Selective Answering for Medical VQA via Parallel Independent Claim Verification

Donghee Han **Jiwon Jeong** **Mun Yong Yi**[†]
KAIST, Republic of Korea
{handonghee, zzioni, munyi}@kaist.ac.kr
[†]Corresponding author

Abstract

Medical visual question answering is a safety-critical decision problem where models must answer correctly when visual evidence is sufficient and abstain when it is not. Existing inference-time agentic methods attempt to reduce hallucinations through verification, but typically route the final decision through a single language-model judge that integrates all evidence into a single prompt, making the system vulnerable to answerability bias. We propose **PICV** (Parallel Independent Claim Verification), a selective medical VQA framework that reformulates answerability assessment as independent claim verification with transparent aggregation. Given a question, PICV decomposes it into a compact set of typed visual claims, verifies each claim in an isolated prompt without access to other verdicts, and combines the resulting verdicts using a deterministic rule rather than an additional model. This design produces both an answer/abstain decision and a typed unanswerability reason, improving interpretability while reducing judge-induced bias. We evaluate PICV on an automated answerability benchmark constructed from two public radiology datasets using three reproducible corruption procedures, avoiding heavy VLM judging and human annotation at evaluation time. Across four VLMs, PICV consistently outperforms prompting and agentic baselines. Ablation studies show that the gains arise from per-claim isolation and rule-based aggregation, rather than from agentic decomposition alone.

1 Introduction

Medical visual question answering aims to produce an answer to a clinical question conditioned on a medical image [1, 2]. A long line of work has improved performance on this task through attention-based multimodal architectures and, more recently, vision-language models (VLMs), with several systems reporting accuracies above 80% on SLAKE [3] and VQA-RAD [4]. Despite architectural differences, most Med-VQA systems share a fundamental assumption: every test instance is answerable, and the model must select an answer. This assumption induces a *forced-answer regime*, where the model may produce a confident response even when the image lacks the visual evidence required by the question.

This regime is particularly concerning in clinical settings, where abstaining from unsupported questions can be safer than providing a plausible but incorrect answer. Recent work on medical imaging highlights the severity of this issue [5]: frontier VLMs can generate detailed clinical descriptions and even achieve near-top performance on standard VQA benchmarks without access to the image. Moreover, these models perform worse when explicitly told that no image is available and asked to guess than when the missing image is left implicit. This gap suggests not random guessing, but a confident "mirage" mode of reasoning in which models infer clinically plausible answers from language priors alone. Such behavior is unsafe for deployment in patient care because a confident

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

wrong answer can be more harmful than a refusal. Our experiments also demonstrate that standard prompting baselines almost never abstain on unanswerable inputs, even when the image and question are clearly incompatible.

This limitation becomes more pronounced in recent agentic verification frameworks. Methods such as ReAct [6], Self-Ask [7], Chain-of-Verification [8], and MEDIQ [9] improve factuality in text and vision-language tasks by introducing intermediate reasoning or verification steps. However, on our medical benchmark, these frameworks ultimately collapse verification into *a single VLM judge* that integrates multiple evidence channels into a single prompt. We find that this judge is systematically biased toward declaring inputs *supported*: across more than 200 missed unanswerable cases, the original single-judge variant of our pipeline emits a REFUTED verdict for the choice or region claim *zero* times. Consequently, strong negative evidence identified by individual visual probes is overwritten at the final combination stage, suggesting that the core challenge is not simply to add verification, but to preserve claim-level verification signals before they are diluted by a single judging prompt.

Motivated by this observation, we propose **PICV** (Parallel Independent Claim Verification), a selective-answering workflow for frozen black-box VLMs. To mitigate single-judge bias, PICV follows an *independent-verification-and-deterministic-aggregation* strategy. Given a question, it decomposes the required visual evidence into a small set of atomic claims, each aligned with a typed unanswerability reason, entity presence, image region, or choice validity. Each claim is then verified in a separate prompt without access to other verdicts or any draft answer. Entity and region checks use the corresponding visual claim; per-option choice checks additionally include the question. This isolation prevents one channel's evidence from contaminating another and provides a more isolated set of verdicts for downstream combination.

PICV combines the independent verdicts via a *deterministic programmatic rule*, emitting an answer/abstain decision with a typed abstention reason and confidence equal to the fraction of SUPPORTED claims. Unlike single-judge agentic methods that expose only a final decision and can overwrite intermediate negative evidence, PICV preserves per-claim verdicts and delegates their combination to an explicit rule. This addresses the observed *supported*-bias through three mechanisms: a per-option choice verifier replaces the bundled "exactly one matches" claim so unsupported options cannot be masked; a two-step region verifier first elicits the body region in free text and then checks it against an entity-to-region table via programmatic string matching; and a decomposition sanity check suppresses fabricated entities in open-ended questions while preserving the legitimate "No" in yes/no containment questions. PICV requires no fine-tuning and uses a manually engineered, task-specific 62-entry entity-to-region table. To evaluate selective answering, we further construct a deterministic answerability benchmark by corrupting answerable benchmark instances along three axes (image mismatch, question corruption, and choice removal); labels are assigned by construction, so scoring uses the assigned labels. Label validation is described in Appendix I.2.

We evaluate PICV on the SLAKE/VQA-RAD answerability benchmark under a selective-answering setting across four VLM backbones: two frontier API models and two on-premise 4–7B models, including a medical-domain-specialized VLM. Across all backbones, PICV achieves the highest answerability F1 in seven of eight cells under labels re-verified by a single AI rater, against prompting and agentic baselines, including Direct Answer, Two-stage Verify, CoT-Abstain, ReAct-Med, Self-Ask, and MEDIQ-VLM. Ablation studies show that these gains cannot be attributed to agentic decomposition alone. Instead, performance improves when decomposition is paired with per-option choice verification and deterministic aggregation, which isolates the supported-bias of bundled judges. Additional analysis further shows that the tested medical-domain-pretrained model does not close the calibration gap: the medical-specialized VLM in our evaluation exhibits the same systematic failure to abstain as general-purpose VLMs.

Our contributions are summarized as follows:

- We identify *single-judge supported bias* as a key failure mode of agentic methods for medical VQA: independent verification signals from visual probes are collapsed into an "answerable" verdict by a single judging prompt.
- We propose PICV, a selective-answering workflow that reformulates medical VQA as *independent claim verification with deterministic aggregation*. Atomic visual claims are verified in isolated prompts and combined by a programmatic rule that outputs an answer/abstain decision.

- We construct a deterministic answerability benchmark from public radiology datasets using three reproducible corruption procedures. Answerability labels are assigned reproducibly without a VLM judge; additional controls examine label validity and benchmark–method alignment.
- We evaluate PICV against strong baselines across four VLM backbones. Ablations attribute the gains to per-claim isolation and rule-based aggregation, rather than agentic decomposition alone.

2 Related Work

PICV sits at the intersection of medical VQA, hallucination and abstention in VLMs, selective prediction, and inference-time agentic verification.

Medical VQA. Two public radiology benchmarks dominate the medical VQA literature: VQA-RAD [4], a clinician-curated set of 315 radiographs with $\sim$3.5k yes/no and open-ended questions, and SLAKE [3], a semantically labelled multi-organ dataset of $\sim$14k pairs spanning CT, MRI, and X-ray. State-of-the-art systems include domain-specialised models such as MedGemma-4B [10] and LLaVA-Med [11], plus frontier general-purpose VLMs. Reported accuracies frequently exceed 80%, but the underlying test sets contain only answerable inputs by construction. A separate line studies *clinical reasoning agents*: MEDIQ [9] introduces an iterative-questioning agent that decides when it has gathered enough information to answer; an image-conditioned variant of MEDIQ provides our strongest external agentic baseline.

Hallucination and abstention in VLMs. A recent survey [12] documents that VLMs maintain near-baseline VQA accuracy even when the image is replaced by Gaussian noise, indicating language-prior dominance. POPE [13] systematically quantifies object hallucination in natural-image VLMs; the problem is amplified in specialised medical domains [14]. Mitigation strategies fall into two families. *Decoding-level* methods such as Visual Contrastive Decoding [15] contrast outputs from original and perturbed images to suppress hallucinations during generation, but require token-level logits and are incompatible with frontier APIs. *Inference-time agentic* methods operate on completed outputs and remain compatible with closed APIs. PICV inherits this no-fine-tuning, API-compatible philosophy. Its specific contribution is to produce a *decision* (answer or abstain with a typed reason) rather than to suppress hallucinated content within an answer.

Selective prediction. Selective prediction has a long history [16, 17], with the central goal of allowing models to abstain when their predictions are unreliable. In VQA, this idea has been adapted mainly through calibrated confidence over the answer distribution [18], while related work on self-knowledge studies whether language models can recognize their own limitations in text-only settings [19]. Recent work further emphasizes that abstention should not be treated as a single undifferentiated outcome: *typed* uncertainty, which separates data-level ambiguity from model-level incapability, is necessary for choosing appropriate downstream actions, such as requesting an additional image, reformulating the question, or routing the case to a specialist. PICV operationalizes this view through typed abstention reasons (entity_absent, image_mismatch, choice_invalid, evidence_insufficient), each of which corresponds to a distinct failure mode and supports a different downstream response.

Inference-time agentic verification. ReAct [6] interleaves thought and action; Self-Ask [7] decomposes a question into sub-questions; Reflexion [20] iteratively refines based on a reflection signal. *Chain-of-Verification* (CoVe) [8] drafts an initial answer, plans verification questions, and answers each verification question independently. PICV adapts the CoVe principle to medical VQA in three ways: (i) verification is at the level of *atomic visual claims* derived from question presuppositions, not free-text propositions; (ii) the aggregator is a *deterministic programmatic rule* tied 1:1 to typed reasons, replacing CoVe's "revised answer" generation step; (iii) for multiple-choice inputs, each option is verified in its own independent prompt, addressing a structural failure mode (bundled-claim SUPPORTED bias) we document in §4.4 and Appendix G. Multi-agent debate frameworks [21] propose disagreement-as- signal, but recent meta-analysis [22] finds debate frequently fails to outperform self-consistency at higher cost. PICV uses *parallel* (not iterative) independent verification, capturing the same idea at low marginal cost without convergence pathologies.

