

TIA-VL: Diagnosing and Correcting Template-Induced Ambiguity in Medical Vision-Language Decisions

Yifan Luo¹, Jiaran Shi², Guoming Wang^{1*}, Siliang Tang¹

¹Zhejiang University

²Department of Cardiology, The Affiliated Lihuili Hospital, Ningbo University
{yifanluo, NB21013}@zju.edu.cn

Abstract

Medical vision-language models (VLMs) are often evaluated by report retrieval, grounding, and explanation quality. We show that these metrics can create a *metric illusion*: a model may appear clinically grounded while exploiting acquisition template, crop, orientation, or protocol rather than patient anatomy. We formalize this failure as **template-induced ambiguity** and introduce **TIA-VL**, a counterfactual benchmark with source, template-family, severity, and compound protocol shifts. We also propose **Rationale-Grounded Template Ambiguity Calibration** (R-GTAC), a plug-in calibration layer that discounts template-dominated evidence while preserving anatomy and textual rationale support. Across strong medical VLM-style baselines, including BioViL-T-style frozen features, adapter features, and anatomy-teacher features, R-GTAC reduces wrong same-template errors to 0.0000 on the main protocol splits. We then test whether systems know when to abstain through readiness and counterfactual rationale deferral. Counterfactual rationale readiness filtering further identifies a reliable low-risk 25% subset reaching 0.8500 Top-1 on the hardest compound shift, while revealing that coupled protocol shifts pose critical frontiers for global risk calibration. Thus, TIA-VL supports a conservative message for medical VLM evaluation: models should be judged not only by whether they answer, but by whether their apparent rationales survive anti-shortcut counterfactuals. The code and benchmark will be available at <https://github.com/YifanLuo0103/TIA-VL>.

1 Introduction

Medical VLMs are increasingly evaluated as if fluent image-text alignment implies clinically meaningful reasoning. Radiology exposes a weakness

in that assumption. Projection images such as fluoroscopy and chest radiography contain the anatomy that clinicians care about, but they also contain acquisition style: view position, crop, burned-in marks, contrast, template family, and institutional protocol. Generic VLM training encourages image-text alignment at scale (Radford et al., 2021), and biomedical VLMs extend this paradigm with radiology images, reports, and biomedical instruction tuning (Zhang et al., 2022; Bannur et al., 2023; Wang et al., 2022; Li et al., 2023). The resulting models can align protocol-like image appearance with protocol-like report language and still produce plausible explanations.

This creates a VLM evaluation problem rather than only a medical-imaging problem. Experienced physicians reason from stable structures such as the clavicles, spine, vertebrae, heart, and lung boundaries, while discounting acquisition artifacts and protocol-specific appearance. A VLM that compares radiographs or explains a decision must make the same separation. It should preserve anatomy and clinically meaningful rationale, but it should not treat shared acquisition style as evidence that two images represent the same patient state. When a shortcut-driven match looks plausible, conventional retrieval, grounding, or explanation scores can improve while the model is relying on a clinically irrelevant cue. We call this gap a *metric illusion*.

We study this problem as *template-induced ambiguity*. Given a query image and a candidate set, a model may select the candidate that shares the query’s acquisition template rather than the one that preserves anatomy or rationale. We make this failure measurable with the wrong same-template rate:

$$WST = \Pr[T(\hat{c}) = T(q) \mid \hat{c} \neq c^+],$$

where q is the query, $\hat{c}$ is the top-ranked candidate,

* Corresponding author.

