

— SCREENING MAMMOGRAPHY

Reconciling Generative and Discriminative Models in Screening Mammography

A rule-constrained orchestration layer for **grounded, auditable draft findings**

AUTHOR

HOPPR

PUBLISHED

August 2026

PIPELINE

Mammo VLM v3

ABSTRACT

A screening mammography pipeline that drafts the findings section of a radiology report can be built from a vision-language model (VLM) fine-tuned on radiologist report text, or from a segmentation-based detector that grounds findings in image coordinates, but neither alone produces an acceptable report. The VLM writes clinically idiomatic prose, chooses the specific finding noun, and distinguishes benign from suspicious patterns, yet its spatial claims are generated rather than detected. The localizer triangulates quadrant, depth, distance from

the nipple, clock position, and bounding boxes from paired craniocaudal and mediolateral-oblique views, yet it has no language model and, as deployed, characterizes every finding it emits as recall-worthy, so it can never assert that a finding is benign. This paper describes the orchestration layer that reconciles the two: a locally hosted instruction-tuned language model, wrapped in deterministic code, that merges VLM prose with localizer geometry under an explicit policy. We state the four invariants the policy enforces, of which the most consequential are that VLM findings content is never deleted and that the BI-RADS assessment may only be made more actionable, never less; we enumerate the twelve action codes into which every reconciliation decision is classified; and we describe the mechanism by which BI-RADS arithmetic is recomputed in code and permitted to override the language model's own output, so that a four-billion-parameter local model cannot move the clinical assessment outside its allowed set. We then identify what the design does not yet resolve. Promotion to BI-RADS Category 0 on the strength of an uncorroborated localizer firing is currently unconditional, with no confidence gate, and it is this injection rate that determines the specificity cost of the pipeline at natural screening prevalence. A schema mismatch between the declared input contract and the localizer's actual output places every injected distance either at risk of validation rejection or wrong by a factor of ten. A separate policy exploration proposes inverting the central invariant, treating localization as a corroboration gate that suppresses uncorroborated VLM findings rather than as an enrichment source, and that question remains open. Because the value of orchestration is distributed unevenly across the action codes, we argue that it cannot be evaluated as a single aggregate number, and specify a per-action-code evaluation on a natural-prevalence cohort instead. Results against the pipeline's stated performance targets are not yet documented and are not claimed here.

4

Invariants the
policy enforces

12

Action codes for
every
reconciliation
decision

~4B

Parameters in the
local language
model

0

The only BI-RADS
category
promotion may
reach

Introduction

A model that drafts the findings section of a screening mammography report has to do two things that are usually the province of different model families. It has to write clinical prose that a radiologist would recognize as report language, choosing among finding nouns whose distinctions carry diagnostic weight (a mass is not an asymmetry, and neither is architectural distortion), and distinguishing calcification patterns that are reassuring from patterns that are not. It also has to say where the finding is, in terms specific enough to be actionable and grounded firmly enough that a reader can verify the claim against the image. A generative vision-language model fine-tuned on report text does the first well and the second only apparently: its spatial language is produced by the same next-token mechanism that produces everything else, and nothing in the architecture ties the phrase “upper outer quadrant” to any particular region of the input. A discriminative detector does the second well and cannot do the first at all.

The Mammo VLM v3 pipeline therefore runs both. A Qwen-family vision-language model, fine-tuned on the findings sections of mammography reports, generates a findings paragraph per breast from the craniocaudal (CC) and mediolateral-oblique (MLO) views, including breast density and a BI-RADS assessment constrained to categories 0, 1, and 2. Independently, and deliberately blind to the VLM’s output, a localizer stack produces spatial geometry: a UNet++ segmentation model on an EfficientNetV2-S encoder yields per-pixel probability maps over four lesion classes, connected-component clustering converts those maps into candidate regions, separate landmark models locate the nipple on each view and the pectoral muscle on MLO, and an emission layer performs Hungarian cross-view matching and triangulation to produce per-breast findings carrying quadrant, depth, distance from the nipple, clock position, and bounding boxes in full-resolution image coordinates.

Running both models raises the problem this paper addresses. The two branches disagree, frequently and by construction, and every disagreement requires a decision with a clinical consequence rather than a technical one. When the VLM describes a mass the localizer did not detect, suppressing the sentence discards a radiologist-trained reading on the evidence of a thresholded detector whose misses are expected behavior; keeping it publishes a spatial claim that no detector supports. When the localizer fires on a breast the VLM read as negative, injecting a finding the VLM never saw forces a recall on the strength of a single model, while declining to inject discards precisely the grounded signal the localizer was

added to provide. The sharpest case is the third: when the VLM reads calcifications as benign and the localizer fires on the calcification axis, the localizer's output cannot distinguish between firing on the same benign calcifications the radiologist described and firing on a different, genuinely suspicious cluster, because the localizer as deployed emits a recall-worthy characterization for every finding it produces and never emits a benign one. No amount of additional signal from the localizer resolves that ambiguity; it has to be resolved by a stated policy whose cost is then measured.

DISAGREEMENT 1

VLM describes a mass the localizer did not detect

Suppress it, or publish a spatial claim no detector supports.

DISAGREEMENT 2

Localizer fires on a breast the VLM read as negative

Inject and force a recall, or discard the grounded signal.

THE SHARPEST CASE

VLM reads benign calcifications; localizer fires

The same benign calcifications, or a different, genuinely suspicious cluster?

The orchestrator is the subsystem in which those policies live. It is an instruction-tuned language model, hosted locally in the inference container rather than called as a hosted service, that receives the VLM's findings text and the localizer's structured output for one breast and returns a reconciled findings paragraph together with structured metadata describing what it did. Its contribution is not primarily accuracy, and this paper does not claim an accuracy result. Its contribution is that the pipeline's output becomes attributable: for any sentence in a generated report, an action code recorded in the audit trail identifies which upstream model contributed the sentence, under which rule, and whether the two models agreed.

That attributability is what makes the remainder of the design tractable, and it is also what makes the design's open questions legible rather than latent. We are able to state, precisely, that promotion to BI-RADS Category 0 on an uncorroborated localizer firing is currently unconditional; that one action code deliberately violates the pipeline's own preservation invariant and does so because it is the only path by which a benign mislabel can be recovered; and that a competing fusion policy exists which inverts the central invariant and has not been decided against. A pipeline without a decision taxonomy would have all of the same behavior and none of the same visibility into it.

The remainder of this paper is organized as follows. Section 2 describes the two upstream models, their asymmetric competences, and the deployment properties of the orchestration layer that constrain its design. Section 3 states the four invariants on which the policy rests.

Section 4 enumerates the reconciliation policy itself: the twelve action codes, the one deliberate exception to the preservation invariant, the role of the localizer’s finding subtype in noun selection, and the construction of the location phrase including a spatial-consistency rule that exists to prevent a specific class of clinically wrong sentence. Section 5 describes how the language model is constrained, in particular how BI-RADS arithmetic is removed from its authority and recomputed in code, and what portion of the decision logic remains inside the model against the intent of the design. Section 6 sets out the known risks and the open clinical and policy questions, including the alternative corroboration-gate fusion policy. Section 7 specifies the evaluation framework. Section 8 states scope, limitations, and regulatory positioning. Section 9 concludes.

02

System context

2.1 Asymmetric competence

The design of the reconciliation policy follows almost entirely from the fact that neither upstream model dominates the other. Their competences are close to disjoint, as summarized in Table 1, and each is authoritative in a region where the other has nothing to contribute.

Capability	VLM	Localizer
Clinical finding noun (“oval mass,” “spiculated nodule,” “focal asymmetry”)	Yes, specific	Coarse: pools mass, asymmetry, and architectural distortion into one slot with a three-value subtype
Benign versus suspicious characterization	Yes	No: emitted recall_worthy for 100% of present findings (394 of 394 verified)
Grounded spatial geometry	No: spatial language is generated, not detected	Yes: triangulated from CC and MLO with landmark models
Bounding boxes for image overlay	No	Yes, in full-resolution pixel coordinates

Capability	VLM	Localizer
Breast density assessment	Yes	No
Findings outside the two axes (implants, lymph nodes, post-surgical change)	Yes	No: outside its vocabulary

Table 1. Competences of the two upstream models. The policy in Section 4 is largely a consequence of this table.