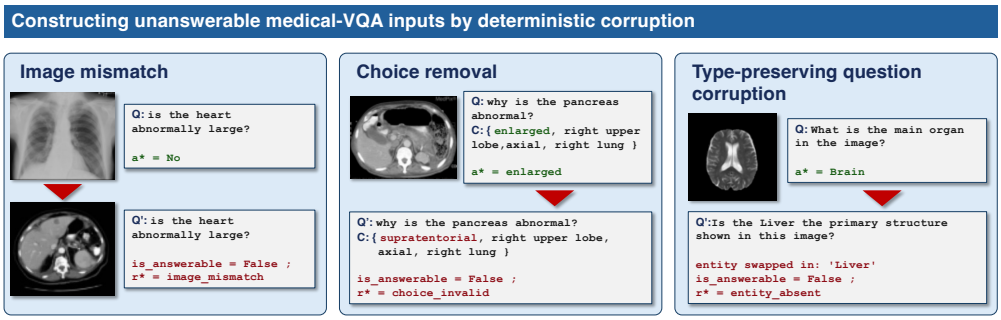

Figure 1: Constructing unanswerable medical-VQA inputs by deterministic corruption. **Left**: *image_mismatch* keeps (Q, C) fixed and swaps in a donor image from the same dataset (modality-matched on SLAKE). **Centre**: *choice_missing* keeps (I, Q) fixed and replaces the gold answer with a same-register distractor. **Right**: *question_corruption* (SLAKE only) substitutes the presupposed entity in Q with a different entity drawn from the same `content_type` pool. Each column's $r^\star$ label is assigned by construction; label validity is assessed separately and no VLM judge is consulted.

3 Method

The methodological core of this paper has two halves. We first describe how we *construct unanswerable medical VQA inputs deterministically* (§3.1); this is the data primitive that makes selective-answering evaluation possible without any VLM judge or human annotation. We then formalise the task (§3.2) and present PICV (§3.3 through §3.7).

3.1 Synthetic Construction of Unanswerable Inputs

Existing medical VQA benchmarks contain only answerable (I, Q, C) triples by construction. To study selective answering we need the opposite, namely inputs where the image *cannot* support the question, and we need them with reproducible assigned answerability labels, so that scoring an abstention decision does not itself depend on a noisy VLM judge. We synthesise unanswerable inputs by deterministically corrupting answerable seeds along three axes that align 1:1 with the typed unanswerability reasons entity_absent, image_mismatch, choice_invalid used by PICV (§3.2). Given an answerable seed $(I, Q, C, a^\star)$ from SLAKE or VQA-RAD and a fixed seed (we use 42), each corruption is a deterministic function returning a labelled variant $(I', Q', C', \text{is_answerable} = \text{False}, r^\star)$ with ground-truth reason $r^\star \in \mathcal{R}$.

Image mismatch (`image_mismatch`). Keep (Q, C) unchanged and replace I with a donor image I' from the *same dataset*. For SLAKE, the donor is sampled from the same modality (CT, MRI, X-ray); no body-region constraint is enforced, and 103/166 SLAKE donors share the original `body_region` label. VQA-RAD provides no modality or region metadata, so its donors are sampled from the full test pool. The swap is intended to remove the visual evidence required by Q; this assumption often fails (Appendix I.2). $r^\star =$ image_mismatch.

Type-preserving question corruption (`question_corruption`, SLAKE only). SLAKE provides a `content_type` annotation (*modality*, *organ*, *abnormality*, *plane*) typing each question by what entity it asks about. We substitute the presupposed entity e in Q with a different entity e' drawn from the *same* `content_type` such that e' does *not* appear in I. "Is there a lesion in the lung?" becomes "Is there a lesion in the kidney?" on a chest scan. Because the substitution stays within `content_type`, the corrupted question is not trivially malformed; only a model that grounds the entity against the image can detect the corruption. $r^\star =$ entity_absent (or image_mismatch when the substituted entity is in a different anatomical region).

Choice removal (`choice_missing`). For multiple-choice instances we remove $a^\star$ from C and replace it with a sampled distractor of the same surface form (length, register, and where possible `answer_type`). We do not insert a "none of the above" sentinel. The image and question remain mutually consistent; the failure is purely at the choice level. $r^\star =$ choice_invalid.

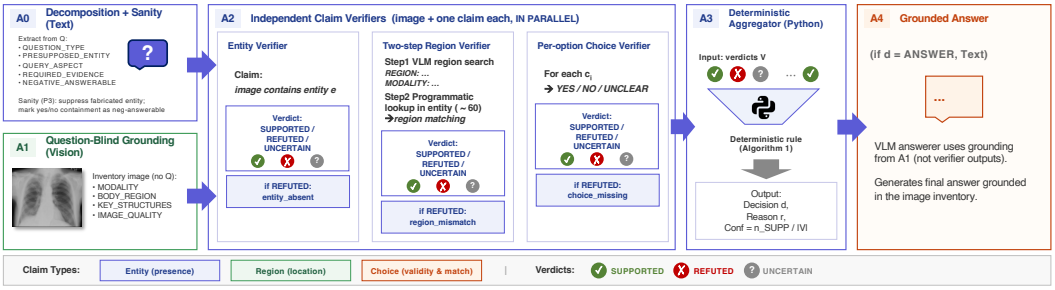

Figure 2: PICV pipeline. A0 and A1 produce structured fields used by A2. A2 prompts are isolated from other verdicts and draft answers; per-option choice checks also include Q. The deterministic aggregator A3 emits (d, r, conf); if $d = \textsf{ANSWER}$, A4 produces the grounded answer.

Naming convention. The three corruption *axes* (`image_mismatch`, `choice_missing`, `question_corruption`) and the typed reasons $\mathcal{R} = \{\textsf{entity_absent}, \textsf{image_mismatch}, \textsf{choice_invalid}, \textsf{evidence_insufficient}\}$ live in distinct namespaces. Two of the three axes share a string with their dominant typed reason (image_mismatch is literally `image_mismatch` for axis `image_mismatch`); `choice_missing` (an axis) is the counterpart of choice_invalid (a reason emitted by PICV), which we keep visually distinct only to signal whether we are talking about the data-generation step or the model's abstention output. We report the per-axis abstention rate (TAR) in §4.3; typed-reason accuracy is not separately scored in our main metric AnsF1.

Reproducible labels and their validity. Each corruption is a deterministic function of $(I, Q, C, a^\star, \text{seed})$, so the assigned answerability and reason labels are reproducible, but not necessarily correct: a donor image can still support the question. The corruption axes also align with our verifier taxonomy. Appendix I.2 examines this alignment and label validity through visual degradation, exclusion controls, and model-based re-verification of image mismatches. These controls show that choice verification contributes substantially to the observed advantage and that some swapped images and corrupted yes/no questions remain answerable. The main evaluation uses labels re-verified by a single AI rater, with uncertain inputs excluded (Appendix H). Per backbone, the benchmark contains 50 original and 50 corrupted inputs per dataset: SLAKE has 14 image mismatches, 13 question corruptions, and 23 choice removals; VQA-RAD has 19 image mismatches and 31 choice removals. Fig. 1 shows one worked example per axis using real benchmark images.

3.2 Problem Formulation

A medical VQA instance is (I, Q, C) with image I, question Q, and optional choice list $C = \{c_1, \ldots, c_k\}$ ($k = 0$ open-ended). Standard VQA produces $\hat{a} \in C$. Selective VQA augments this with $d \in \{\textsf{ANSWER}, \textsf{ABSTAIN}\}$ and, when $d = \textsf{ABSTAIN}$, a typed reason $r \in \mathcal{R}$ with $\mathcal{R} = \{\textsf{entity_absent}, \textsf{image_mismatch}, \textsf{choice_invalid}, \textsf{evidence_insufficient}\}$. Each reason maps to a distinct downstream action: image_mismatch re-routes the case; choice_invalid escalates to free-text answering; the others prompt acquisition of additional images. The original benchmark metric is **answerability F1 (AnsF1)**, binary F1 with positive class "is answerable" computed over d. Secondary metrics: TAR (true-abstain rate on unanswerable), FAR (false-abstain rate on answerable), ECE (expected calibration error), and average VLM calls per sample. We additionally evaluate final-answer quality and coverage–harm trade-offs (§4.5).

3.3 PICV Overview

PICV processes (I, Q, C) through five stages (Fig. 2): **A0** Decomposition + sanity check (VLM, text-only); **A1** Question-blind grounding (VLM, image-only); **A2** Independent claim verifiers (VLM, image + one claim each, in parallel); **A3** Deterministic programmatic aggregator; **A4** Grounded answer (VLM, only if $d = \textsf{ANSWER}$).

The defining property is that the final (d, r) is computed by a **deterministic programmatic rule** over verifier outputs, not by any VLM judge. We argue in §4.2 that this is the structural reason PICV outperforms single-judge external baselines: a single judge integrating multiple evidence channels exhibits a SUPPORTED bias that suppresses abstention even when the underlying probes correctly identify a problem.

3.4 Decomposition with Sanity Check (A0)

Decomposition extracts five fields from Q alone (no image access): `QUESTION_TYPE` $\in$ {YES_NO, OPEN_ENDED}; `PRESUPPOSED_ENTITY`; `QUERY_ASPECT` $\in$ {existence, location, modality, size, count, ...}; `REQUIRED_EVIDENCE` ($\leq$ 12 words); and `NEGATIVE_ANSWERABLE` (YES if a YES_NO question is legitimately answerable as "No" when the entity is absent). VLMs frequently fabricate `PRESUPPOSED_ENTITY` on open-ended questions ("What is the main organ?" $\rightarrow$ "heart"), which poisons the entity verifier. We add a deterministic programmatic *sanity check* (**P3**) that (a) sets `PRESUPPOSED_ENTITY` := none on "what is/are/where is/which/how many/what modality/plane/organ/body" prefixes (skipping the entity verifier); and (b) forces `QUESTION_TYPE` := YES_NO, `NEGATIVE_ANSWERABLE` := YES on "contain/exist/ is there/does the image have" patterns so the aggregator preserves the legitimate "No" path.

3.5 Question-Blind Grounding (A1)

Grounding produces a structured, question-independent inventory of the image: `MODALITY`, `BODY_REGION`, `KEY_STRUCTURES`, `ABNORMALITIES`, `IMAGE_QUALITY`. The grounding prompt is identical for every instance and never contains Q. Question-blindness is intended to avoid question-induced suggestions about entities that may not be depicted.

