meaning
K	Number of static support image-mask pairs fixed before inference.
B, B_t	Vector of hard interaction limits; remaining budget at step t.
$b(a), c(a)$	Hard resource use and residual soft burden of action a.
s_t, h_t, u_t	Episode state; interaction history; multi-level risk evidence.
$d_t \in \{A, Q, D\}$	Accept, query, defer.
$\mathcal{T}$	Bounded, reversible update operator.
$\mathcal{M}(s), \mathcal{M}^{\mathcal{C}}(s)$	Reachable set and correctable set at state s.
$\hat{\rho}(s)$	Estimated probability that the case is correctable at s.
$\mathcal{R}, \varepsilon_R$	Protected reference set and retention tolerance.
$\mathcal{P}_a$	Ambiguity set of response distributions used in cold start.
L_{clin}, C_D	Clinically weighted loss; cost of full expert review.

References

[1] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots. One-Shot Learning for Semantic Segmentation. In *BMVC*, 2017.

[2] J. Snell, K. Swersky, and R. S. Zemel. Prototypical Networks for Few-Shot Learning. In *NeurIPS*, vol. 30, 2017.

[3] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng. PANet: Few-Shot Image Semantic Segmentation with Prototype Alignment. In *ICCV*, pp. 9197–9206, 2019.

[4] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger. 'Squeeze & Excite' Guided Few-Shot Segmentation of Volumetric Images. *Medical Image Analysis*, 59:101587, 2020.

[5] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert. Self-Supervised Learning for Few-Shot Medical Image Segmentation. *IEEE TMI*, 41(7):1837–1848, 2022.

[6] S. Hansen, S. Gautam, R. Jenssen, and M. Kampffmeyer. Anomaly Detection-Inspired Few-Shot Medical Image Segmentation Through Self-Supervision with Supervoxels. *Medical Image Analysis*, 78:102385, 2022.

[7] Y. Zhu, S. Wang, T. Xin, and H. Zhang. Few-Shot Medical Image Segmentation via a Region-Enhanced Prototypical Transformer. In *MICCAI*, LNCS 14223, pp. 271–280, 2023.

[8] Y. Zhu, S. Wang, T. Xin, Z. Zhang, and H. Zhang. Partition-A-Medical-Image: Extracting Multiple Representative Subregions for Few-Shot Medical Image Segmentation. *IEEE TIM*, 73:1–12, 2024.

[9] Y. Zhu, Z. Cheng, S. Wang, and H. Zhang. Learning De-Biased Prototypes for Few-Shot Medical Image Segmentation. *Pattern Recognition Letters*, 183:71–77, 2024.

[10] S. Lei, X. Zhang, J. He, F. Chen, B. Du, and C.-T. Lu. Cross-Domain Few-Shot Semantic Segmentation. In *ECCV*, pp. 73–90, 2022.

[11] Y. Zhu, M. Li, Q. Ye, S. Wang, T. Xin, and H. Zhang. RobustEMD: Domain Robust Matching for Cross-Domain Few-Shot Medical Image Segmentation. *Artificial Intelligence in Medicine*, 167:103197, 2025.

[12] Y. Zhu, S. Wang, T. Zhou, Z. Li, H. Zhang, and L. Shao. Cross-Domain Few-Shot Medical Image Segmentation via Dynamic Semantic Matching. *IEEE TIP*, 34:6669–6682, 2025.

[13] Y. Zhu, S. Wang, T. Xin, and H. Zhang. Adversarial Prototypical Perturbation for Cross-Domain Few-Shot Medical Image Segmentation. *ACM TOMM*, 22(8):1–27, 2026.

[14] Y. Zhu and H. Zhang. MAUP: Training-Free Multi-Center Adaptive Uncertainty-Aware Prompting for Cross-Domain Few-Shot Medical Image Segmentation. In *MICCAI*, LNCS 15966, pp. 326–336, 2025.

[15] G. Wang, W. Li, M. A. Zuluaga, et al. Interactive Medical Image Segmentation Using Deep Learning with Image-Specific Fine Tuning. *IEEE TMI*, 37(7):1562–1573, 2018.

[16] G. Wang, M. A. Zuluaga, W. Li, et al. DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation. *IEEE TPAMI*, 41(7):1559–1572, 2019.