Two entries in that table do most of the work. The first is that the localizer cannot say “benign.” This is not a tuning artifact to be corrected later but a property of what the localizer was built to do: it emits regions worth a second look, and the absence of a benign category means that a localizer firing on a breast the VLM read as benign carries no information about which of the two competing explanations is correct. Every subtlety in the calcification axis of Section 4 descends from this single fact, and the current prompt states it explicitly, as a defensive measure, precisely because a future maintainer who added a rule trusting the localizer’s characterization field would silently break the benign and suspicious logic without any test failing.

394/394

Verified present findings the localizer emitted as `recall_worthy`

KEY FACT

The localizer cannot say “benign.” It emits regions worth a second look.

The second is that the VLM’s spatial language is generated. This is what makes enrichment worth doing at all: when both models fire, the VLM contributes the noun, the size, and the hedging, and the localizer contributes the coordinates, and the resulting sentence is better than either model produces alone. It is also why the VLM’s noun is preserved rather than overwritten whenever the two disagree about what kind of finding is present. The radiologist-trained model’s word choice is the more informative signal; the localizer’s three-value subtype is a pooled bucket.

2.2 Pipeline and deployment properties

The orchestrator is the third branch of a three-branch pipeline. A study of standard screening views is first reduced by view selection, which requires both a CC and an MLO view per breast, applies a priority ranking that prefers screening over diagnostic acquisitions

and implant-displaced views over standard ones, and drops any laterality for which both required views are unavailable. A YOLOX-Nano detector then locates the breast region, and the resulting crop is shared by both downstream branches so that the same breast region is analyzed by each. The VLM branch resizes and normalizes the crop and generates findings text per laterality under beam search with a bounded token budget. The localizer branch resizes to its own working resolution, runs segmentation with temperature-scaled outputs, clusters connected components into candidate regions subject to a per-image cap, detects landmarks, and emits per-breast findings through a matching cascade whose primary path is Hungarian assignment between CC and MLO candidates, with documented fallbacks for single-view detections and for the case in which the segmentation model assigns the same lesion to different classes on the two views. The orchestrator consumes the outputs of both branches and produces the response the platform returns.

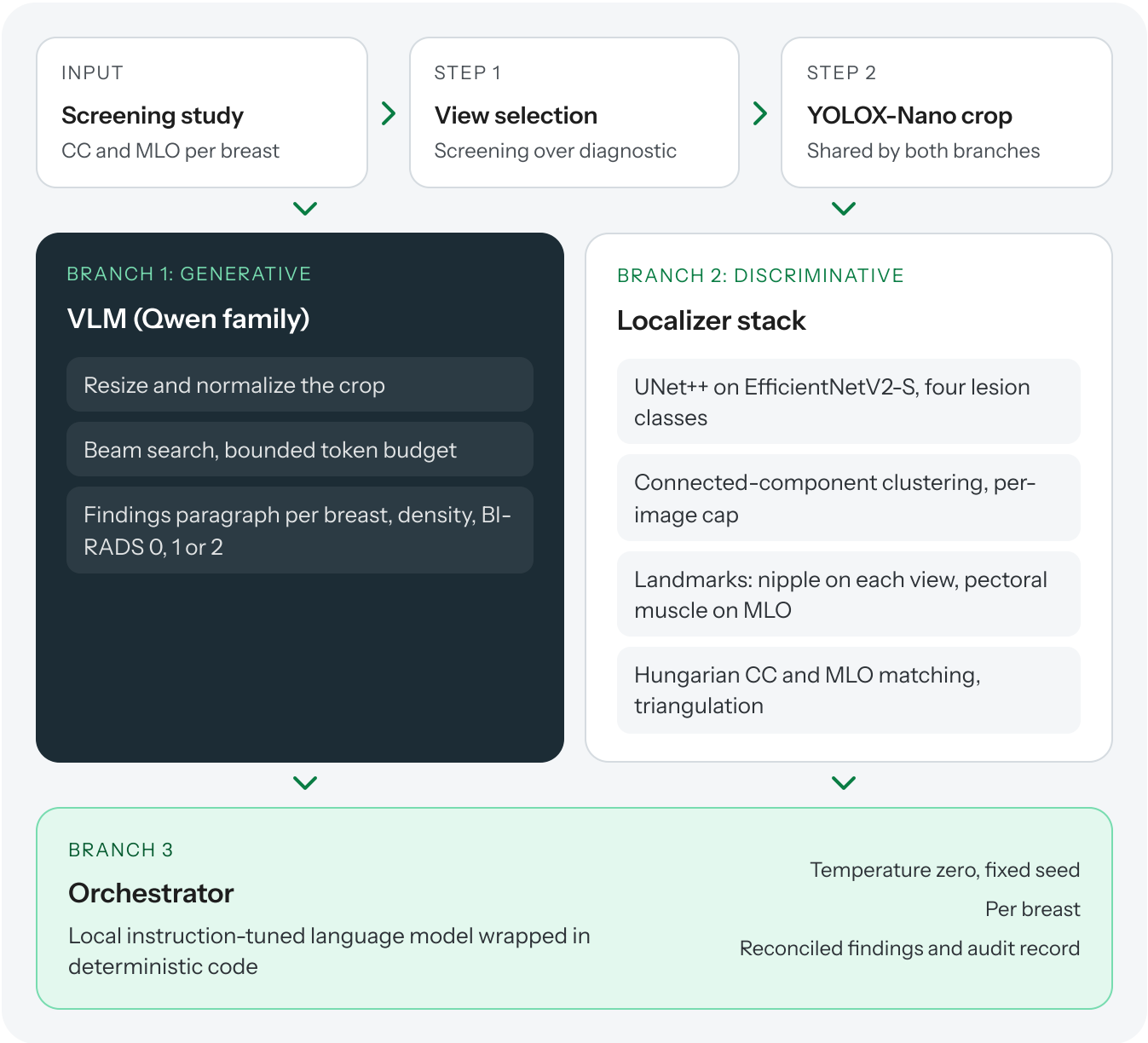


Figure 1. The three-branch pipeline. The same breast crop is analyzed by each branch, and the localizer is deliberately blind to the VLM’s output.

Three deployment properties of the orchestration layer are load-bearing for its design rather than incidental to it.

The orchestrator’s language model runs locally in the inference container, in reduced precision, rather than being called as a hosted service. The immediate motivation is that no study-derived text crosses the network to an external provider. The design consequence is more interesting: the local model is small, at roughly four billion parameters, which directly bounds how much reasoning can safely be delegated to it. A larger hosted model would tolerate more rule-following burden; the choice to run locally is therefore also an argument for moving decision logic into deterministic code, and Section 5 describes how far that has been taken and how far it has not.

The orchestrator is deterministic, at temperature zero with a fixed seed. This is a precondition for evaluation rather than a preference. The quantity of interest in any evaluation of this pipeline is the difference between the VLM alone and the VLM combined with the localizer; if the reconciliation step contributed its own per-row sampling variance, that variance would sit directly on top of the signal being measured and would swamp it.

The orchestrator operates per breast. Each study-and-laterality pair is reconciled independently, and there is no cross-laterality reasoning in the current design. A consequence worth naming is that a VLM sentence referring to the contralateral breast, which the VLM occasionally produces, is treated as text noise: it is passed through verbatim and never matched against the localizer, since matching it would attach geometry from one breast to a sentence about the other.

LOCAL

No study-derived text crosses the network; roughly four billion parameters.

DETERMINISTIC

Temperature zero with a fixed seed, a precondition for evaluation.

PER BREAST

Each study-and-laterality pair is reconciled independently.

Four invariants

The policy enumerated in Section 4 has twelve action codes, but it rests on four invariants, and the codes are best understood as the enumeration of what those invariants require in each combination of upstream states.

1

Augment only: VLM findings content is never deleted.