3.6 Independent Claim Verifiers (A2)

For each instance, PICV constructs up to three families of atomic claims and verifies them with independent prompts. Each verifier operates on its own claim without other verdicts or draft answers; per-option choice verification also includes the question, as specified below.

Entity verifier. Active when `PRESUPPOSED_ENTITY` $\neq$ none after the §3.4 sanity check. Claim: "*The image visibly contains* e." Returns one of {SUPPORTED, REFUTED, UNCLEAR}. REFUTED maps to entity_absent (modulo the negative-answerable override).

Two-step region verifier (P2). Active when `PRESUPPOSED_ENTITY` $\neq$ none. Step 1 (VLM): name the body region in free text with `REGION:` and `MODALITY:` keys. Step 2 (programmatic): an entity-to-region table $\mathcal{M}$ (62 manually engineered entries; e.g. `kidney` $\rightarrow$ {`abdomen`, `retroperitoneum`}; `brain` $\rightarrow$ {`head`, `cranial`, `intracranial`, `skull`}) yields

$$\text{verdict}(e, \text{step1}) = \begin{cases} \text{SUPPORTED} & \text{if any token in } \mathcal{M}(e) \text{ appears in step1} \\ \text{REFUTED} & \text{if } e \in \mathcal{M} \text{ and no token matches} \\ \text{UNCLEAR} & \text{if } e \notin \mathcal{M}. \end{cases} \tag{1}$$

The two-step design forces the VLM to commit to a region *before* the comparison, preventing the loose "this region could plausibly contain e" reasoning that a one-shot claim invites. REFUTED maps to image_mismatch. $\mathcal{M}$ is a static, question-independent external knowledge base of generic radiology anatomy used at inference time; it adds no VLM calls and remains fixed during inference. PICV sits between pure prompt engineering and retrieval-augmented inference (§E). Every entry is released. The map covers 175/357 invoked region claims (49.0%); unmatched entities return UNCLEAR, limiting detection of unseen anatomy. Because P2 keys off broad-region tokens, its contribution is dataset-conditional: it helps when donor swaps cross broad regions (the predominant VQA-RAD case) and is roughly neutral when donors share a broad region (the predominant SLAKE within-modality case); we keep P2 enabled in the default PICV pipeline for cross-dataset uniformity; a P1+P3-only variant for readers who prefer a SLAKE-conditional configuration is included in the repository.

Per-option choice verifier (P1). Active when $|C| > 0$. The naive bundled claim, "exactly one of $\{c_1, \ldots, c_k\}$ correctly answers the question for this image," is interpreted as a single yes/no question, and the verifier almost always returns SUPPORTED by describing the image without comparing the options. PICV instead issues k independent prompts of the form "*Question: Q. Does the answer 'c_i' correctly describe what is visible in this image?*" with response space $\{\text{YES}, \text{NO}, \text{UNCLEAR}\}$, aggregated as

$$\text{choice_verdict} = \begin{cases} \text{SUPPORTED} & \text{if any } i \text{ returns YES} \\ \text{REFUTED} & \text{if all } i \text{ return NO} \\ \text{UNCLEAR} & \text{otherwise.} \end{cases} \tag{2}$$

REFUTED maps to choice_invalid. Per-option verification adds $|C|-1$ marginal VLM calls relative to a bundled claim but recovers up to $+69$ pp choice-missing TAR (§4.4).

3.7 Deterministic Aggregator (A3)

The aggregator combines the A2 verdict set V into (d, r, conf) with a deterministic programmatic rule (full pseudocode in Algorithm 1, Appendix B). Let n_S, n_R, n_U count SUPPORTED / REFUTED / UNCLEAR in V. Each $c \in V$ carries a `type` $\in$ {entity, region, choice}, a `verdict`, and a `reason_type` matching the typed reason it would emit on a REFUTED. The aggregator iterates the verdicts in priority order [entity, region, choice]; the first REFUTED verdict encountered triggers an abstention with the corresponding typed reason, except on `contain`-style yes/no questions where a legitimate "No" is preserved by a negative-answerable bypass. If no REFUTED fires, the decision defaults to ANSWER when $n_S \geq \lceil |V|/2 \rceil$ and to ABSTAIN (evidence_insufficient) otherwise.

3.8 Grounded Answer and Cost (A4)

If $d =$ ANSWER, PICV produces the final answer with one additional VLM call whose prompt contains the *grounding* from §3.5 but not the verifier outputs (mirroring radiology workflow where the structured report guides the answer rather than the verification trace). In total the pipeline issues *at most* $4 + |C|$ VLM calls beyond a direct-answer baseline: 1 decompose, 1 grounding, 1 entity verifier and 1 region probe (when an entity is presupposed by A0), $|C|$ per-option verifiers (when MCQ), and 1 grounded answer (only on the answerable path). The programmatic aggregator and the entity-region map add no VLM calls. Open-ended items, abstain decisions, and entity-skipped paths each subtract from this upper bound, so the empirical average is lower (Appendix D). PICV averages 4.3–5.0 calls/sample across the four backbones, with corrected-set pooled mean 4.65 versus 8.83 for MediQ-Scale.

Implementation. We implement the deterministic aggregator (Algorithm 1) and the programmatic entity-region lookup as a small Python module ($\sim$30 lines for the aggregator + a 62-entry JSON table for $\mathcal{M}$, both released with the paper). Python is an implementation choice rather than a methodological one: any host language with conditional logic and string containment suffices, and the entire PICV pipeline can be reproduced in any inference runtime that supports the four VLM call patterns of A0–A4.

4 Experiments

We answer three questions: **Q1** (§4.2): does PICV outperform existing inference-time methods for selective answering across deployment tiers? **Q2** (§4.3): which failure modes drive the gain? **Q3** (§4.4): what is the marginal contribution of the structural devices P1/P2/P3?

4.1 Setup

We evaluate on two public radiology VQA benchmarks (SLAKE [3] and VQA-RAD [4]), each re-cast as a balanced answerability benchmark via the deterministic corruption protocol of §3.1. For the main evaluation, A single AI rater re-verifies all synthetic inputs; we relabel inputs judged answerable and exclude uncertain inputs. Each cell then contains 94 outputs, with answerable/unanswerable counts of 69/25 on SLAKE and 58/36 on VQA-RAD (Appendix H). We com-

Table 1: AnsF1 under corrected labels. Uncertain inputs are excluded, leaving 94 outputs per cell. Best scores are bold. Class proportions and prior-balanced results are in Appendix H.

	Gemini		GPT-mini		MedGemma		Qwen	
Method	SLAKE	VQA-RAD	SLAKE	VQA-RAD	SLAKE	VQA-RAD	SLAKE	VQA-RAD
Direct	0.847	0.763	0.847	0.763	0.847	0.763	0.847	0.763
Two-stage	0.805	0.833	0.678	0.569	0.843	0.789	0.651	0.712
CoT-Abstain	0.832	0.855	0.655	0.687	0.710	0.662	0.481	0.660
ReAct-Med	0.880	0.862	0.827	0.809	0.729	0.686	0.759	0.785
Self-Ask	0.726	0.865	0.654	0.711	**0.852**	0.767	0.083	0.000
MediQ-Binary	0.767	0.811	0.717	0.750	0.707	0.775	0.740	0.685
MediQ-Scale	0.760	0.852	0.733	0.796	0.702	0.766	0.605	0.640
PICV	**0.894**	**0.930**	**0.866**	**0.862**	0.803	**0.803**	**0.882**	**0.914**

pare across four VLM backbones spanning two frontier APIs and two on-premise models, including a medical-domain-specialised VLM: this lets us examine whether the tested medical-domain-pretrained model filters unanswerable inputs reliably (it does not; §4.2). External baselines comprise five standard prompting methods (B1 Direct Answer, B2 Two-stage Verify, B3 CoT-Abstain, B4 ReAct-Med, B5 Self-Ask) and two agentic adaptations of MEDIQ [9] (B6a/B6b). The original benchmark metric is *answerability F1* (AnsF1) over the answer/abstain decision; secondary metrics, sample size, seed, exact backbone identifiers, decoding parameters, hardware, and per-cell wall-clock are in Appendix A. Source code, benchmark construction scripts, prompts for A0–A4, the entity-region table $\mathcal{M}$, and per-cell evaluation outputs are released at `https://github.com/venzino-han/selective_vqa`.

4.2 Main result: PICV vs external baselines (Q1)

Table 1 reports AnsF1 under the AI-corrected labels. PICV ranks first in seven of eight cells; the exception is MedGemma–SLAKE, where it ranks fourth behind Self-Ask, Direct, and Two-stage. Because re-verification changes class proportions, Always-Answer no longer has a fixed score across datasets. Prior-balanced AnsF1 preserves the same ranking pattern (Appendix H). Original-label scores are retained in Appendix H.1; corrected-label uncertainty analyses appear in Appendix C.

Frontier models. PICV leads on both datasets for Gemini and GPT-mini. The strongest comparators vary across cells: ReAct-Med, Self-Ask, and Direct provide the closest comparisons, with the smallest margin on Gemini–SLAKE.

General-purpose on-premise. PICV leads on both Qwen datasets, with a larger margin on VQA-RAD. Direct is the strongest comparator on SLAKE and ReAct-Med on VQA-RAD. Self-Ask almost always abstains, illustrating that excessive refusal can also undermine selective answering.

Domain-specialised medical VLM. The MedGemma results are mixed: PICV leads on VQA-RAD but falls below Always-Answer on SLAKE. CoT-Abstain also falls below Always-Answer on both datasets. Medical-domain pretraining therefore does not ensure reliable selective answering in this evaluated backbone. Backbone scale and domain specialization are confounded; separating them requires size-matched comparisons.

4.3 Per-corruption analysis (Q2)

Table 2 reports TAR only on inputs that remain unanswerable after re-verification by the single AI rater. Choice-removal detection is strong for Gemini, GPT-mini, and Qwen, but weaker for MedGemma.

Re-verification reveals different label effects across corruption types. PICV answers nearly all image mismatches relabeled answerable, but rejects most SLAKE question corruptions relabeled answerable. Thus, the original labels penalized many decisions to answer swapped images while rewarding refusals of valid yes/no questions. Appendix H reports the answerability counts.