[17] X. Luo, G. Wang, T. Song, et al. MIDeepSeg: Minimally Interactive Segmentation of Unseen Objects from Medical Images Using Deep Learning. *Medical Image Analysis*, 72:102102, 2021.

[18] N. A. Koohbanani, M. Jahanifar, N. Z. Tajadin, and N. Rajpoot. NuClick: A Deep Learning Framework for Interactive Segmentation of Microscopic Images. *Medical Image Analysis*, 65:101771, 2020.

[19] H. E. Wong, M. Rakic, J. Guttag, and A. V. Dalca. ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Biomedical Image. In *ECCV*, pp. 207–229, 2024.

[20] Q. Liu, M. Zheng, B. Planche, et al. PseudoClick: Interactive Image Segmentation with Click Imitation. In *ECCV*, pp. 728–745, 2022.

[21] C. Shen, W. Li, Q. Xu, et al. Interactive Medical Image Segmentation with Self-Adaptive Confidence Calibration. *Frontiers of Information Technology & Electronic Engineering*, 24(9):1332–1348, 2023.

[22] I. A. Mikhailov, B. Chauveau, N. Bourdel, et al. A Deep Learning-Based Interactive Medical Image Segmentation Framework with Sequential Memory. *Computer Methods and Programs in Biomedicine*, 245:108038, 2024.

[23] R. Feng, X. Zheng, T. Gao, J. Chen, W. Wang, D. Z. Chen, and J. Wu. Interactive Few-Shot Learning: Limited Supervision, Better Medical Image Segmentation. *IEEE TMI*, 40(10):2575–2588, 2021.

[24] D. Al Chanti, V. G. Duque, M. Crouzier, et al. IFSS-Net: Interactive Few-Shot Siamese Network for Faster Muscle Segmentation and Propagation in Volumetric Ultrasound. *IEEE TMI*, 40(10):2615–2628, 2021.

[25] T. Kontogianni, E. Celikkan, S. Tang, and K. Schindler. Continuous Adaptation for Interactive Object Segmentation by Learning from Corrections. In *ECCV*, pp. 579–596, 2020.

[26] S. Hu, Z. Liao, Z. Liu, and Y. Xia. Towards Clinician-Preferred Segmentation: Leveraging Human-in-the-Loop for Test-Time Adaptation in Medical Image Segmentation. arXiv:2405.08270, 2024.

[27] J. Zhu and J. Wu. MedUHIP: Towards Human-in-the-Loop Medical Segmentation. arXiv:2408.01620, 2024.

[28] G. Torpey, J. Guttag, and H. E. Wong. Learning What to Ask For When: Image Ordering for In-Context Interactive Medical Image Segmentation. In *CVPR Workshops*, pp. 6418–6426, 2026.

[29] Y. Gehad, E. Zerefa, K. Kabra, and G. Balakrishnan. Context Matters: Support Set Selection and Failure Detection for In-Context Medical Image Segmentation. arXiv:2608.05333, 2026.

[30] Q. Tian, H. Sun, Y. Wang, Y. Shi, and Y. Yin. DeferredSeg: A Multi-Expert Deferral Framework for Medical Image Segmentation. *Pattern Recognition*, 182:114811, 2026.

[31] W. Xu, Z. Liang, H. Anthony, Y. Ibrahim, F. Cohen, G. Yang, and K. Kamnitsas. You Point, I Learn: Online Adaptation of Interactive Segmentation Models for Handling Distribution Shifts in Medical Imaging. In *ICLR*, 2026.

[32] G. Zhao, Y. Wang, C. Gong, et al. User-Preference Alignment with Uncertainty-Aware Interactive Rectification for Liver Organ and Tumor Segmentation and Analysis from CT Images. *npj Digital Medicine*, 9:418, 2026.

[33] D. J. C. MacKay. Information-Based Objective Functions for Active Data Selection. *Neural Computation*, 4(4):590–604, 1992.

[34] Y. Gal, R. Islam, and Z. Ghahramani. Deep Bayesian Active Learning with Image Data. In *ICML*, PMLR 70, pp. 1183–1192, 2017.

[35] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. In *MICCAI*, pp. 399–407, 2017.