The VLM is trained on radiologist-written findings; the localizer is a detector operating at a fixed threshold. A detector that does not fire has not established that nothing is present, and the localizer's thresholds are set for an operating point rather than for exhaustive recall. The orchestrator may therefore enrich a VLM sentence with geometry, or add a sentence the VLM did not write, but it may not remove one. This invariant has exactly one deliberate exception, discussed in Section 4.3, and the reasoning behind that exception is the clearest illustration of why the invariant exists.

2

The BI-RADS assessment never becomes less urgent.

The only permitted change is promotion to Category 0, "Incomplete: Needs additional imaging evaluation." An earlier revision of the prompt justified this to the model by describing Category 0 as "the only category higher than 1 or 2," which invites a model to reason numerically and conclude that 2 exceeds 0. The current wording frames the constraint in terms of actionability instead: Category 0 is more actionable than 1 or 2, never less. The substance is unchanged and the failure mode it guards against is a model that follows the letter of a numerically stated rule.

3

The two finding axes are independent.

Mass-like findings and calcifications are reconciled separately. A study may carry findings on both axes with no interaction between them, and either axis may independently require promotion to Category 0. The final assessment is the disjunction: if either axis promotes, the result is Category 0, and promotion happens once regardless of how many axes demanded it.

4

No geometry is invented.

The orchestrator may use only the quadrant, depth, distance, and clock values present in the localizer's output for that breast. When a field is null or unspecified, the corresponding phrase is removed in full, including its units and its connecting words, so that the sentence reads grammatically with whatever remains. The output never contains a literal null placeholder and never a fragment such as “approximately cm from the nipple.” In the degenerate case where the localizer fires but no geometry field is populated, an injected sentence reduces to naming the breast.

04

The reconciliation policy

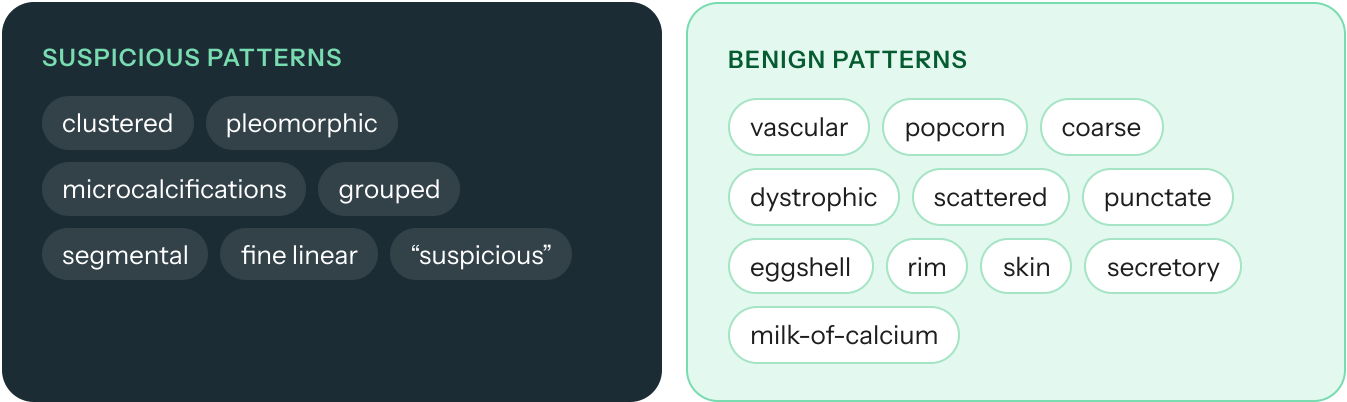
4.1 Action codes

For each axis independently, the orchestrator classifies the pair consisting of the VLM's state on that axis and the localizer's status into exactly one action code, and applies the transformation that code specifies. Tables 2 and 3 give the two taxonomies.

Code	Trigger	Effect	BI-RADS
M1_enriched	VLM non-negated mass-like noun, localizer present	Append geometry as a single subordinate clause; keep the VLM's noun unchanged	unchanged
M2_keep_vlm_no_loc_support	VLM positive, localizer absent	Keep VLM text verbatim; warn	unchanged
M3_inject_loc_only	VLM silent or negating, localizer present	Inject one sentence whose noun is selected by subtype	promote 1 or 2 to 0
M4_no_finding	Both silent	No change	unchanged
M5_multi_vlm_one_loc	Several positive VLM sentences, localizer fires once	Enrich the subtype-matching sentence; keep the others verbatim	unchanged

Table 2. Mass-like axis action codes.

Classification of the VLM’s calcification state is lexical, keyed on the descriptors the VLM actually produces. Suspicious patterns include clustered, pleomorphic, microcalcifications, grouped, segmental, and fine linear, together with a generic use of “suspicious.” Benign patterns include vascular, popcorn, coarse, dystrophic, scattered, punctate, eggshell, rim, skin, secretory, and milk-of-calcium. Everything outside the two axes (implants, lymph nodes, skin thickening, scar and biopsy markers, post-surgical change, fat necrosis, pacemakers) passes through verbatim and is never matched against the localizer, as does the density sentence.



Code	Trigger	Effect	BI-RADS
C1_enriched_suspicious	VLM suspicious-pattern calcification, localizer present	Enrich with geometry; keep the VLM’s descriptor	unchanged
C2_keep_vlm_no_loc_support	VLM suspicious, localizer absent	Keep verbatim; warn	unchanged
C3_inject_calc_loc_only	VLM silent or negating, localizer present	Inject “Suspicious calcifications are identified in . . .”	promote 1 or 2 to 0
C4_no_finding	Both silent	No change	unchanged
C5_benign_calc_passthrough	VLM benign-pattern calcification, localizer absent	Keep verbatim	unchanged
C6_benign_plus_loc	VLM benign-pattern calcification, localizer present	Drop the benign sentence and replace it with the injected suspicious sentence; warn	promote 1 or 2 to 0

Code	Trigger	Effect	BI-RADS
C7_split_susp_benign_with_loc	VLM has both benign and suspicious sentences, localizer present	Enrich the suspicious sentence; pass the benign sentence through	unchanged

Table 3. Calcification axis action codes.

4.2 Enrichment as a single clause

The enrichment codes attach geometry to an existing VLM sentence, and the manner of attachment is specified rather than left to the model: the geometry is appended as a single subordinate clause, not split into a second sentence. The reason is that a second sentence describing the same lesion reads as a second finding. A report stating that there is an oval mass in the left breast and then separately stating that a finding is present in the upper outer quadrant nine centimeters from the nipple has, to a reader, described two things. The single-clause requirement is a guard against inflating the apparent finding count through purely typographic means, and it is paired with an explicit instruction against producing a run-on sentence in the attempt.

4.3 One deliberate exception: C6

C6 is the only code that removes VLM text, and it therefore deserves to be justified rather than merely stated. Its trigger is a VLM sentence describing benign-pattern calcifications together with a localizer firing on the calcification axis, and its effect is to drop the benign sentence and replace it with the standard suspicious-calcification injection, promoting the assessment to Category O.

The reason it cannot simply be deleted from the policy is the localizer’s inability to characterize a finding as benign, established in Section 2.1. Because the localizer never says benign, C6 is the only path in the entire policy through which a VLM mislabel of a genuinely suspicious cluster as benign can be recovered. Removing C6 would leave that failure mode with no mitigation anywhere in the pipeline.

The reason it is nonetheless the policy’s largest single risk is that benign calcifications are ubiquitous in screening mammography, so the conjunction that triggers C6 is common, and a localizer firing on the same breast as a set of benign calcifications is frequently a different and coincidental cluster rather than a reinterpretation of the same one. C6 therefore converts a common, usually benign co-occurrence into a recall.

The rule’s behavior has already changed once, and the change is instructive. The original specification kept the VLM’s benign sentence and appended a separate hedged sentence describing the localizer’s region, on the reasoning that the two hypotheses were indistinguishable and that keeping both statements was the conservative choice. The current revision drops the benign sentence and replaces it, on the reasoning that a report asserting both benign-appearing coarse calcifications and suspicious calcifications in the same breast is internally incoherent, and that a reader cannot act on a document that contradicts itself. The trigger was narrowed at the same time, from benign calcification or benign mass language to benign calcification only, since the calcification axis is the only one where the argument applies.