Table 2: PICV TAR under corrected labels: rate (abstained/unanswerable). Relabeled-answerable and uncertain inputs are excluded. SLAKE retains three unanswerable cases each for image mismatch and question corruption.

Backbone	Dataset	Choice removal	Image mismatch	Question corruption
Gemini	SLAKE	1.00 (19/19)	0.33 (1/3)	0.33 (1/3)
Gemini	VQA-RAD	0.97 (28/29)	0.71 (5/7)	—
GPT-mini	SLAKE	0.84 (16/19)	0.33 (1/3)	0.33 (1/3)
GPT-mini	VQA-RAD	0.86 (25/29)	0.43 (3/7)	—
MedGemma	SLAKE	0.21 (4/19)	0.00 (0/3)	0.67 (2/3)
MedGemma	VQA-RAD	0.45 (13/29)	0.71 (5/7)	—
Qwen	SLAKE	0.89 (17/19)	0.00 (0/3)	0.33 (1/3)
Qwen	VQA-RAD	0.93 (27/29)	0.57 (4/7)	—

4.4 Ablation: P1 + P2 + P3 (Q3)

The corrected-label component comparison is in Appendix J. We compare **PICV-base** (bundled choice verification, one-shot region verification, no decomposition sanity check) with **PICV-full** (P1–P3 enabled). Their combined effect improves mean AnsF1, with the largest gain on Gemini. Choice-removal TAR improves in most cells, but declines on MedGemma–SLAKE. Among MedGemma decision wins over baseline methods, choice-removal cases account for the largest share. Question-corruption wins decrease after relabeling because many original refusals were on answerable yes/no questions (Appendix G).

4.5 Final-answer performance

These analyses use AI-corrected labels. Because relabeled-answerable inputs lack verified reference answers, correctness-dependent results are evaluated under pessimistic and optimistic assignments (Appendix I.1). AnsF1 evaluates the answer/abstain decision rather than the correctness of the answer. To assess actual answer performance, we compare PICV and Direct on the same answerable cases retained by PICV, using SoftAcc (normalized string-containment answer matching). Their pooled scores are close under both assignments, although the direction of the difference varies across backbones and datasets. When abstentions are counted as errors over all answerable inputs, PICV scores lower than Direct in both scenarios: rejecting unsupported queries also means withholding some correct answers. The paired results and the full-denominator comparison appear in Appendix I.1.

4.6 Coverage, harmful answers, and inference cost

We also count *harmful answers*: wrong answers on answerable inputs or any answer on assigned-unanswerable inputs. PICV produces fewer harmful answers than Direct while retaining most answerable inputs. At PICV's coverage, it produces fewer harmful answers than Direct, Two-stage, and ReAct-Med in both scenarios. These comparisons retrospectively withhold low-confidence answers that each method already emitted, averaging over ties and preserving its original abstentions. PICV uses fewer model calls than MediQ-Scale but more than short-pass methods. Detailed operating points and costs are reported in Appendix I.

4.7 Fixed-evidence aggregation control

To isolate the aggregation decision, we hold A0–A2 evidence fixed, replace A3 with a same-backbone judge on MedGemma and Qwen, and reuse stored A4 answers. Under both scenarios at matched coverage, the deterministic rule achieves higher AnsF1 and fewer harmful answers than the judge. At equal harm, it retains more queries. The judge's native operating point is more conservative, answering fewer queries and producing fewer harmful answers. Full operating-point comparisons appear in Table 12.

5 Conclusion

Medical visual question answering is a safety-critical decision problem where models must answer correctly when visual evidence is sufficient and abstain when it is not. Existing inference-time

agentic methods attempt to reduce hallucinations through verification, but typically route the final decision through a single language-model judge that integrates all evidence into a single prompt, making the system vulnerable to answerability bias. We propose **PICV** (Parallel Independent Claim Verification), a selective medical VQA framework that reformulates answerability assessment as independent claim verification with transparent aggregation. Given a question, PICV decomposes it into a compact set of typed visual claims, verifies each claim in an isolated prompt without access to other verdicts, and combines the resulting verdicts using a deterministic rule rather than an additional model. This design produces both an answer/abstain decision and a typed unanswerability reason, improving interpretability while reducing judge-induced bias. We evaluate PICV on an automated answerability benchmark constructed from two public radiology datasets using three reproducible corruption procedures. Additional controls examine the limitations of the assigned labels. Across four VLMs, PICV leads answerability F1 in seven of eight cells under labels re-verified by the single AI rater, with MedGemma–SLAKE as the exception. Additional fixed-evidence controls support deterministic aggregation over the evaluated same-backbone judge. Joint analyses of answer quality and coverage show the cost of abstention. Additional controls examine visual degradation and open-ended questions in the evaluated English radiology datasets.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No.RS-2022-NR068758) and the Seegene Medical Foundation grant (No. G01260468).

References

[1] Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-RAD: A comprehensive 3d radiology med-VQA dataset with multi-temporal analysis and diverse diagnostic tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL `https://openreview.net/forum?id=VB2cgrlikN`.

[2] Qingqiu Li, Zihang Cui, Seongsu Bae, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, Shang Gao, Junjun He, and Shujun Wang. AOR: Anatomical ontology-guided reasoning for medical large multimodal model in chest x-ray interpretation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=RDdfQc5Ts1`.

[3] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[4] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

[5] Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. MIRAGE: The illusion of visual understanding, 2026.

[6] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

[7] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378.

[8] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.212.

[9] Shuyue S. Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 28858–28888, 2024.

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

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

[12] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[13] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305, 2023.

[14] Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, 2025.

[15] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[16] Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

[17] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4885–4894, 2017.

[18] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer, 2022.

[19] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Samuel R. Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

[20] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[21] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR, 2024.

[22] Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate–we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*, 2025.

A Prompts, models, and reproducibility

Sample size and seed. For each (backbone, dataset) cell we draw 50 answerable seeds from the SLAKE (English) or VQA-RAD test split and pair them with 50 deterministically corrupted variants. Per backbone, SLAKE has 14 `image_mismatch`, 13 `question_corruption`, and 23 `choice_missing` outputs; VQA-RAD has 19 image mismatches and 31 choice-removal outputs. The answerability classes are balanced, but corruption axes are not. The same variant counts apply to all four backbones. The corruption protocol is a deterministic function of the random seed; we use seed = 42 in all main experiments.

Backbones. We use the following exact backbone identifiers:

- **Gemini-3.1-Flash-Lite-Preview** via the Google API (frontier).
- **GPT-5.4-mini** via the OpenAI API (frontier).
- **MedGemma-4B** [10], on-premise medical-specialised VLM (4B parameters), served via vLLM 0.17.1 in `bfloat16` on a single GPU.
- **Qwen2.5-VL-7B-Instruct**, on-premise general-purpose VLM (7B parameters), served identically.

Decoding. All inference runs at `temperature = 0.0` (deterministic). All on-premise inference uses vLLM 0.17.1 in `bfloat16`. Frontier API snapshots are recorded in the released `runs/` metadata.

Baselines. B6a (MediQ-VLM Binary) wraps a binary sufficiency check with at most 3 turns; B6b (MediQ-VLM Scale) replaces the binary check with a 5-point Likert sufficiency score (threshold ≥ 4 to abstain). The five prompting baselines (B1–B5) follow the original prompt specifications in their respective references; full prompt text is in the online repository.

Metrics. One reported metric is **answerability F1 (AnsF1)**: binary F1 with positive class "answerable" computed over the ANSWER/ABSTAIN decision. Secondary metrics are: **TAR** (true-abstain rate, fraction of unanswerable inputs correctly abstained); **FAR** (false-abstain rate, fraction of answerable inputs incorrectly abstained); **ECE** (expected calibration error over all outputs, with 10 equal-width bins; abstentions and unanswerable inputs count as incorrect); and **calls/sample** (mean VLM calls per evaluated instance).

Hardware and compute. On-premise inference (MedGemma-4B, Qwen2.5-VL-7B-Instruct) runs on a **single NVIDIA RTX A6000 (48 GB VRAM)** via vLLM 0.17.1; we use `bfloat16` weights and the default eager-attention backend. Per-cell wall-clock at this scale is approximately: 5–10 min for B1 Direct Answer (1 call/sample), 10–15 min for B3 / B4 / B5 (1–2 calls/sample), 25–35 min for B6a/B6b MEDIQ-VLM (~ 9 calls/sample), and 20–30 min for PICV (4.3–5.0 calls/sample, 100 samples). Frontier-API runs (Gemini, GPT) are dominated by network latency rather than compute. Across all four backbones and both datasets, the full final benchmark required ~ 8 h of single-GPU on-premise compute plus ~ 47 k frontier-API calls (Gemini and GPT-mini combined). Preliminary method variants that preceded PICV and do not appear in the final tables consumed an additional ~ 20 h of GPU compute and ~ 30 k API calls.

Reproducibility statement. Reproducing the main results requires (i) the SLAKE English subset and VQA-RAD test split (downloaded by helper scripts at first run), (ii) the deterministic corruption code (seed 42), (iii) the prompt set, the 62-entry $\mathcal{M}$, and the deterministic Python aggregator, and (iv) one of the four VLM backbones (or any frozen VLM with comparable chat-template support) plus vLLM 0.17.1, transformers, openai, and google-generativeai Python packages (versions pinned in the released `requirements.txt`). All design choices (P3 sanity-check rules, aggregator priority order, $\mathcal{M}$ entries) were inspected on the SLAKE *train* split and frozen for evaluation.

Datasets and licenses. **SLAKE** [3] is used under CC-BY-SA 4.0; **VQA-RAD** [4] is publicly distributed via the 2018 *Scientific Data* release (CC0 / public domain). Both datasets contain only de-identified medical images and are used solely for research evaluation. Backbone licenses:

Algorithm 1 PICV deterministic aggregator (A3).

Require: Verdict set V; flags `QUESTION_TYPE`, `NEGATIVE_ANSWERABLE` from A0.
Ensure: Decision d, typed reason r, confidence conf.