[36] X. Li, M. Xia, J. Jiao, S. Zhou, C. Chang, Y. Wang, and Y. Guo. HAL-IA: A Hybrid Active Learning Framework Using Interactive Annotation for Medical Image Segmentation. *Medical Image Analysis*, 88:102862, 2023.

[37] V. I. Butoi, J. J. Gonzalez Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca. UniverSeg: Universal Medical Image Segmentation. In *ICCV*, pp. 21438–21451, 2023.

[38] H. E. Wong, J. J. Gonzalez Ortiz, J. V. Guttag, and A. V. Dalca. MultiverSeg: Scalable Interactive Segmentation of Biomedical Imaging Datasets with In-Context Guidance. In *ICCV*, pp. 20966–20980, 2025.

[39] R. A. Howard. Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

[40] Y. Geifman and R. El-Yaniv. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *ICML*, PMLR 97, pp. 2151–2159, 2019.

[41] Y. Ding, J. Liu, X. Xu, M. Huang, J. Zhuang, J. Xiong, and Y. Shi. Uncertainty-Aware Training of Neural Networks for Selective Medical Image Segmentation. In *MIDL*, PMLR 121, pp. 156–173, 2020.

[42] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. *IEEE TMI*, 23(7):903–921, 2004.

[43] S. A. A. Kohl, B. Romera-Paredes, C. Meyer, et al. A Probabilistic U-Net for Segmentation of Ambiguous Images. In *NeurIPS*, vol. 31, pp. 6965–6975, 2018.

[44] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. In *ICML*, PMLR 70, pp. 1321–1330, 2017.

[45] Y. Ovadia, E. Fertig, J. Ren, et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *NeurIPS*, vol. 32, 2019.

[46] A. Mehrtash, W. M. Wells III, C. M. Tempany, P. Abolmaesumi, and T. Kapur. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *IEEE TMI*, 39(12):3868–3878, 2020.

[47] A. Jungo, F. Balsiger, and M. Reyes. Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation. *Frontiers in Neuroscience*, 14:282, 2020.

[48] M. Zenk, D. Zimmerer, F. Isensee, et al. Comparative Benchmarking of Failure Detection Methods in Medical Image Segmentation: Unveiling the Role of Confidence Aggregation. *Medical Image Analysis*, 101:103392, 2025.

[49] Y. Du, F. Bai, T. Huang, and B. Zhao. SegVol: Universal and Interactive Volumetric Medical Image Segmentation. In *NeurIPS*, vol. 37, 2024.

[50] A. Kirillov, E. Mintun, N. Ravi, et al. Segment Anything. In *ICCV*, 2023.

[51] S. D. Jain and K. Grauman. Predicting Sufficient Annotation Strength for Interactive Foreground Segmentation. In *ICCV*, pp. 1313–1320, 2013.

[52] N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu. Test-Time Adaptable Neural Networks for Robust Medical Image Segmentation. *Medical Image Analysis*, 68:101907, 2021.

[53] M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. Ben Ayed. Source-Free Domain Adaptation for Image Segmentation. *Medical Image Analysis*, 82:102617, 2022.

[54] Q. Wang, O. Fink, L. Van Gool, and D. Dai. Continual Test-Time Domain Adaptation. In *CVPR*, 2022.

[55] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan. Efficient Test-Time Model Adaptation without Forgetting. In *ICML*, PMLR 162, pp. 16888–16905, 2022.

[56] P. Tschandl, C. Rinner, Z. Apalla, et al. Human–Computer Collaboration for Skin Cancer Recognition. *Nature Medicine*, 26:1229–1234, 2020.

[57] S. Gaube, H. Suresh, M. Raue, et al. Do as AI Say: Susceptibility in Deployment of Clinical Decision-Aids. *npj Digital Medicine*, 4:31, 2021.

[58] F. Yu, A. Moehring, O. Banerjee, et al. Heterogeneity and Predictors of the Effects of AI Assistance on Radiologists. *Nature Medicine*, 30:837–849, 2024.

[59] S.-L. Lu, F.-R. Xiao, J. C.-H. Cheng, et al. Randomized Multi-Reader Evaluation of Automated Detection and Segmentation of Brain Tumors in Stereotactic Radiosurgery with Deep Neural Networks. *Neuro-Oncology*, 23(9):1560–1568, 2021.