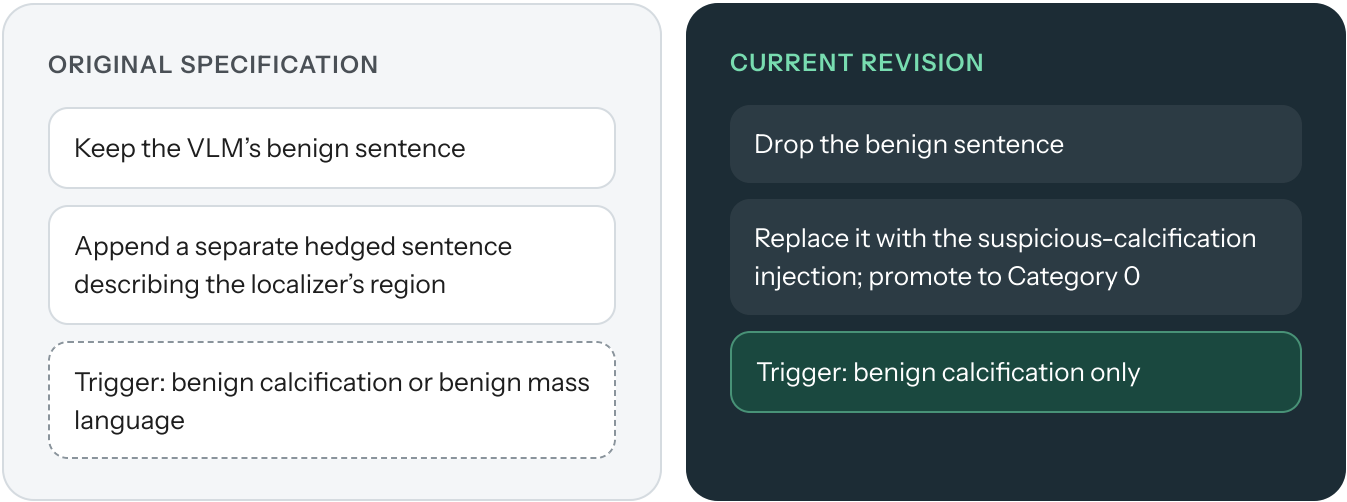


Figure 2. The two versions of `C6`. Neither version has been validated.

Neither version has been validated. What `C6` needs is a measurement of its recall-rate contribution in isolation, which Section 7 specifies, and an answer to a question that measurement alone will not settle: whether automatic promotion is the right response when the VLM has explicitly read the calcifications as benign, or whether localizer-versus-benign disagreement should instead flag the study for human re-read without triggering a recall on its own.

4.4 Subtype-driven noun selection

The localizer’s mass-like slot carries a subtype field taking the values mass, asymmetry, architectural distortion, or null. On a 759-row evaluation file the distribution was asymmetry 227, mass 107, and architectural distortion 55. Calcification entries carry no subtype, the only available upstream value being redundant with the slot name.

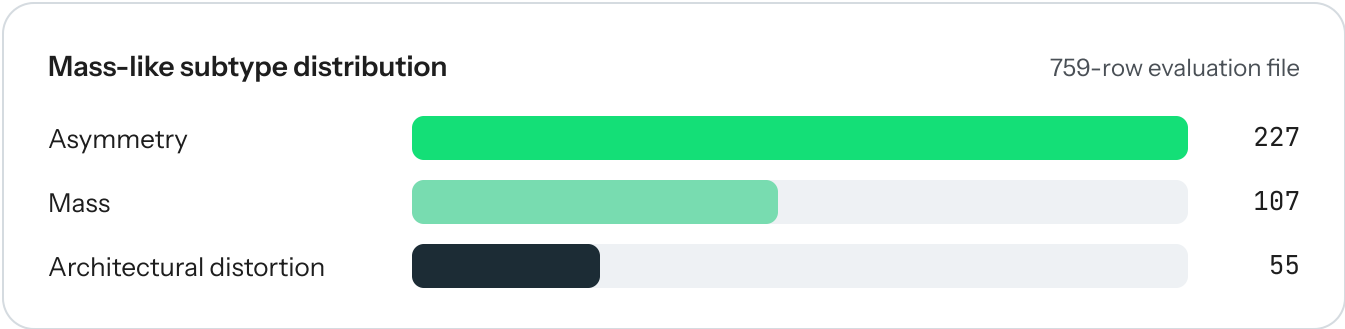


Figure 3. Subtype values on the localizer’s mass-like slot. Bars are drawn to scale from the counts reported in the text.

Subtype is consumed in exactly three places, and the differences between them are deliberate. In **M3**, subtype selects the noun of the injected sentence, since in **M3** the VLM has contributed no noun and the localizer’s is the only one available; when subtype is null the fallback is the asymmetry template, asymmetry being the lowest-commitment of the three finding nouns. In **M5**, subtype selects which of several positive VLM sentences receives the geometry, with a priority order as fallback when no sentence matches. In **M1**, subtype is used only to raise a warning: when the VLM’s noun and the localizer’s subtype refer to different ACR categories, the orchestrator keeps the VLM’s noun, emits a disagreement warning, and does nothing else.

M3

Selects the noun

Null falls back to the asymmetry template

M5

Selects the sentence

Priority order as fallback

M1

Raises a warning only

Keeps the VLM’s noun

The **M1** treatment is worth dwelling on because the alternative is tempting and wrong. If the VLM describes a mass and the localizer’s subtype says asymmetry, one could split the case into an enrichment of the VLM’s mass sentence plus an injection of a separate asymmetry sentence. That would report one lesion twice. The localizer slot fired once and is presumed to describe the same lesion the VLM described; the disagreement is about what to call it, not about how many there are, and the policy resolves naming disagreements in favor of the radiologist-trained model. Synonymy is handled explicitly so the warning does not fire spuriously: mass, nodule, and lesion all match the mass subtype, and distortion matches architectural distortion.

The M5 priority order was corrected during review. The original order placed mass above architectural distortion, which attaches the localizer's geometry to the lower-yield finding when the VLM has described both, since architectural distortion carries among the highest positive predictive values for malignancy of any mammographic finding. The revised order places architectural distortion at least equal to mass and both above asymmetry, and applies only as a fallback when subtype matching has failed.

REVISED M5 FALLBACK

Architectural distortion

≥

Mass

>

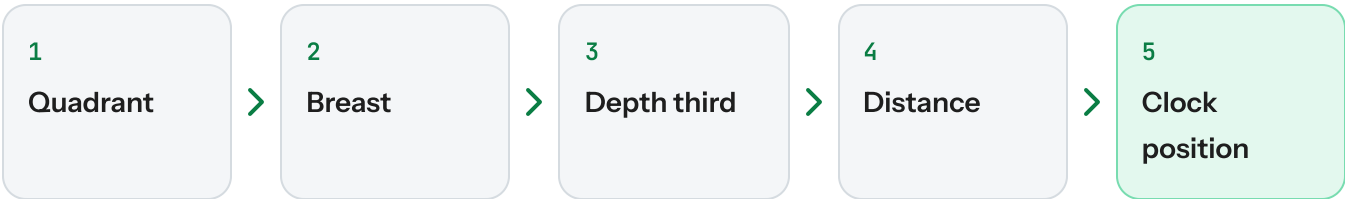
Asymmetry

The scoping of the M1 warning is enforced twice, once in the prompt and once in code, which is a reasonable illustration of how much of this system's reliability is structural rather than model-dependent. The prompt reserves the warning for M1 and states explicitly that M3 must not emit it, giving the reason: in M3 the VLM has no noun to disagree with, so subtype-driven sentence selection is the normal path and not an anomaly. Code then strips the warning after the fact whenever the recorded action is not M1 and records a separate audit note that it did so. The prompt instruction is the mechanism; the code check is the guarantee.

The prompt instruction is the mechanism; the code check is the guarantee.