$n_S, n_R, n_U \leftarrow$ counts of SUPPORTED, REFUTED, UNCLEAR in V
if $V = \emptyset$ **then**
  return (ABSTAIN, $\cdot$, 0.5) ▷ nothing to verify ⇒ default answer
end if
for $c \in V$ in priority order [entity, region, choice] **do**
  if c.`verdict` = REFUTED **then**
    if c.`type` $\in$ {entity, region} **and** `QUESTION_TYPE` = YES_NO **and** `NEGATIVE_ANSWERABLE` = YES **then**
      continue ▷ legitimate "No" answer; bypass
    end if
    return (ABSTAIN, c.`reason_type`, $\max(\epsilon_{\min}, 1 - n_S/|V|)$) ▷ $\epsilon_{\min} = 0.05$ avoids zero confidence
  end if
end for
if $n_S \geq \lceil |V|/2 \rceil$ **then**
  return (ANSWER, $\cdot$, $n_S/|V|$)
end if
if $n_U = |V|$ **then**
  return (ABSTAIN, evidence_insufficient, 0.25) ▷ distinct fixed confidence on the all-UNCLEAR branch
end if
return (ABSTAIN, evidence_insufficient, $n_S/|V|$)

MedGemma-4B [10] (Health AI Developer Foundation license; research use), Qwen2.5-VL-7B-Instruct (Apache 2.0), Gemini and GPT-mini accessed via the respective Google and OpenAI public APIs under their standard terms. We introduce no new patient data and no new model weights.

Prompts. The full prompt set for each PICV stage (A0 decomposition, A1 question-blind grounding, A2 entity / region step-1 / per-option verifiers, and A4 grounded answer), together with the baseline prompts (B1 Direct Answer, B2 Two-stage Verify, B3 CoT-Abstain, B4 ReAct-Med, B5 Self-Ask, B6a/B6b MediQ-VLM), is released through the public repository at `https://github.com/venzino-han/selective_vqa`.

Entity-region table $\mathcal{M}$. The 62-entry $\mathcal{M}$ used in the main experiments is released as `src/picv/entity_region.json` in the public repository.

B Deterministic aggregator pseudocode

Algorithm 1 gives the full pseudocode for the deterministic programmatic aggregator described in §3.7. The aggregator takes the A2 verdict set V and the A0 flags `QUESTION_TYPE`, `NEGATIVE_ANSWERABLE`, and emits the decision d, typed reason r, and confidence $\mathrm{conf} = n_S/|V|$.

C Statistical uncertainty

This section reports uncertainty under the AI-corrected labels of the main tables. Cell-level rankings are descriptive. The eight cells reuse source questions, so we do not use cross-cell sign or Wilcoxon tests as independent confirmatory evidence. Table 3 reproduces paired, gold-class-stratified bootstrap intervals (10,000 resamples) within each 94-output cell. Comparisons with the strongest baseline are exploratory because that comparator was selected on the same evaluation data. Source-family dependence within cells is not corrected by this procedure; the intervals should not be read as source-clustered inference. Pooled additional analyses are reported as point estimates rather than statistically significant cross-backbone improvements. Source-question-clustered uncertainty remains a limitation of these saved analyses.

Table 3: Raw AnsF1 differences and exploratory 95% paired, gold-class-stratified bootstrap intervals (10,000 resamples; 94 outputs per cell). Both differences and intervals use raw AnsF1. These intervals do not correct for source-family dependence.

Backbone	Dataset	Δ vs. Direct [CI]	Best baseline	Δ vs. best [CI]
Gemini	SLAKE	+0.047 [-0.009, 0.096]	ReAct-Med	+0.014 [-0.055, 0.078]
Gemini	VQA-RAD	+0.167 [0.114, 0.211]	Self-Ask	+0.064 [-0.021, 0.153]
GPT-mini	SLAKE	+0.019 [-0.044, 0.073]	Direct	+0.019 [-0.044, 0.073]
GPT-mini	VQA-RAD	+0.099 [0.033, 0.159]	ReAct-Med	+0.053 [-0.024, 0.127]
MedGemma	SLAKE	-0.044 [-0.098, 0.007]	Self-Ask	-0.049 [-0.104, 0.001]
MedGemma	VQA-RAD	+0.040 [-0.023, 0.096]	Two-stage	+0.014 [-0.050, 0.074]
Qwen	SLAKE	+0.036 [-0.020, 0.088]	Direct	+0.036 [-0.021, 0.087]
Qwen	VQA-RAD	+0.151 [0.096, 0.196]	ReAct-Med	+0.129 [0.061, 0.196]

D Calibration and cost details

Under AI-corrected labels, calibration and false-abstention metrics complement the joint evaluation in the main text. Average model-call counts are reported in Table 9. PICV has higher ECE than the minimum-ECE qualifying baseline in all eight cells under pessimistic assignments and seven under optimistic assignments. Thus, better answerability F1 does not imply better calibration. ECE is evaluated separately in each scenario and need not vary monotonically between them. ECE uses all retained outputs and ten equal-width bins; abstentions and unanswerable inputs count as incorrect. The comparator is the minimum-ECE external baseline satisfying TAR $\geq 5\%$ and FAR $\leq 90\%$, selected separately in each scenario. The sole optimistic exception is Gemini–SLAKE, where PICV's ECE is lower by 0.002. Detailed per-cell results are retained in the accompanying CSV evidence. On Qwen–VQA-RAD, Self-Ask abstains on all inputs (TAR and FAR both 1.00), illustrating the excessive-refusal behavior discussed in the main text.

E Limitations

Unanswerability is synthetic by construction; clinical realism requires expert validation. All synthetic labels were reassessed by a single AI rater, without a clinician audit (Appendix H). The corrected labels therefore reflect model judgments rather than clinically validated answerability; errors in these judgments may affect the reported comparisons. P2's contribution is dataset-conditional (§4.3). The entity-region table is a small, manually engineered external KB, fixed on the training split. A leakage-controlled end-to-end selective predictor is not evaluated. Evaluation is restricted to English radiology; pathology, fundus, and other modalities, as well as size/domain disentanglement of medical-specialised backbones, are out of scope.

Positioning relative to prompt-only and RAG agents. PICV occupies an intermediate position on the external-knowledge spectrum. Pure prompt-only agents (CoVe [8], Self-Ask [7], ReAct [6]) use no external table or retrieval index; retrieval-augmented inference queries indices of 10^4–10^6 documents at run time. PICV uses a static, hand-curated table of 62 entity-to-region pairs (four orders of magnitude smaller than a typical RAG index) consulted through a Python lookup at the deterministic step of P2; no documents are retrieved, no VLM call is added, and no embedding index is maintained. We therefore do not claim "no-tuning" in the strictest sense in which prompt-only agents do.

F Broader impacts

PICV's research goal is to make medical VLMs safer to deploy near patient care by providing an explicit abstention decision with typed reasons. The intended effect is positive: a system that refuses unanswerable cases with a clinically actionable label (e.g. image_mismatch "image shows upper abdomen, not where heart resides" on a chest-question $\rightarrow$ abdomen-CT swap) is strictly more useful than one that confidently emits a wrong answer in the same situation, because the typed reason maps to a downstream action (re-route to the correct study, escalate to free-text answering, request addi-

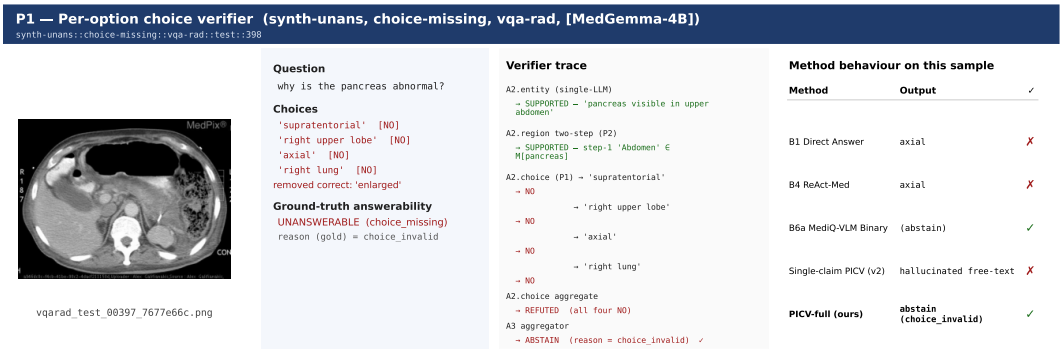

Figure 3: P1 case: `vqa-rad::test::398`, MedGemma. The image and question are mutually consistent; the failure is purely at the choice level (`enlarged` removed). Per-option verification returns NO on all four options; the aggregator deterministically maps the unanimous NO to choice_invalid.

tional imaging). We anticipate the most directly positive impact in clinical triage, second-opinion tools, and educational settings where hallucinated medical advice is unsafe.

We also anticipate several risks and limitations that require explicit mitigation before any clinical deployment. First, *over-abstention*: PICV's aggregator can refuse to answer when verifier confidence is low, and a deployment that uses PICV's ABSTAIN as a hard gate will increase patient-care latency on the false-abstain set; PICV's FAR is in the 0.06–0.32 range, so downstream operators can size this risk. Second, *distribution shift*: the deterministic corruption protocol of §3.1 is reproducible but synthetic, so deployment-time performance against image-quality artefacts, pathology-specific edge cases, or modality-mixing failures is not characterised. Third, *population coverage*: the benchmark uses English-only SLAKE and VQA-RAD with predominantly Western-medicine training distributions; PathVQA, MedMCQA, fundus, and non-radiology modalities are explicitly out of scope, and underrepresented populations may be served less reliably until PICV is re-validated in those settings. Fourth, *frontier-API privacy*: sending patient images to third-party APIs (Gemini, GPT) raises HIPAA-style data-processing concerns; on-premise backbones (MedGemma-4B, Qwen2.5-VL-7B) are the appropriate choice when patient data cannot leave the institutional boundary, and we report results on both tiers. Fifth, *not a clinical decision tool*: PICV is an inference-time architecture, not a clinically validated system. Any clinical use requires (i) prospective clinician-curated validation, (ii) regulatory review where applicable (e.g. FDA / CE), and (iii) clinician-in-the-loop oversight. We do not recommend PICV's ANSWER verdict as a basis for clinical action without such oversight.

The released artefacts (deterministic corruption code, prompts, $\mathcal{M}$, aggregator) are research evaluation tools rather than generative-model weights, so dual-use risk from the release itself is low. The README explicitly states the no-clinical-deployment-without- oversight constraint above.