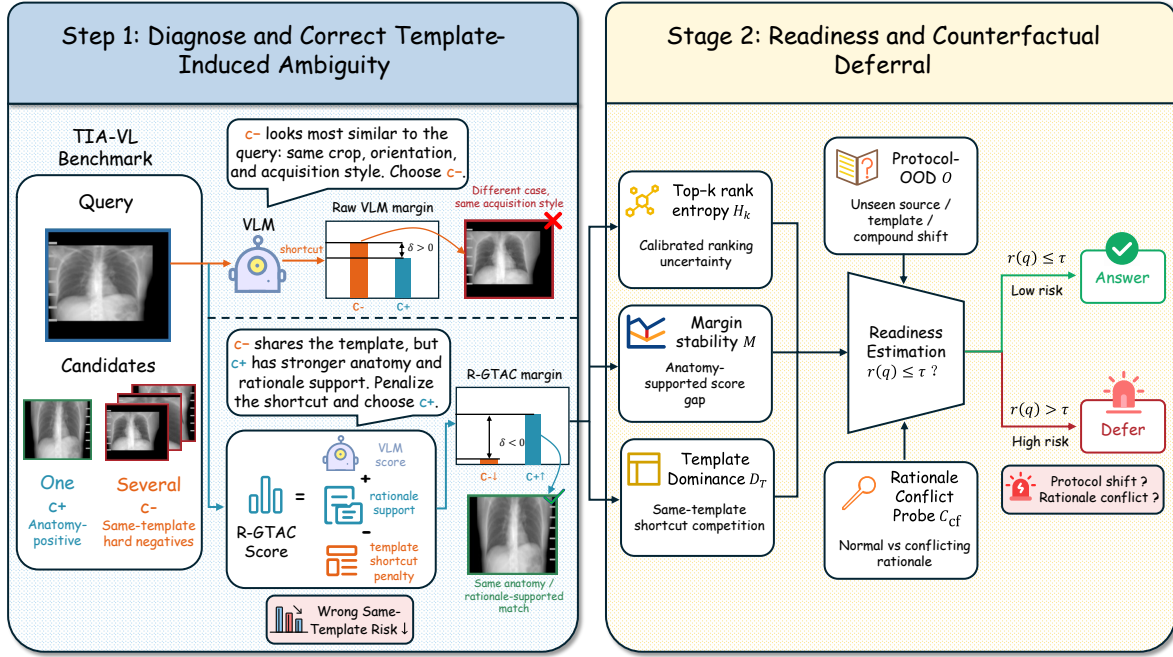


Figure 1: Overview of the TIA-VL and R-GTAC workflow. Step 1 diagnoses template-induced ambiguity with counterfactual TIA pairs and same-template hard negatives: raw VLM margins reveal shortcut failures, while R-GTAC opens the anatomy/rationale-supported margin and drives wrong same-template risk down. The diagnosed shortcut and calibrated R-GTAC scores then feed Step 2, where readiness features, expert-rationale support, and counterfactual-rationale probes decide whether a case should be answered as low risk or deferred under protocol shift.

c^+ is the anatomically matched positive, and $T(\cdot)$ denotes the acquisition-template family.

We make three key contributions. First, **TIA-VL** turns template-induced ambiguity into controlled public tasks with source, template-family, severity, and compound shifts, and measures wrong same-template errors. Second, **R-GTAC** reduces template-dominated evidence while preserving anatomy and rationale support. Third, counterfactual-rationale deferral shows that shortcut correction and confidence awareness are distinct: R-GTAC can remove same-template errors, while compound protocol shifts remain difficult for refusal calibration. This is the central NLP question: if a conflicting textual rationale does not lower readiness, the VLM has not grounded confidence in the rationale it appears to use.

Our claims are intentionally bounded. We do not claim CT-fluoroscopy registration, autonomous navigation, or clinical deployment. Instead, we show that VLMs need explicit benchmark tests for shortcut safety and refusal behavior before they can be trusted in clinical decision-support settings. Figure 1 summarizes the TIA-VL evaluation loop and the R-GTAC correction mechanism.

2 Related Work

Medical image-text representation learning.

Medical VLMs extend the contrastive image-text paradigm into radiology. ConVIRT learns visual representations from paired radiographs and reports (Zhang et al., 2022). MedCLIP relaxes the need for exact image-report pairing by contrastive learning from unpaired medical images and text (Wang et al., 2022). Recent chest radiograph and biomedical VLP systems further exploit report supervision, zero-shot natural-language supervision, expanded image-text construction, and medical knowledge grounding (Tiu et al., 2022; Wu et al., 2023; You et al., 2023). BioViL-T incorporates temporal structure from prior images and reports for biomedical vision-language processing (Bannur et al., 2023). Instruction-tuned and generalist systems such as LLaVA-Med, Med-Flamingo, RadFM, Med-PaLM M, CheXagent, and BiomedGPT further move biomedical VLMs toward open-ended VQA, report generation, dialogue, and multi-task reasoning (Li et al., 2023; Moor et al., 2023; Wu et al., 2025; Tu et al., 2024; Chen et al., 2024; Zhang et al., 2024). These models improve transfer, grounding, report-related tasks, and interactive explanation,

Asset	Scale / role
Protocol axes	source, template, severity, compound
Leaderboard	143 method/split rows
Failure atlas	1020 protocol-error rows
Split summary	48 aggregate metric rows
Private trace probe	88 traces, 20 case cards

Table 1: TIA-VL release assets. Leaderboard rows report retrieval safety; failure-atlas rows support error analysis; private traces are used only as expert-rationale alignment probes.

but our results show that stronger medical VLM features can still be unsafe under template-induced ambiguity.