4.5 Location phrase construction and spatial consistency

Injected and enriched sentences share a single location-phrase builder, rendered in fixed order: the quadrant clause, then the breast, then the depth third, then the distance, then the clock position. Depth is rendered in the anatomically idiomatic form, “in the anterior third” rather than “at anterior depth,” and the internal slot label “mass-like finding” is banned from output entirely, since it is a data-structure name and not a clinical term.



SPATIAL-CONSISTENCY RULE	
Full quadrant and clock position	Consistent by construction; both rendered
Hemisphere-level quadrant and clock position	Hemisphere clause dropped; the clock is more specific
Clock absent	Quadrant or hemisphere rendered as available

Figure 4. The fixed render order of the location phrase, and the rule that resolves a quadrant and a clock position by specificity.

One rule in this builder is a clinical-safety fix rather than a matter of style. The localizer can emit a hemisphere-level quadrant, such as upper or outer, together with a clock position, and the two can be inconsistent: an upper hemisphere together with a clock position of four describes a lower-breast location for a left breast. Rendering both produces a sentence that asserts two contradictory locations, and, critically, a reader cannot detect the contradiction without independently recomputing the mapping from clock position to quadrant. The rule resolves this by specificity. A full quadrant and a clock position are consistent by construction and both are rendered. A hemisphere-level quadrant together with a clock position causes the hemisphere clause to be dropped, since the clock is the more specific of the two signals and a vague hemisphere should not be permitted to contradict it. When the clock is absent, the quadrant or hemisphere is rendered as available. The governing instruction, stated directly in the prompt, is never to assert a quadrant and a clock that point to different regions.

05

Constraining the language model

5.1 Deterministic BI-RADS arithmetic

The most important structural property of the orchestrator is that its language model does not hold final authority over the clinical assessment.

The model returns a JSON object containing the reconciled findings text, a final BI-RADS value, an object per axis recording the action taken and whether it promoted, and a list of warnings. That object is validated against a schema. The final BI-RADS value is then recomputed in code from the VLM's input assessment and the two action codes, according to a rule that admits exactly one transition: if the VLM's assessment was 1 or 2, and the mass axis injected or the calcification axis injected or replaced, the result is 0; otherwise the result is the VLM's original value. Values outside the constrained set pass through unchanged, and a missing input assessment yields a missing output.

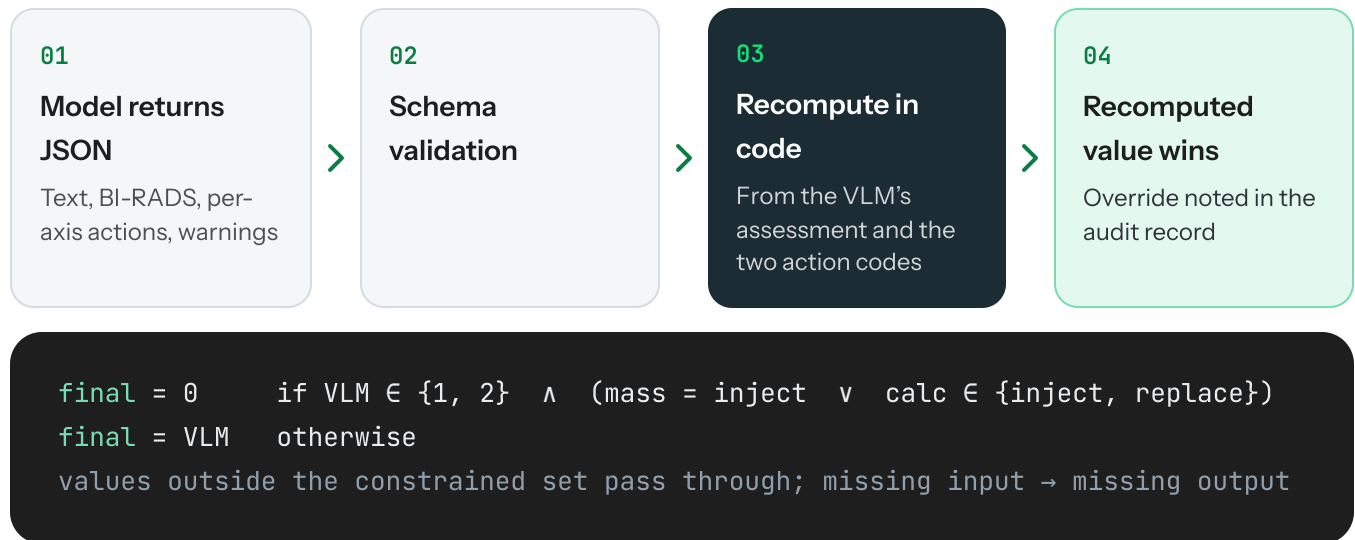


Figure 5. BI-RADS arithmetic is removed from the model's authority. The rule admits exactly one transition.

When the model's own reported value disagrees with the recomputed one, the recomputed value wins and a warning is appended to the audit record noting the override. The practical effect is that a four-billion-parameter local model cannot move the BI-RADS assessment outside the permitted set regardless of what it generates. It can be wrong about the action code, and Section 6 treats that as a live risk, but it cannot be wrong about the arithmetic that follows from the action code.

5.2 The three-layer target, and what remains undone

Removing BI-RADS arithmetic from the model is a partial implementation of a larger structural intent identified during review, and the unfinished remainder is the highest-leverage open item in the subsystem.

The action code is itself deterministic given the VLM's state on an axis and the localizer's status. Detecting whether a VLM sentence negates a finding, matching calcification descriptors against benign and suspicious pattern lists, extracting the BI-RADS suffix, and

reading the localizer’s status are all mechanically decidable operations. Asking the language model to perform that classification in addition to the prose rewrite introduces an error class that is hard to audit, and, more damagingly, it contaminates the per-scenario evaluation that Section 7 depends on. If results under **M3** are poor, the cause may be a poorly constructed injected sentence or it may be that cases were misclassified into **M3** in the first place, and the evaluation as currently constituted cannot distinguish the two.

The intended end state separates three layers. Code performs BI-RADS extraction, negation and pattern detection, and localizer status reading, and emits the action code and the final assessment deterministically. The language model performs prose rewriting only, for an action that has already been decided. Schema validation then checks prose-only output. BI-RADS arithmetic has moved into the first layer. Action classification has not, and until it does, per-scenario numbers should be read as measuring the reconciliation policy and the classifier of that policy jointly rather than separately.

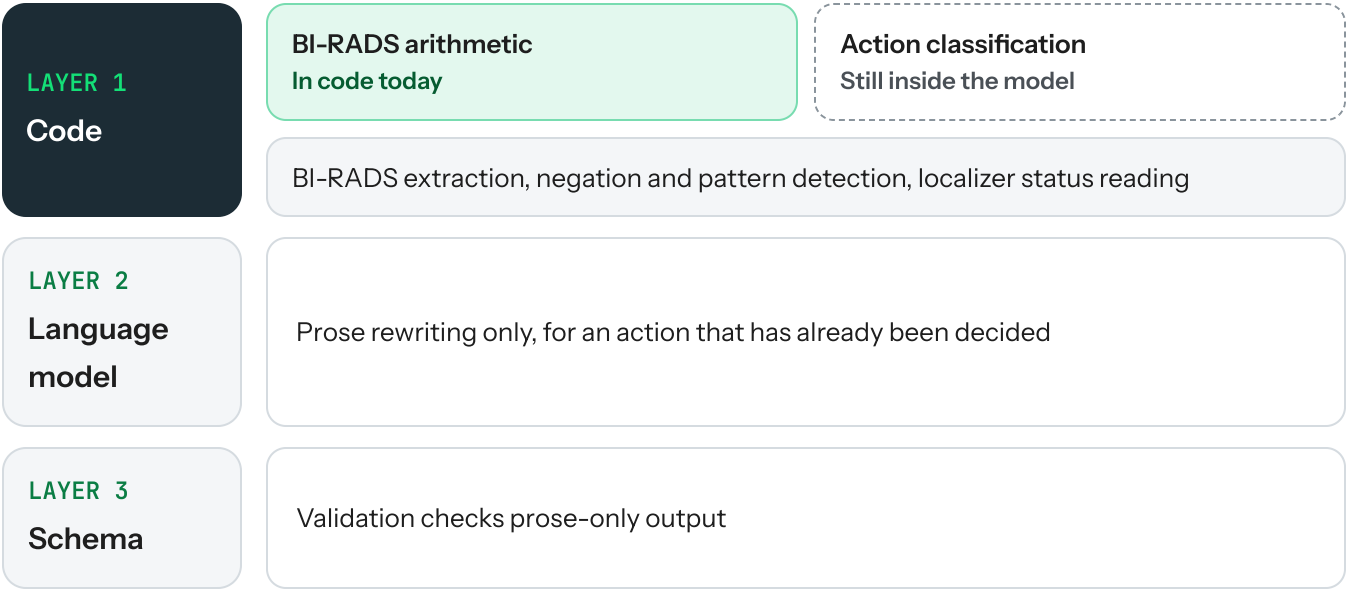


Figure 6. The intended three-layer end state. BI-RADS arithmetic has moved into the first layer; action classification has not.