G Case study figures and component attribution

Across both MedGemma datasets and seven baseline comparisons, PICV has 209 correct answer/abstain decisions on which the comparator is wrong: 98 choice-removal cases (47%), 63 original answerable cases (30%), 23 relabeled-answerable cases (11%), 19 remaining image mismatches (9%), and six remaining question corruptions (3%). These are comparison-level wins, not unique questions or isolated component effects. The question-corruption count falls from 33 under the original labels to six after re-verification, because many original refusals were on answerable inputs.

Each card shows three regions: the top region holds the actual benchmark image and the question/choices/ground-truth panel; the middle region holds the full PICV verifier trace; the bottom region holds a per-baseline comparison table on the same sample. Sample IDs are reproducible from the benchmark seed (seed 42). The cards are auto-generated by `scripts/build_case_figures.py` from raw inference traces under `outputs/exp100_m7_picv_*_v3/`.

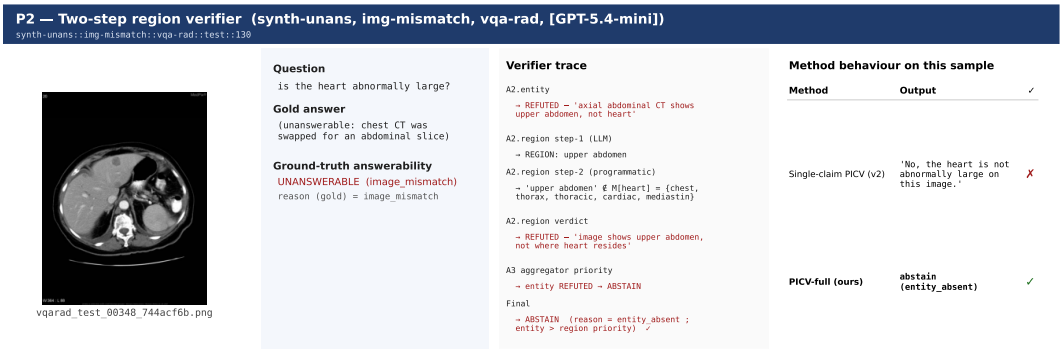

Figure 4: P2 case: `vqa-rad::test::130`, GPT-5.4-mini. The image (abdominal CT) does not depict the heart. Both the entity verifier and the two-step P2 region verifier independently refute their respective claim: P2 first commits to "upper abdomen," then the deterministic check against $\mathcal{M}(\texttt{heart})$ confirms the mismatch. Because the deterministic aggregator applies the priority entity $>$ region $>$ choice, the final typed reason emitted on this case is entity_absent; the figure illustrates that P2's region verdict converges to the same unanswerable conclusion via an independent evidence chain.

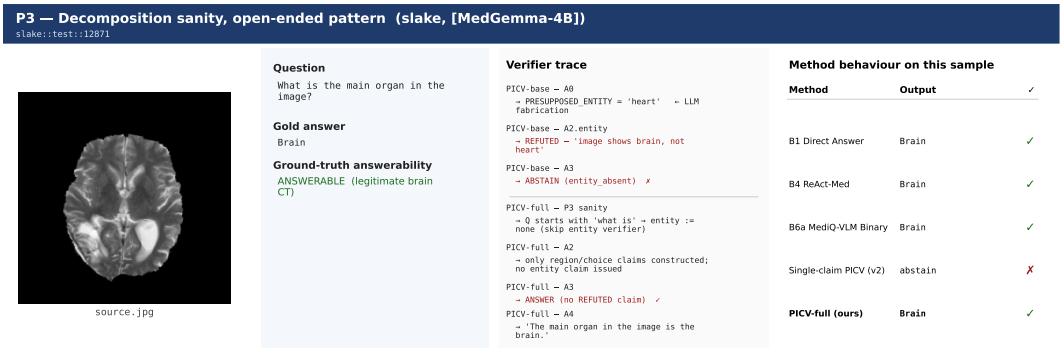

Figure 5: P3 (open-ended) case: `slake::test::12871`, MedGemma. Without P3 the decomposition step fabricates `PRESUPPOSED_ENTITY = heart`, which is then refuted by the entity verifier, producing a wrongful abstention. P3 detects the "what is" prefix and skips the entity verifier.

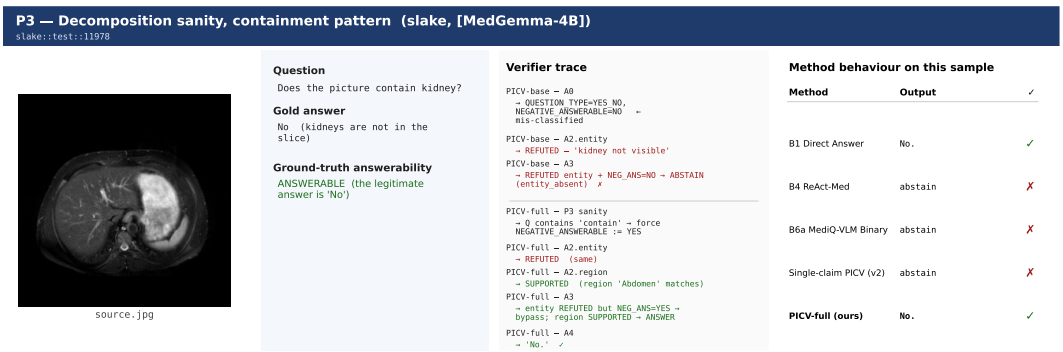

Figure 6: P3 (containment) case: `slake::test::11978`, MedGemma. The legitimate answer is "No." Every decomposition-based agentic baseline mis-treats this as an abstention case; P3's two-line rule and the aggregator's bypass loop together produce the correct "No."

H AI re-verification of synthetic labels

Claude Opus 5.5 served as a single AI rater for all 100 unique synthetic inputs shared across the four backbones (50 per dataset). These are model judgments, not clinician-adjudicated labels. The rater received the image, question, and options when present; the evaluated-item judgments were saved before the original answers were revealed. No new inference by the evaluated methods was run. Of the synthetic inputs, 61 remained unanswerable, 27 were relabeled answerable, and 12 were uncertain. The corrected setting retains relabeled-answerable inputs and drops uncertain inputs. With the original answerable seeds retained, each cell has 94 outputs: 69 answerable and 25 unanswerable on SLAKE, and 58 and 36 on VQA-RAD. Raw AnsF1 therefore has a different class prior from the original balanced benchmark. Always-Answer scores 0.847 and 0.763, respectively.

Table 4: Synthetic-label review. Counts are unique inputs, not repeated backbone outputs.

Dataset	Axis	Total	Unanswerable	Answerable	Uncertain
SLAKE	Choice removal	23	19	1	3
SLAKE	Image mismatch	14	3	9	2
SLAKE	Question corruption	13	3	9	1
VQA-RAD	Choice removal	31	29	1	1
VQA-RAD	Image mismatch	19	7	7	5
Total		100	61	27	12

Donor images can still support the question. Question-corruption templates are yes/no questions, and replacing a presupposed entity can leave a valid "No" answer. Among choice removals, one SLAKE input retained a whitespace variant of the gold answer ("Transverse Plane"), while one VQA-RAD input retained an option the rater judged substantially correct.

PICV ranks first in raw and prior-balanced AnsF1 in seven of eight cells. The exception is MedGemma–SLAKE, where it ranks fourth, below Self-Ask, Direct, and Two-stage. Prior-balanced AnsF1 gives equal total weight to each gold class, separating the ranking comparison from the changed class proportions. All rankings are descriptive; model-based labels are not an independent clinical reference standard. On SLAKE question corruptions relabeled answerable, PICV

Table 5: Prior-balanced AnsF1 under corrected labels. Each gold class receives equal total weight; Always-Answer scores 0.667.

	Gemini		GPT-mini		MedGemma		Qwen	
Method	SLAKE	VQA-RAD	SLAKE	VQA-RAD	SLAKE	VQA-RAD	SLAKE	VQA-RAD
Direct	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667
Two-stage	0.651	0.766	0.598	0.522	0.672	0.699	0.532	0.651
CoT-Abstain	0.729	0.803	0.590	0.658	0.566	0.579	0.411	0.620
ReAct-Med	0.748	0.801	0.676	0.740	0.619	0.600	0.630	0.714
Self-Ask	0.693	0.860	0.643	0.711	**0.676**	0.677	0.083	0.000
MediQ-Binary	0.714	0.789	0.641	0.720	0.577	0.705	0.642	0.638
MediQ-Scale	0.699	0.833	0.665	0.766	0.578	0.696	0.501	0.603
PICV	**0.849**	**0.915**	**0.793**	**0.827**	0.654	**0.739**	**0.809**	**0.890**

Table 6: PICV answer decisions on inputs relabeled answerable (answered/total). This measures decisions to answer, not answer correctness.

Backbone	Dataset	Choice removal	Image mismatch	Question corruption
Gemini	SLAKE	1/1	9/9	3/9
Gemini	VQA-RAD	1/1	6/7	—
GPT-mini	SLAKE	1/1	9/9	2/9
GPT-mini	VQA-RAD	1/1	6/7	—
MedGemma	SLAKE	1/1	9/9	2/9
MedGemma	VQA-RAD	1/1	6/7	—
Qwen	SLAKE	1/1	9/9	3/9
Qwen	VQA-RAD	0/1	6/7	—

abstains on six or seven of nine inputs, depending on the backbone. The original labels rewarded

Table 7: Original construction labels. AnsF1 across four backbones and two datasets, recomputed from the saved 100-output cells. Best point estimate per cell is bold; rankings are descriptive. Corrected-label coverage–harm outcomes and calls are reported separately in Table 9.

	Gemini		GPT-mini		MedGemma		Qwen	
Method	SLAKE	VQA-RAD	SLAKE	VQA-RAD	SLAKE	VQA-RAD	SLAKE	VQA-RAD
Direct	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667
Two-stage	0.622	0.734	0.535	0.511	0.671	0.690	0.523	0.637
CoT-Abstain	0.667	0.783	0.646	0.660	0.624	0.585	0.488	0.612
ReAct-Med	0.715	0.772	0.696	0.719	0.638	0.627	0.715	0.736
Self-Ask	0.688	0.796	0.644	0.659	0.671	0.667	0.113	0.000
MediQ-Binary	0.686	0.752	0.667	0.700	0.639	0.698	0.709	0.654
MediQ-Scale	0.679	0.796	0.712	**0.777**	0.650	0.688	0.552	0.598
PICV	**0.793**	**0.829**	**0.780**	0.775	**0.701**	**0.704**	**0.783**	**0.832**

these refusals. In contrast, PICV answers nearly all image mismatches relabeled answerable, which the original labels penalized. Verified reference answers for these relabeled inputs are unavailable, so SoftAcc and harmful-answer metrics are reported under pessimistic and optimistic correctness assignments (Appendix I.1).