[60] E. P. P. Pang, H. Q. Tan, F. Wang, et al. Multicentre Evaluation of Deep Learning CT Autosegmentation of the Head and Neck Region for Radiotherapy. *npj Digital Medicine*, 8:312, 2025.

[61] N. Arjmandi, A. R. Sebzari, F. Molaei, et al. Clinical Validation of AI-Assisted Contouring in Prostate Radiation Therapy Treatment Planning: Highlighting Automation Bias and the Need for Standardized Quality Assurance. *Journal of Applied Clinical Medical Physics*, 27(1):e70425, 2026.

[62] H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, and D. Sontag. Who Should Predict? Exact Algorithms for Learning to Defer to Humans. In *AISTATS*, PMLR 206, pp. 10520–10545, 2023.

[63] D. Tailor, A. Patra, R. Verma, P. Manggala, and E. Nalisnick. Learning to Defer to a Population: A Meta-Learning Approach. In *AISTATS*, PMLR 238, pp. 3475–3483, 2024.

[64] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang. Segment Anything Model for Medical Image Analysis: An Experimental Study. *Medical Image Analysis*, 89:102918, 2023.

[65] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment Anything in Medical Images. *Nature Communications*, 15:654, 2024.

[66] J. Wu, Z. Wang, M. Hong, W. Ji, H. Fu, Y. Xu, M. Xu, and Y. Jin. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *Medical Image Analysis*, 102:103547, 2025.

[67] Z. Xu, C. Chen, D. Lu, et al. FM-ABS: Promptable Foundation Model Drives Active Barely Supervised Learning for 3D Medical Image Segmentation. In *MICCAI*, pp. 294–304, 2024.

[68] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study. *PLOS Medicine*, 15(11):e1002683, 2018.

[69] C. Ouyang, C. Chen, S. Li, Z. Li, C. Qin, W. Bai, and D. Rueckert. Causality-Inspired Single-Source Domain Generalization for Medical Image Segmentation. *IEEE TMI*, 42(4):1095–1106, 2023.

[70] L. Maier-Hein, A. Reinke, P. Godau, et al. Metrics Reloaded: Recommendations for Image Analysis Validation. *Nature Methods*, 21:195–212, 2024.

[71] S. Nikolov, S. Blackwell, A. Zverovitch, et al. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. *Journal of Medical Internet Research*, 23(7):e26151, 2021.

[72] A. J. Vickers and E. B. Elkin. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, 26(6):565–574, 2006.

[73] P. Kohli, H. Nickisch, C. Rother, and C. Rhemann. User-Centric Learning and Evaluation of Interactive Segmentation Systems. *International Journal of Computer Vision*, 100(3):261–274, 2012.

[74] M. Amrehn, S. Steidl, R. Kortekaas, et al. A Semi-Automated Usability Evaluation Framework for Interactive Image Segmentation Systems. *International Journal of Biomedical Imaging*, 2019:1464592, 2019.

[75] A. Ramkumar, J. Dolz, H. A. Kirisli, et al. User Interaction in Semi-Automatic Segmentation of Organs at Risk: A Case Study in Radiotherapy. *Journal of Digital Imaging*, 29(2):264–277, 2016.

[76] B. Vasey, M. Nagendran, B. Campbell, et al. Reporting Guideline for the Early-Stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI. *Nature Medicine*, 28:924–933, 2022.

[77] B. Settles. Active Learning Literature Survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

[78] D. Madras, T. Pitassi, and R. Zemel. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *NeurIPS*, vol. 31, 2018.

[79] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan. Distribution-Free, Risk-Controlling Prediction Sets. *Journal of the ACM*, 68(6):1–34, 2021.

[80] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, et al. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. In *MICCAI*, pp. 119–127, 2019.

[81] A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *NeurIPS*, vol. 30, 2017.

[82] D. Hendrycks and K. Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*, 2017.

[83] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker. Domain Generalization via Model-Agnostic Learning of Semantic Features. In *NeurIPS*, vol. 32, 2019.

[84] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain Generalization: A Survey. *IEEE TPAMI*, 45(4):4396–4415, 2023.

[85] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nature Methods*, 18:203–211, 2021.