An interim mitigation is in place, chosen specifically because the weakest link is a small model on rule-dense input. The canonical hard case is a single sentence containing both a negation and a positive finding, of the form “no discrete mass is identified, though an asymmetry is noted,” where a classifier that keys on the negation alone will reach **M3** and a classifier that keys on the finding noun alone will reach **M1**. The system prompt therefore carries eight inlined worked examples, covering hard negation, the multi-sentence tiebreak, the benign-replacement case, the split of benign and suspicious sentences, the fully

degenerate localizer output in which only laterality is known, and the quadrant-and-clock consistency rule. These are few-shot demonstrations doing the work that a deterministic classifier should do, and they should be understood as scaffolding rather than as the design.

THE CANONICAL HARD CASE

“no discrete mass is identified, though an asymmetry is noted”

Keys on the negation alone: reaches M3

Keys on the finding noun alone: reaches M1

5.3 Auditability

Every reconciled breast emits an audit record containing the model’s raw response, both action codes, the warning list, and the sampling seed. The warning vocabulary is deliberately partitioned by origin, as shown in Table 4: some warnings are emitted by the language model as part of its response, and others are appended by code, either because the condition is only observable outside the model or because the code is correcting the model.

Warning	Source	Condition
loc_disagrees_with_vlm_mass	model	M2 : VLM positive on mass-like, localizer silent
loc_disagrees_with_vlm_susp_calc	model	C2 : VLM suspicious calcification, localizer silent
loc_present_with_vlm_benign_pattern_calc	model	C6 : a benign calcification sentence was replaced
loc_subtype_disagrees_with_vlm_noun	model	M1 : VLM noun and localizer subtype are different ACR categories
loc_laterality_mismatch	model	Inner localizer laterality is a third value; rare by construction
llm_fallback_used	code	Schema retries exhausted; hard fallback taken
llm_birads_overridden_by_code	code	Model’s assessment disagreed with the deterministic computation

Warning	Source	Condition
<code>dropped_subtype_disagrees_</code> <code>warning_off_m1</code>	<code>code</code>	An <code>M1</code> -only warning was stripped from a non- <code>M1</code> action

Table 4. Warning vocabulary, partitioned by whether the warning originates in the language model’s response or is appended by surrounding code.

Two consequences of this partition are worth naming. The first is that the two disagreement warnings are the localizer’s own miss log. Their rate is not primarily a signal to a human reviewer; it is a training signal, identifying the cases in which a radiologist-trained model saw something the detector did not, which is exactly the set a subsequent localizer training round should be evaluated against. The second is that because the action codes were held stable across the prompt rewrite that changed `C6`, the depth rendering, the `M3` noun selection, and the `M5` priority order, audit records and downstream arithmetic remain comparable across prompt revisions. The taxonomy functions as a versioning interface between the policy and everything that consumes it, which is a stronger claim than saying it is documentation.

06

Known risks and open questions

6.1 Promotion to Category 0 has no confidence gate

A single localizer firing with no VLM corroboration is sufficient to force a recall. The findings the orchestrator consumes were thresholded at fixed cut-offs applied upstream, at submission time, and are not tunable at orchestration time; a per-finding confidence score exists upstream but the orchestrator’s input payload does not currently carry it into the decision. At natural screening prevalence, where cancer occurs in roughly half a percent of studies, the rate at which the localizer fires on VLM-silent breasts is the dominant term in the pipeline’s specificity cost, because it is the only path that creates a recall out of a negative read.

This is the single most consequential open question in the subsystem, and it has two parts. The engineering part is to pass the confidence score through the payload and make the promotion threshold a configuration parameter rather than an implicit consequence of upstream thresholding. The clinical part is to decide what confidence justifies a recall on a

breast that a radiologist-trained model read as negative, and whether the mass-like and calcification axes should use the same threshold, given that their base rates, their detector reliability, and the consequences of a false recall differ.

ENGINEERING PART

Pass the confidence score through the payload; make the promotion threshold a configuration parameter.

CLINICAL PART

Decide what confidence justifies a recall, and whether the two axes share a threshold.

6.2 **C6** recall inflation

Section 4.3 states the argument in full. In risk terms: **C6** fires on a common co-occurrence, converts it into a recall, and cannot be removed because it is the only recovery path for a benign mislabel. It requires a confidence gate for the same reason **M3** and **C3** do, and it requires its recall-rate contribution to be measured on its own rather than inside an aggregate, because the aggregate will not reveal whether the policy's specificity cost is concentrated in this one rule.

6.3 Schema and unit mismatch between contract and implementation

The declared input schema for a localizer region does not match what the localizer emits, and the discrepancy was verified against 394 present findings. The schema declares depth as a required three-value enumeration, while the localizer frequently emits an unspecified value that is not a member of that enumeration and is not declared optional. It declares distance from the nipple as a required float in millimetres, while the localizer emits a nullable float in centimetres. It expects a per-finding type field, where the localizer buckets findings into mass-like and calcification arrays carrying no type field. And it does not model the status, characterization, or clock fields at all.

Field	Declared schema	Localizer emits
Depth	Required three-value enumeration	Frequently an unspecified value outside the enumeration
Distance	Required float in millimetres	Nullable float in centimetres
Type	Per-finding type field	Mass-like and calcification arrays, no type field
Status, characterization, clock	Not modeled	Present

Figure 7. The declared input contract against the localizer’s actual output, verified against 394 present findings.

The two possible consequences are mutually exclusive and both are serious. If the schema validates the localizer’s input, most present findings are rejected on the depth field and never reach the reconciliation policy. If it does not validate, every injected distance is wrong by a factor of ten, and wrong in the direction of producing a plausible-looking number: a lesion nine centimetres from the nipple rendered as ninety would be obviously wrong, but the mm-declared, cm-supplied direction yields understated distances that read as reasonable. This is a code fix rather than a prompt fix, and the correct target is a model of the localizer’s actual output shape, in centimetres, admitting unspecified and null values.

6.4 Multi-region orchestration is not supported

The payload builder retains only the first mass-like region and the first calcification region from the localizer’s output for a breast; subsequent regions are dropped silently. The localizer’s emission layer supports up to four regions per bucket per breast, and the emission design explicitly notes that multiple lesions per breast are clinically common. The orchestrator is consequently the narrowest point in the pipeline: geometry that the localizer computed, and that the pipeline paid for, is discarded at the reconciliation boundary. The drop is silent, which means it does not appear in the audit record and cannot be counted from the outputs.

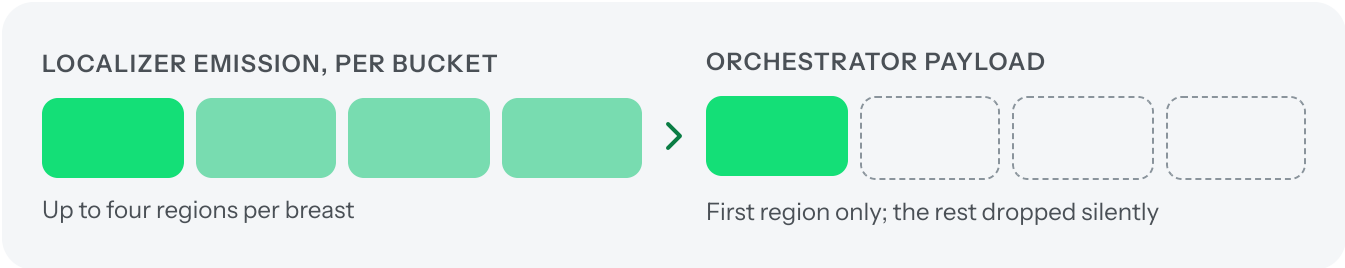


Figure 8. The orchestrator is the narrowest point in the pipeline.