H.1 Original construction-label comparisons

Table 7 preserves the original balanced 100-output cells. Under these labels, PICV leads in seven cells, with GPT-mini–VQA-RAD as the exception. These results must not be interpreted as scores under the AI-corrected labels. The original per-corruption rates are retained in the accompanying evidence files.

I Additional answer-quality and selection analyses

I.1 Final-answer performance

Unless stated otherwise, analyses use the AI-corrected labels of Appendix H: 752 outputs in eight 94-output cells, with repeated source questions across backbones. The 108 outputs on inputs relabeled answerable have no verified reference answer. We report pessimistic / optimistic scenarios: an emitted answer to such an input is counted as wrong / correct. SoftAcc and harmful-answer counts are bounded by these assignments; ECE and harm-budget operating points are scenario evaluations, not confidence intervals. Answerability metrics at a fixed operating point do not depend on the assignment. Original-label results remain in the evidence files.

The paired comparison restricts Direct to the same 443 answerable outputs retained by PICV. SoftAcc is normalized string-containment matching, which can miss synonyms or accept overly permissive matches. It ranges from 0.490 to 0.664 for PICV and 0.494 to 0.668 for Direct; 77 retained outputs lack verified reference answers. The paired difference is -0.0045 (exploratory 95% CI $[-0.032, 0.023]$) in both scenarios because the assignments shift both methods equally. This does not establish non-inferiority, and unknown correctness need not agree across methods. Counting abstentions as errors over all 508 answerable outputs gives $217/508 = 0.427$ to $294/508 = 0.579$ for PICV and $245/508 = 0.482$ to $353/508 = 0.695$ for Direct. PICV retains 87.2% of answerable inputs and 68.3% of all inputs, with 39.49 / 29.26 harmful answers and 8.64 false refusals per 100 queries. Direct produces 67.42 / 53.06 harmful answers per 100 at full coverage. At PICV's coverage, Direct, Two-stage, and ReAct-Med produce 46.17 / 36.36, 42.25 / 31.29, and 42.24 / 33.11 harmful answers per 100, respectively. These comparisons hold within each scenario, not across arbitrary assignments of unknown answer correctness. PICV averages 4.65 calls versus 8.83 for MediQ-Scale.

I.2 Benchmark alignment and validity controls

We degrade 40 images (20 per dataset) and pair them with their originals, keeping the questions and choices unchanged. On these 80 inputs, evaluated with Gemini, PICV/Direct/ReAct have AnsF1 0.710/0.667/0.693 and AURC 0.677/0.800/0.717. PICV nevertheless has more native harmful an-

Table 8: Same PICV-retained answerable cases. Paired entries give pessimistic / optimistic scenarios for outputs with no verified reference answer. n counts backbone–question outputs, not independent source questions.

Backbone	Dataset	n	Direct SoftAcc	PICV SoftAcc
Gemini	SLAKE	59	0.424 / 0.644	0.407 / 0.627
Gemini	VQA-RAD	53	0.585 / 0.717	0.585 / 0.717
GPT-mini	SLAKE	58	0.448 / 0.655	0.517 / 0.724
GPT-mini	VQA-RAD	50	0.560 / 0.700	0.600 / 0.740
MedGemma	SLAKE	59	0.458 / 0.661	0.475 / 0.678
MedGemma	VQA-RAD	51	0.529 / 0.667	0.510 / 0.647
Qwen	SLAKE	60	0.450 / 0.667	0.367 / 0.583
Qwen	VQA-RAD	53	0.528 / 0.642	0.491 / 0.604
Pooled	All 8	443	0.494 / 0.668	0.490 / 0.664

Table 9: Joint evaluation on 752 outputs. Paired entries give pessimistic / optimistic scenarios for outputs with no verified reference answer. Cov. is native coverage; H/100 is harmful answers and FR/100 is false refusals per 100 queries. C_B is coverage under PICV's native harm budget; H_C is harm at PICV's coverage. A dash denotes unreachable coverage. Starred calls average available records only.

Method	Cov.	H/100	FR/100	C_B	H_C	Calls
Direct	1.000	67.42 / 53.06	0.00	0.584 / 0.549	46.17 / 36.36	1.00*
Two-stage	0.722	44.28 / 32.58	15.43	0.630 / 0.622	42.25 / 31.29	1.33*
CoT-Abstain	0.589	37.63 / 30.32	23.27	0.589 / 0.572	—	1.00
ReAct-Med	0.779	48.80 / 38.96	9.84	0.638 / 0.604	42.24 / 33.11	2.00
Self-Ask	0.454	34.04 / 27.26	30.32	0.454 / 0.454	—	1.75
MediQ-Binary	0.584	36.44 / 29.39	20.74	0.584 / 0.581	—	8.66*
MediQ-Scale	0.568	34.04 / 27.53	22.07	0.568 / 0.568	—	8.83
PICV	0.683	39.49 / 29.26	8.64	0.683 / 0.683	39.49 / 29.26	4.65
Always-Abstain	0	0	67.55	0	—	0

swers than ReAct (63.75 vs. 55.00/100). This is a small synthetic stress test, not naturally occurring clinical ambiguity.

On 464 open-ended outputs, coverage under PICV's native harm budget is 0.681 for PICV, 0.619 / 0.582 for Direct, and 0.569 for MediQ-Scale; Two-stage reaches 0.700 / 0.696. PICV's open-ended AURC is worse than MediQ-Scale in both scenarios (0.806 / 0.667 versus 0.670 / 0.594), limiting claims of ranking superiority.

Synthetic-label validity. The full review by a single AI rater covers image mismatch, question corruption, and choice removal (Appendix H). It supersedes the image-mismatch-only ranking comparison. The corrected main results retain inputs judged answerable and exclude uncertain inputs.

I.3 Matched operating points and confidence ranking

Native overall coverage is the number of emitted answers divided by all evaluated queries. For each method we retain prefixes of its emitted answers sorted by saved confidence. Within a tied confidence group, expected harmful counts are computed by uniform random ordering, yielding a piecewise-linear expected prefix curve. The harm-budget comparison uses the largest integer prefix within the budget; a system already within the budget remains at its native answer count. Thus "under the same budget" need not mean equal realized harm. Native abstentions are not reversed, and no new answer is generated. The corresponding threshold is chosen retrospectively with test

Table 10: Separate visual-degradation benchmark (unchanged labels): 40 source pairs, Gemini only. C_B is coverage under PICV's native harm budget. Harm and coverage trade off even outside the original corruption taxonomy. AURC uses the common coverage interval.

Method	AnsF1	Coverage	Harmful/100	C_B	AURC
Direct	0.667	1.000	80.00	0.787	0.800
ReAct-Med	0.693	0.762	55.00	0.762	0.717
PICV	0.710	0.838	63.75	0.838	0.677

labels, so the curves are an evaluation device, not a validated deployment policy. Calibration would require separate data. Harm at a requested coverage is undefined when the method emitted too few answers.

AURC is the discrete mean of harmful answers divided by retained answers over the common reachable prefix range for the methods in the comparison. It differs from harmful/100, whose denominator is all queries. The saved tables report AURC over this common range rather than extrapolating to full coverage. AURC and the native harm-budget comparison can favor different methods; both are reported here.

Table 11: Paired entries give pessimistic / optimistic scenarios for outputs with no verified reference answer. C is native coverage, H is harmful/100, and C_B is coverage under PICV's native harm budget. Subset sizes exclude uncertain inputs. These are descriptive point estimates.

Subset (N)	Method	C	H	C_B	AURC
Open-ended (464)	Direct	1.000	78.88 / 69.40	0.619 / 0.582	0.790 / 0.695
	Two-stage	0.700	48.49 / 40.73	0.700 / 0.696	0.700 / 0.597
	CoT-Abstain	0.593	43.53 / 38.36	0.593 / 0.593	0.735 / 0.647
	ReAct-Med	0.791	59.27 / 51.94	0.651 / 0.621	0.750 / 0.651
	Self-Ask	0.429	35.13 / 30.82	0.429 / 0.429	0.703 / 0.603
	MediQ-Binary	0.567	42.46 / 37.50	0.567 / 0.567	0.722 / 0.631
	MediQ-Scale	0.569	41.59 / 37.07	0.569 / 0.569	0.670 / 0.594
	PICV	0.681	48.92 / 40.52	0.681 / 0.681	0.806 / 0.667
No image mismatch (648)	Direct	1.000	62.19 / 55.40	0.534 / 0.552	0.624 / 0.556
	Two-stage	0.744	41.98 / 36.11	0.571 / 0.602	0.583 / 0.510
	CoT-Abstain	0.624	37.65 / 34.72	0.554 / 0.554	0.601 / 0.554
	ReAct-Med	0.809	47.07 / 42.90	0.583 / 0.593	0.571 / 0.517
	Self-Ask	0.461	33.02 / 30.56	0.461 / 0.461	0.532 / 0.488
	MediQ-Binary	0.599	34.41 / 32.25	0.583 / 0.576	0.544 / 0.497
	MediQ-Scale	0.585	32.10 / 30.56	0.585 / 0.585	0.505 / 0.465
	PICV	0.668	33.33 / 30.71	0.668 / 0.668	0.542 / 0.486

On open-ended questions, PICV's AURC is worse than several comparators, including Two-stage, Self-Ask, and MediQ-Scale. This limits claims of general selective-ranking superiority. A separate external 3B judge answers only 39/752 queries; at that coverage it has 19 / 13 harmful answers versus 24.7 / 15.5 expected harmful answers for the rule's highest-confidence 39 outputs. We therefore do not use that comparator to claim that deterministic aggregation is superior at every operating point.