Radiology datasets and anatomy grounding. MIMIC-CXR provides a large de-identified chest radiograph database with free-text reports (Johnson et al., 2019). Chest ImaGenome adds anatomy-centered scene graphs and anatomical bounding boxes for clinical reasoning (Wu et al., 2021). RadGenome-Anatomy provides large-scale anatomy labels by projecting volumetric structures into radiographic space (Ye et al., 2026). TIA-VL uses this line of public anatomy and metadata resources to build controlled counterfactual protocol variants rather than only evaluating natural dataset splits.

Shortcut learning and selective prediction. Shortcut learning describes models exploiting easy but spurious signals instead of task-relevant evidence (Geirhos et al., 2020). Selective prediction and reject-option models aim to abstain when predictions are unreliable (Geifman and El-Yaniv, 2019). Conformal prediction formalizes distribution-free uncertainty under exchangeability assumptions (Vovk et al., 2005; Lei et al., 2018). Our setting differs because the unsafe signal is structured: protocol/template similarity can produce confident, semantically plausible errors. We therefore evaluate both shortcut removal and selective answering.

3 TIA-VL Benchmark

3.1 Task

TIA-VL evaluates cross-view medical vision-language decisions under controlled acquisition-template ambiguity. Each query has an anatomy-preserving positive and a candidate gallery containing hard negatives. The benchmark asks whether

a method ranks anatomy/rationale support above acquisition-template similarity.

We model an image as a mixture of factors:

$$I = A + T + G + Q + R,$$

where A is anatomy, T is acquisition template, G is geometry such as crop or orientation, Q is image quality, and R is explicit or implicit clinical rationale. The desired representation should preserve A and useful R , discount spurious T , handle G without overfitting, and defer when Q or protocol shift makes the decision unreliable.

The key benchmark object is therefore not a single image, but a *counterfactual pair set*. For each query q , TIA-VL includes an anatomy-preserving positive c^+ and hard negatives $c^- \in \mathcal{N}(q)$ that can share source, template family, crop, or severity with q . Detailed benchmark scale, candidate construction parameters, and confound diagnostics are provided in Appendix B. Operationally, TIA-VL compares whether a VLM ranks the same-anatomy candidate above a different-case candidate that looks procedurally familiar. It is therefore a test of anatomy-grounded matching against protocol-style shortcuts, not a comparison of two arbitrary image augmentations. A template-induced error occurs when a negative with stronger protocol agreement outranks the positive:

$$\begin{aligned} \exists c^- \in \mathcal{N}(q) \text{ s.t. } T(c^-) = T(q), \\ s(q, c^-) > s(q, c^+). \end{aligned}$$

This definition is stricter than ordinary retrieval failure. It asks whether the system is wrong *because the wrong candidate looks procedurally familiar*. That distinction is important for VLM evaluation: an error driven by a missed anatomy token and an error driven by shared acquisition style have different scientific and clinical meanings.

3.2 Splits and Metrics

TIA-VL includes four protocol generalization axes:

- **Unseen source:** held-out public source domain.
- **Unseen template family:** held-out template/crop family.
- **Unseen severity:** held-out geometry severity band.
- **Compound shift:** held-out source, template family, and severity combination.

The hard-negative ladder used for inspection has five protocol-style dimensions: L1 same template,

L2 same orientation, L3 same crop family, L4 severe zoom/geometry, and L5 source split. These dimensions are a cumulative stress ladder rather than five mutually exclusive datasets. Figure 4 gives public examples of the query, anatomy-positive candidate, and shortcut hard negative for each dimension.