6.5 An unresolved alternative: localization as a corroboration gate

The policy described in this paper is augment-only by invariant. A subsequent policy exploration sketches a materially different fusion policy in which localization acts as a corroboration gate: a suspicious VLM finding that the localizer does not corroborate is dropped rather than kept and flagged, and a localizer firing on a VLM-silent breast produces no report prose at all rather than an injected sentence and a promotion. Table 5 sets the two policies side by side.

Condition	Augment-only (current)	Corroboration gate
VLM positive mass-like, localizer absent	Keep verbatim and warn (M2)	Drop the VLM sentence
VLM suspicious calcification, localizer absent	Keep verbatim and warn (C2)	Drop the VLM sentence
VLM silent, localizer present	Inject and promote to 0 (M3 , C3)	Keep VLM verbatim; no prose emitted
VLM benign calcification, localizer present	Drop benign, inject suspicious, promote (C6)	Keep VLM verbatim but drop the benign sentence
Both models fire	Enrich, keeping the VLM's noun	Enrich, keeping the VLM's noun

Table 5. Two fusion policies. They agree only on the agreement case, and invert on both disagreement cases.

The two policies agree only where both models agree. They invert on both disagreement cases, and the inversion is not a matter of degree. The current policy maximizes sensitivity by trusting the radiologist-trained model and granting the localizer authority to add findings; the gate maximizes precision by requiring two-model corroboration before any suspicious claim reaches the report. Which is correct depends on where the pipeline sits in a clinical

workflow, what the reviewing radiologist is expected to do with a draft, and what the tolerable recall rate is: product and clinical questions rather than technical ones. This remains undecided, and the policy sketch is presented in its source as an example of a gating policy rather than as an adopted decision.

The framing that exploration contributes is worth carrying forward independently of which policy wins. Before writing any rules, state whether localization is a gating signal, an enrichment source, or a tie-breaker, and describe how agreement, disagreement, and single-model firing are each to be handled. The current policy’s answer is that localization is an enrichment source with injection authority. The gate’s answer is that it is a gate. The rule tables in each case follow from that choice, rather than motivating it, and a rule table written without stating the choice first is a policy whose commitments are implicit.

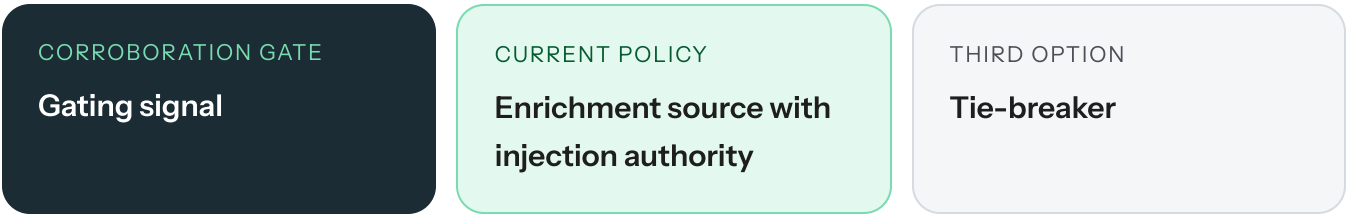
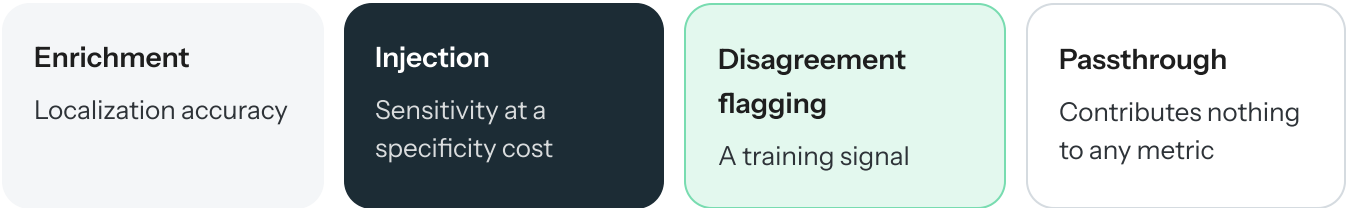


Figure 9. State the role of localization before writing any rules.

07

Evaluation framework

The orchestrator cannot usefully be evaluated as a single aggregate number, because its effect is distributed unevenly and in different units across the action codes. Enrichment adds spatial precision to a finding both models already agree exists, and its value is measured in localization accuracy. Injection adds sensitivity at a specificity cost, and its value is a trade-off rather than a quantity. Disagreement flagging changes no report text at all and produces a training signal instead. Passthrough is correct behavior that contributes nothing to any metric. An aggregate score averages these into a number that no decision depends on.



The evaluation is therefore specified per action code, on a natural-prevalence cohort, with bootstrap confidence intervals, as set out in Table 6.

Codes	Measured quantity	Decision informed
M1 , C1	Change in F1, quadrant accuracy, and depth accuracy relative to the VLM alone, on cases where both models agree	The localizer’s additive value
M3 , C3	Sensitivity gained against recall rate inflated	The core deployment number
C6	Recall-rate cost in isolation	Whether C6 ships, ships gated, or is replaced by a review flag
M2 , C2	Localizer miss rate relative to the VLM	Where to spend localizer training effort
All	Disagreement rate, and which model is correct (requires ground truth)	Overall allocation of effort

Table 6. Per-action-code evaluation. Each row measures a different quantity and informs a different decision.

The natural-prevalence requirement is not a refinement. Evaluating injection behavior on an enriched cohort understates its specificity cost by construction, since the denominator of the false-recall rate is precisely the population of negative studies that an enriched cohort under-represents, and that specificity cost is the number on which the deployability of localizer-only injection turns.

Two further measurements are specified. The first is provider agreement: running a stronger reference model against the shipping local model and reporting agreement on the action code, separately from agreement on the final text. This isolates the classification decision from the prose, and, given Section 5.2, it is the closest available proxy for the reliability of the component that should not be inside the model at all. The second is verification of determinism in the evaluation harness itself, asserting temperature and seed rather than trusting that configuration matches intent.

The pipeline’s stated performance targets are a suspicious mass F1 of at least 0.5, a calcification F1 of at least 0.65, density accuracy above 0.7, and localization weighted F1 of at least 0.5. Results against these targets are recorded as open items in the design specification and are not documented in any linked artifact at the time of writing. No performance claim is made in this paper.

STATED PERFORMANCE TARGETS

RESULTS NOT YET DOCUMENTED

≥ 0.5

Suspicious mass F1

≥ 0.65

Calcification F1

> 0.7

Density accuracy

≥ 0.5

Localization weighted F1

08

Scope, limitations, and regulatory positioning

Several capabilities are explicitly outside the current design and should not be inferred from it. Cancer probability classification is wired but disabled, with the reported probability currently derived from the segmentation maximum rather than from a trained classifier. Prior-study comparison is a separate effort and is not part of this release; preliminary work indicates that using prior images alone is complicated by inter-study tissue movement and compression, which is a substantive obstacle rather than an engineering delay. Three-dimensional tomosynthesis is out of scope, as the pipeline handles two-dimensional screening acquisitions only. Diagnostic workup views are filtered out at view selection. Multi-region orchestration is capped at one region per bucket, as described in Section 6.4. And there is no cross-laterality reasoning; each breast is reconciled independently.