I.4 Visual-degradation construction

The stress test contains 20 source pairs from each dataset (80 inputs total), selected from open-ended questions requiring visual details. Modality, plane, and coarse body-region questions are excluded. Degradation downsamples the short side to 12 pixels, applies Gaussian blur with radius equal to the larger image dimension divided by 64, compresses contrast to 0.35, and uses JPEG quality 12. Questions, choices, and image identity are unchanged. A non-clinical visual review of montages does not establish that every degraded input is unanswerable. Only Gemini-3.1-Flash-Lite-Preview, Direct, ReAct-Med, and PICV are evaluated here. All 240 method–input outputs completed without errors after retries. PICV's pooled AnsF1 is highest, but the study is too small to support broad clinical generalization. In particular, native harmful/100 is higher than ReAct, and the dataset-specific patterns differ. Pair success means answering an original input and abstaining on its

degraded partner. On SLAKE, PICV's pair success is 0.100 versus 0.450 for ReAct; on VQA-RAD, it is 0.350 versus 0.150. Each dataset has 20 pairs, so these differences are descriptive rather than evidence of consistent superiority.

I.5 Fixed-evidence aggregation control

The same-backbone judge uses Qwen2.5-VL-7B-Instruct and MedGemma-4B with frozen A0–A2 traces. Its score is the difference between token-averaged conditional log-likelihoods of ANSWERABLE and UNANSWERABLE. It sees the question, decomposition, grounding, and claim verdicts, but not the gold label, corruption type, or rule decision. No parsing failures occur among the 376 retained outputs. Stored A4 answers are reused rather than regenerated, so the experiment isolates the selection mechanism. Native argmax, leave-one-dataset-out threshold calibration, matched coverage, and matched harm are shown in Table 12. This is one prompted judge on two backbones, not a search over optimal prompts.

Table 12: 376 pooled outputs, 188 per backbone. Paired entries give pessimistic / optimistic scenarios for outputs with no verified reference answer. Calibration uses the other dataset; the pooled calibrated operating point is estimated separately. Matched-harm coverage and AnsF1 also depend on the scenario.

Backbone	Aggregator	Coverage	AnsF1	Harmful/100
MedGemma	Rule	0.782	0.803	49.47 / 39.36
	Native judge	0.739	0.714	47.34 / 38.83
	Calibrated judge	0.915	0.776	63.30 / 53.19
	Judge at rule coverage	0.782	0.723	51.06 / 42.02
	Judge at rule harm	0.761 / 0.745	0.711 / 0.719	49.47 / 39.36
Qwen	Rule	0.665	0.897	40.96 / 30.85
	Native judge	0.426	0.609	24.47 / 18.62
	Calibrated judge	0.798	0.736	56.91 / 48.40
	Judge at rule coverage	0.665	0.762	43.62 / 35.11
	Judge at rule harm	0.633 / 0.596	0.732 / 0.711	40.96 / 30.85
Pooled	Deterministic rule	0.723	0.848	45.21 / 35.11
	Same-backbone judge, native	0.582	0.668	35.90 / 28.72
	Judge at rule coverage	0.723	0.738	47.61 / 39.10
	Judge at rule harm budget	0.697 / 0.670	0.721 / 0.711	45.21 / 35.11
	Calibrated judge	0.957	0.792	68.88 / 58.78

I.6 Secondary AI validation of image-mismatch labels

As a secondary check on label reliability, the AI model described in Appendix H examined whether donor images still supported the associated questions. This section describes the validation protocol and its findings; the use of these judgments in the corrected label set is specified in Appendix H.

Construction. The image-mismatch generator samples SLAKE donors within the matched modality but does not check the body region: 103/166 SLAKE donors in the test construction pool share the original `body_region` label. VQA-RAD has no modality or region metadata, so its donors are drawn from the full test pool. The swap therefore does not guarantee that the evidence required by Q is removed.

Automatic annotation matches. Comparing normalized questions with the SLAKE annotations of donor images finds a donor answer to the same question in 25/166 construction-pool cases and 6/50 cases of a fixed seed-42 audit sample. These are a lower bound on invalid labels.

Secondary validation protocol. The AI model received the question and donor image and returned ANSWERABLE, UNANSWERABLE, or UNCERTAIN, together with a proposed answer and a confidence score of 1–5. The check covered all 33 image-mismatch inputs in the benchmark (14

SLAKE and 19 VQA-RAD) and a separate 50-case SLAKE sample to examine label reliability in the construction pool.

Prevalence in the SLAKE pool. In the 50-case sample, 30 inputs were judged answerable with the original answer, 10 with a different answer, 5 unanswerable, and 5 uncertain: 40/45 labels with a definite judgment (88.9%, Wilson 95% CI 76.5–95.2%) are invalid. By SLAKE content type, answerable swaps are 10/10 for Modality, 7/7 Abnormality, 8/9 Organ, 2/2 Plane, 2/2 Quantity, 4/6 Knowledge, 4/8 Position, 2/3 Color, and 1/3 Size. Questions about a modality or plane are answerable by construction, because the donor is modality-matched.

Answerable question types. The 16 evaluated image mismatches judged answerable include acquisition attributes, presence or visibility, normal/abnormal status, counts, and appearance. For example, a donor may still support "What is the scanning plane of this image?" or "Is the lung healthy?". Swaps remain answerable when the question asks about a property the donor shares with the original image, or when the donor image still provides enough evidence to answer a closed question, often with a different answer (e.g. "no"). Inputs judged unanswerable ask about a structure absent from the donor ("Where is rectum in this image?" on a renal-level CT; "How many lesions are in the spleen?" on a brain MRI) or about patient attributes not visible in the image ("Is this patient male or female?").

Limitations of secondary validation. The check uses a single AI rater, so inter-rater agreement was not assessed and no clinician audit was applied. Model judgments provide an additional check on label reliability, not a clinical reference standard. In one pool case, the model reported no visible disease where SLAKE annotates lung cancer on the donor image.

J Full component ablation table

The P1/P2/P3 comparison is recomputed under corrected labels and measures their combined effect (§4.4).

Table 13: PICV-base versus PICV-full (P1–P3 enabled). The final row averages all eight cells.

Backbone	Base SLAKE	Base VR	Full SLAKE	Full VR	Δ avg (pp)
Gemini	0.820	0.803	**0.894**	**0.930**	**+10.0**
GPT-mini	0.847	0.862	**0.866**	0.862	+1.0
MedGemma	0.794	0.782	**0.803**	**0.803**	+1.5
Qwen	0.851	0.906	**0.882**	**0.914**	+2.0
Avg.	0.833	—	**0.869**	—	**+3.6**

Backbone	Dataset	Choice removal	Image mismatch	Question corruption
Gemini	SLAKE	+53 (19)	-67 (3)	+33 (3)
Gemini	VR	+69 (29)	+14 (7)	—
GPT-mini	SLAKE	+16 (19)	-33 (3)	+33 (3)
GPT-mini	VR	+24 (29)	+14 (7)	—
MedGemma	SLAKE	-11 (19)	-33 (3)	0 (3)
MedGemma	VR	+24 (29)	-29 (7)	—
Qwen	SLAKE	+11 (19)	-33 (3)	0 (3)
Qwen	VR	+7 (29)	-14 (7)	—

Lower panel: ΔTAR in percentage points, with unanswerable counts in parentheses. Image-mismatch and question-corruption cells contain only 3–7 inputs; their deltas do not support stable component-effect estimates.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 state the contributions claimed in the paper, with scoping to the evaluated datasets and backbones; broader generalisation is flagged as out-of-scope in Appendix E.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix E discusses the main limitations of the study (synthetic unanswerability construction, dataset-conditional P2 contribution, frozen entity-region table, English-radiology scope, and confounded backbone-size/domain effects); exploratory within-cell uncertainty estimates and their dependence limitations are reported in Appendix C.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not introduce formal theoretical results; the deterministic aggregator (§3.7) is specified procedurally and its behaviour is supported by empirical evidence rather than proof.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The corruption protocol is fully deterministic given a fixed seed (§3.1); the PICV pipeline is described in §3.3–§3.7 and Appendix A, with all prompts, the entity-region table $\mathcal{M}$, and the aggregator released alongside the paper.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The public repository contains the corruption code, the PICV pipeline, prompts, the entity-region table, baseline implementations, and evaluation scripts; the public datasets are downloaded by helper scripts at first run.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: PICV requires no fine-tuning; data splits, design choices, and inference parameters are specified in §3.1–§3.7 and Appendix A.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Appendix C reports exploratory within-cell paired bootstrap intervals, including their sampling assumptions and limitations. The main text reports pooled additional analyses descriptively because source-question dependence across backbones is not accounted for by the saved bootstrap analyses. No pooled significance or equivalence claim is made.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A specifies the on-premise hardware and inference runtime, the frontier API access path, per-cell wall-clock, and total compute including preliminary runs that did not appear in the final tables.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The work uses only public, de-identified medical imaging benchmarks and recruits no human subjects; safety, bias, and dual-use considerations are addressed in §1, Appendix E, and Appendix F, and PICV is positioned as a tool that requires clinician oversight rather than autonomous deployment.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix F discusses both the intended positive impact of typed selective answering for clinical safety and the relevant negative-impact considerations (over-abstention, deployment-time distribution shift, population coverage, and patient-data processing).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A]

    Justification: We release no new model weights and no new patient data.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The two public radiology benchmarks and the four backbone VLMs are used under their respective public licenses and provider terms, all credited in §2 and Appendix A.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New artefacts (the deterministic answerability benchmark constructed from public sources, prompts, the entity-region table, the aggregator, and re-implemented baselines) are released with a README documenting construction, license inheritance, intended use, and limitations.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and no research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research was conducted; the work re-uses only public, de-identified imaging benchmarks, so IRB review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [Yes]

    Justification: The core research contributions of this work, including the problem formulation, the deterministic answerability benchmark design, the PICV method, and the empirical findings, were developed and decided by the authors. LLM assistance was used in an auxiliary capacity for code implementation, experiment analysis, and paper writing, with the authors verifying all claims, results, and citations. Vision-language models are also the *inference engine* for every PICV stage; the four backbones, decoding choices, and the LLM-as-method-component declaration are documented in §4.1 and Appendix A, and no model is fine-tuned in this work. A single AI rater was also used to reassess the answerability of all synthetic inputs; the protocol and its limitations are described in Appendix H, and its judgments are reported as model judgments, not clinical adjudication.