Retrieval metrics are Top-1, Top-5, mean reciprocal rank (MRR), and wrong same-template rate. Selective-answering metrics are AUROC, AUPRC, calibration error (ECE), risk-coverage curves, and low-risk subset Top-1. Counterfactual rationale deferral measures whether risk rises when a rationale conflicts with template or anatomy evidence. The release used here contains 143 leaderboard rows, 1020 failure-atlas rows, 48 split-metric rows, and 88 private trace-derived expert-alignment examples. Fold-level confidence intervals and split diagnostics across 5-fold evaluations are detailed in Appendix B.

We report both accuracy and safety metrics because they can move in opposite directions. A representation can improve Top-1 by learning protocol regularities, while simultaneously increasing WST. Conversely, a conservative anti-template correction can reduce dangerous shortcut errors even when it does not fully solve patient-level matching. TIA-VL is designed to expose this tradeoff rather than collapse it into a single leaderboard number.

3.3 Release Package

The release package separates deployable rows from oracle/reference rows. For example, anatomy-teacher rows estimate headroom from stronger anatomy supervision and are not treated as deployable clinical claims. Table 1 summarizes the assets used in this paper. The point of the release format is not only to report a single retrieval leaderboard, but to make each method row analyzable by split, protocol condition, shortcut error, and selective-answering behavior. This lets the same benchmark support ranking evaluation, failure-atlas inspection, and refusal calibration.

4 Method: R-GTAC

R-GTAC is a score-level safety correction for existing image or image-text encoders. It asks whether a pairwise score is supported by anatomy and rationale, or dominated by acquisition template. We use the same evidence decomposition as the bench-

mark:

$$I(q, c) = A(q, c) + T(q, c) + G(q, c) + Q(q, c) + R(q, c),$$

where A is anatomy, T is acquisition template, G is crop/zoom geometry, Q is image quality, and R is textual or expert rationale. Clinical matching should preserve A and R while discounting T or protocol-correlated G when they dominate a fluent explanation.

Let $s_{\text{img}}(q, c)$ be the image or VLM similarity between query q and candidate c . Let $s_{\text{rat}}(q, c)$ be anatomy/rationale support, and let $m_T(q, c)$ indicate template-related similarity. A simplified calibrated score is:

$$s_{R-GTAC}(q, c) = s_{\text{img}}(q, c) + \alpha s_{\text{rat}}(q, c) - \beta g(q, c) m_T(q, c),$$

where $g(q, c)$ is a learned or selected anti-shortcut gate that increases the penalty for likely same-template shortcuts.

Margin view. For a positive c^+ and same-template hard negative c^- , define the raw shortcut margin

$$\delta_{\text{img}} = s_{\text{img}}(q, c^-) - s_{\text{img}}(q, c^+).$$

The raw encoder fails when $\delta_{\text{img}} > 0$. R-GTAC reverses this pair if rationale support and template penalty overcome that margin:

$$\alpha \Delta_{\text{rat}} + \beta \Delta_T > \delta_{\text{img}},$$

where $\Delta_{\text{rat}} = s_{\text{rat}}(q, c^+) - s_{\text{rat}}(q, c^-)$ and $\Delta_T = g(q, c^-) m_T(q, c^-) - g(q, c^+) m_T(q, c^+)$. This is a mechanism view, not proof of clinical correctness: the positive needs stronger anatomy/rationale support, or the negative needs a larger template penalty. We interpret the ablations through this margin. This also explains why a post-hoc correction can be useful even when the underlying encoder is fixed. If the encoder already contains some anatomy signal, R-GTAC can expose it by suppressing template-dominated evidence at ranking time. If the encoder contains little anatomy signal, the margin condition will fail, which is why we report both weak frozen features and stronger anatomy-teacher rows.

Rationale grounding. Public anatomy masks and tokens become rationale support over the lung, heart, clavicle, spine, and vertebrae. Private doctor traces are parsed as reasoning components rather

than dense segmentations, matching how clinicians mark stable landmarks. This choice is important for projection images: an expert trace often marks a line, anchor, or region used to reason about geometry, not a complete organ segmentation. The rationale channel therefore represents clinically meaningful support rather than another global visual embedding.