Cancer probability

Wired but disabled

Prior-study comparison

A separate effort

Tomosynthesis

2D screening acquisitions only

Diagnostic workup views

Filtered out at view selection

Multi-region orchestration

One region per bucket

Cross-laterality reasoning

Each breast reconciled independently

The regulatory positioning bears directly on the orchestrator's design rather than merely surrounding it. The software is intended for distribution as a component integrated into a customer's medical device, not as a standalone medical device, and the integrating customer is responsible for obtaining and maintaining clearance, defining intended use, and performing clinical validation. A requirement follows from this positioning: the output must be modifiable for a customer's application without retraining the vision-language model.

The orchestrator prompt is what satisfies that requirement, which means the prompt is not an implementation detail but the product's customization surface. Two properties of the design exist because of this. The invariant that VLM findings content is preserved rather than removed is what makes offering that customization surface defensible, since it bounds what a prompt modification can do to the clinical content. The action-code taxonomy is what makes a modification's effect verifiable, since it renders the behavioral change countable rather than merely observable in prose.

DEFENSIBLE

The preservation invariant bounds what a prompt modification can do to the clinical content.

VERIFIABLE

The action-code taxonomy renders the behavioral change countable rather than merely observable in prose.

— 09

Conclusion

This paper described the orchestration layer that reconciles a generative vision-language model with a discriminative localizer in a screening mammography pipeline whose output is a draft findings section. The two upstream models have close to disjoint competences, and the resulting disagreements are not noise to be smoothed but decisions with clinical consequences, of which the sharpest arises from the localizer's structural inability to characterize a finding as benign. The layer resolves these disagreements under four invariants: VLM findings content is never deleted, the BI-RADS assessment only ever becomes more actionable, the mass-like and calcification axes are reconciled independently, and no geometry is used that the localizer did not supply. Twelve action codes enumerate what those invariants require in each combination of upstream states, and one code deliberately violates the first invariant because it is the only path by which a benign mislabel can be recovered. Deterministic code recomputes the BI-RADS assessment from the action codes and overrides the language model when the two disagree, so that a small local model cannot move the clinical assessment outside its permitted set.

The design's principal contribution is attributability rather than accuracy: every sentence in a generated report is traceable to an upstream model and a rule, which is what makes both the policy and its unresolved questions legible. Those questions are stated here rather than

deferred. Promotion to Category 0 on an uncorroborated localizer firing is unconditional, and it is that injection rate, not any component metric, that determines the pipeline’s specificity at natural prevalence. A declared input schema disagrees with the localizer’s actual output in a way that places every injected distance at risk. Action classification remains inside the language model against the intent of the design, which means per-scenario evaluation currently measures the policy and its classifier jointly. And a competing fusion policy exists that inverts the central invariant, treating localization as a corroboration gate rather than an enrichment source, on a question that is clinical rather than technical and that has not been decided.

Two directions follow. The narrower one is to complete the separation already begun for BI-RADS arithmetic, moving action classification into deterministic code so that the language model performs prose rewriting alone and per-scenario numbers become trustworthy. The broader one is suggested by a parallel exploration in which a single vision-language model is replaced by an encoder carrying many finding classifiers, each firing classifier triggering downstream segmentation and localization through an orchestrator, with each structured finding enriched and assembled into the same target schema. The claimed advantages are modularity, independent per-component improvement, and isolated failure modes; the costs are component count and deployment surface. Both directions converge on the same observation. If orchestration generalizes from reconciling two models to assembling many model outputs into a structured report, then the orchestrator ceases to be a reconciliation patch and becomes the pipeline’s assembly layer, at which point the reliability of its decision logic matters more than the fluency of its prose, and the case for moving that logic out of the language model strengthens accordingly.

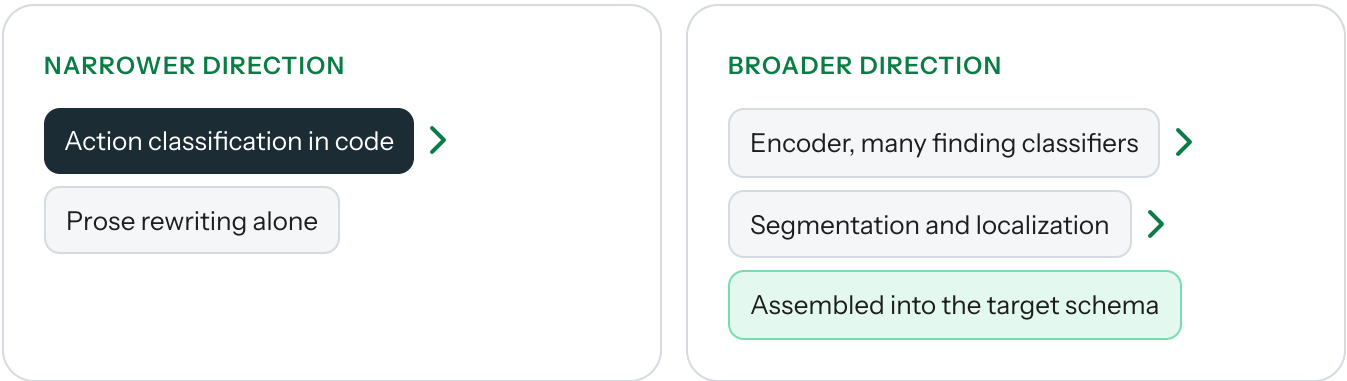


Figure 10. Two directions that converge on the orchestrator as the pipeline’s assembly layer.

THE PRINCIPAL CONTRIBUTION

Every sentence in a generated report is traceable to **an upstream model and a rule.**

APPENDIX A

BI-RADS aggregation

The final assessment is the disjunction of the two axes, and promotion occurs at most once. When neither axis promotes, the VLM’s original assessment sentence is preserved verbatim rather than regenerated.

Mass-like axis	Calcification axis	Final assessment
no change	no change	unchanged (the VLM’s value)
promote to 0	no change	Category 0
no change	promote to 0	Category 0
promote to 0	promote to 0	Category 0 (promoted once)

Table 7. BI-RADS aggregation across the two axes.

The canonical assessment sentences are fixed strings rather than generated text: Category 0 renders as “Incomplete: Needs additional imaging evaluation,” Category 1 as “Negative: No findings of concern,” and Category 2 as “Benign finding: Non-cancerous finding.”

CATEGORY 0

Incomplete: Needs additional imaging evaluation

CATEGORY 1

Negative: No findings of concern

CATEGORY 2

Benign finding: Non-cancerous finding

Edge case handling

#	Trigger	Action
E1	No parseable assessment suffix in the VLM text	Pass through verbatim; final assessment null; skip enrichment and injection; log a parse error
E2	No localizer record for this study and side	Treat both axes as absent; no promotion is possible; log the omission
E3	Localizer present, quadrant unspecified and clock null	Omit both clauses; never emit a literal placeholder string
E4	Distance null	Drop the distance phrase including its units
E5	Only depth is populated	Render the depth clause alone
E6	Hemisphere-level quadrant	Render as an aspect rather than a quadrant, subject to the spatial-consistency rule of Section 4.5
E7	Localizer laterality disagrees with the breast being reconciled	Drop the localizer slot for this breast; log the mismatch
E8	VLM text refers to the contralateral breast	Keep verbatim; never match against a localizer region
E9	VLM finding parsed as belonging to neither axis	Pass verbatim; attempt no enrichment
E10	Both axes inject	Emit the mass-like sentence first, then the calcification sentence; promote once

Table 8. Edge case handling.

Sources

This paper is a synthesis of internal design documentation rather than a report of new work. Its substantive content derives from the Mammo VLM v3 Software Design Specification and its companion architecture diagrams; the current revision of the orchestrator system prompt together with its changelog; the scenario catalogue enumerating the VLM-by-localizer combinations with worked examples; a prompt review conducted against the localizer's real output schema, which is the source of the schema mismatch of Section 6.3, the spatial-consistency rule of Section 4.5, the priority-order correction of Section 4.4, and the structural argument of Section 5.2; a subsequent fusion-policy exploration, which is the source of Section 6.5; and a spike specification for the multi-classifier orchestration direction discussed in Section 9. Implementation references are the orchestrator module's prompt builder, its reconciliation entry point, its deterministic assessment computation, and its input and output schemas.