Selective answering. Readiness models estimate risk instead of forcing every query to return an answer. v9 combines case uncertainty, conformal-style tail risk, and counterfactual sensitivity:

$$r(q) = h(H_k(q), M(q), O(q), D_T(q), C_{cf}(q)),$$

where H_k is top- k rank entropy, M is margin stability, O is protocol-OOD evidence, D_T is same-template dominance, and C_{cf} is counterfactual-rationale sensitivity. A threshold τ answers only when $r(q) \leq \tau$. Changing the rationale while holding the image pair fixed then tests whether confidence is behaviorally tied to the stated clinical reason. Risk-coverage curves test the practical operating point: if the model refuses a high-risk subset, do the remaining answered cases become more reliable? Counterfactual rationale deferral tests a stricter language-grounded behavior: if the stated reason contradicts visible anatomy or protocol evidence, readiness should fall rather than remain a fluent post-hoc explanation.

Gate variants. We report fixed visual G-TAC, a full learned gate, and ablations that remove protocol-OOD, combo-OOD, or template-dominance features. The ablations isolate the core retrieval mechanism: explicit template penalty, with richer gate features retained for refusal and diagnostics. This is why we do not overclaim that every gate feature is necessary for every retrieval split. The retrieval gain mostly tests whether template evidence is penalized; the richer features matter when the same score is used for selective answering and deployment-time risk aggregation.

5 Experiments

5.1 Main Retrieval Safety

Table 2 gives the central safety result. Across the main TIA-VL protocol splits, the learned gate improves or preserves Top-1 over fixed visual G-TAC and drives wrong same-template errors to 0.0000. The gains are not confined to an easy split: Top-1 improves by 0.0760 on severity, 0.0438 on source,

Split	Method	Top-1	WST
Severity	Fixed G-TAC	0.5625	0.0379
Severity	Learned gate	0.6385	0.0000
Source	Fixed G-TAC	0.5781	0.0469
Source	Learned gate	0.6219	0.0000
Compound	Fixed G-TAC	0.5781	0.0271
Compound	Learned gate	0.6359	0.0000
Template	Fixed G-TAC	0.5781	0.0334
Template	Learned gate	0.6328	0.0000

Table 2: Retrieval safety on TIA-VL. R-GTAC with the full learned gate removes wrong same-template errors across the main protocol splits.

0.0578 on compound, and 0.0547 on template-family shift. Thus R-GTAC changes the ranking order in the intended direction: anatomy/rationale-supported positives move above same-template hard negatives. The central point is not only higher retrieval accuracy, but removal of a specific dangerous shortcut. A method could improve Top-1 by learning protocol regularities; WST asks whether those gains come from the wrong same-template candidate. The learned gate improves accuracy while also removing the measured same-template errors.

5.2 Gate Ablation

Table 3 evaluates the empirical necessity of individual learned gate components. Compared with fixed visual G-TAC, the full gate raises average Top-1 from 0.5742 to 0.6323 across the four protocol families and reduces average WST from 0.0363 to 0.0000. A bias-only template penalty is a strong baseline, but it leaves nonzero shortcut risk on source, template-family, and compound shifts. Removing individual OOD features does not hurt on the current retrieval split. This supports a conservative mechanism claim: explicit anti-template calibration is the essential retrieval component, while OOD and dominance features are retained for refusal and deployment diagnostics. The full learned gate gives the model a controlled way to combine shortcut penalty, rationale support, and protocol-risk evidence without treating every gate feature as equally necessary for the retrieval table.

5.3 R-GTAC Ablation

To explicitly decouple the roles of individual components in R-GTAC, Table 4 presents a full 2^3 factorial ablation over template penalty (T), anatomy support (A), and rationale-consistency proxy (R) on the compound split.

Variant	Avg. Top-1	Avg. WST
Fixed visual G-TAC	0.5742	0.0363
Bias-only penalty	0.6268	0.0338
Full learned gate	0.6323	0.0000
No protocol-OOD	0.6323	0.0000
No combo-OOD	0.6323	0.0000
No template dominance	0.6323	0.0000

Table 3: Mechanism ablation averaged over severity, source, template-family, and compound splits.

Variant	Top-1 $\uparrow$	Top-5 $\uparrow$	MRR $\uparrow$	WST $\downarrow$
Base	0.0250	0.0984	0.0931	0.9687
T	0.1313	0.4266	0.2705	0.0656
A	0.7172	0.9359	0.8088	0.6794
R	0.0813	0.2281	0.1724	0.9964
A+R	0.6281	0.8922	0.7420	0.7894
T+A	0.9391	0.9953	0.9643	0.0000
T+R	0.4844	0.7328	0.6015	0.0016
T+A+R	0.9469	0.9984	0.9690	0.0292

Table 4: 2^3 component ablation over template penalty (T), anatomy support (A), and rationale proxy (R) on the compound split.

The results cleanly separate the component roles: ***T* acts primarily as the shortcut-suppression signal**, drastically reducing WST from 0.9687 to 0.0656; ***A* serves as the primary anatomy-ranking driver**, providing the strongest standalone accuracy gain (Top-1 jumps from 0.0250 to 0.7172); and ***R* acts as a secondary consistency support signal** when combined with *T* and *A*. Notably, the *T* + *A* and *T* + *A* + *R* configurations achieve the strongest overall performance (0.9391 and 0.9469 Top-1, respectively), confirming that shortcut suppression and anatomical support are highly complementary. This confirms our framing of R-GTAC as a transparent post-hoc calibration and re-ranking baseline for a measurable protocol-template shortcut, rather than an opaque or monolithic architecture.

5.4 Medical VLM Baselines

Table 5 asks whether stronger VLM-style features remove template shortcuts by themselves. They do not. Frozen BioViL-T-style features have very high WST; adapter and anatomy-teacher features improve retrieval but still leave same-template errors. Adding R-GTAC drives WST to 0.0000 across these rows. On the compound split, anatomy-teacher features reach 0.9047 Top-1 but 0.3113 WST; after R-GTAC, they reach 0.9453 Top-1 and 0.0000 WST. This shows that better anatomy su-

pervision and anti-template calibration are complementary: one strengthens clinical signal, the other prevents remaining protocol evidence from dominating. The comparison is deliberately narrow. It does not claim that a larger VLM is clinically ready, or that the post-hoc score proves patient identity. It shows that stronger medical VLM-style features alone do not guarantee shortcut safety, while a plug-in anti-template layer can remove the measured failure mode.

Because R-GTAC explicitly penalizes same-template candidates, $WST = 0$ alone could be self-confirming if viewed in isolation. We therefore report the full Top-1 error decomposition beyond WST across 2560 query exposures in Table 6. The results demonstrate that R-GTAC not only eliminates same-template errors across all feature families (e.g., from 94.5% to 0.0% for frozen BioViL-T, and from 21.4% to 0.0% for the VLM adapter), but also substantially lowers the total error rate (BioViL-T error drops from 97.5% to 51.2%, VLM adapter from 46.7% to 22.6%, and Anatomy teacher from 9.5% to 5.4%). At the same time, the breakdown transparently shows that some residual errors shift to other negative candidates, confirming that R-GTAC acts as an effective post-hoc shortcut filter rather than an infallible solver.

5.5 Selective Answering

Table 7 summarizes readiness and refusal. High-precision deferral works well on source, severity, and template-family shifts, but the hardest compound shift remains substantially less calibrated. Case-level uncertainty improves the trusted low-risk subset on the compound split to 0.8500 Top-1, but its AUROC remains 0.7458 and ECE is 0.2366. Thus v9 can identify a better low-risk subset, but global risk ranking under compound shift remains open. TIA-VL therefore reports both coverage-conditioned accuracy and calibration metrics. This distinction matters for deployment-style evaluation. The low-risk 25% result answers whether the answered subset becomes more reliable when the system can defer many cases. AUROC and ECE ask the stronger scientific question: can the model globally order all safe and unsafe cases under compound shift? Here the first improves before the second is solved; Figure 3 illustrates the corresponding risk-coverage trade-off on the compound split.

Split	Method	Top-1	Top-5	MRR	WST
Compound	BioViL-T frozen	0.0250	0.0984	0.0931	0.9687
Compound	BioViL-T + R-GTAC	0.4844	0.7547	0.6099	0.0000
Compound	VLM adapter	0.5328	0.8859	0.6839	0.4223
Compound	VLM adapter + R-GTAC	0.7781	0.9625	0.8582	0.0000
Compound	VLM adapter TAC	0.6312	0.9281	0.7574	0.0063
Compound	VLM adapter TAC + R-GTAC	0.7844	0.9625	0.8611	0.0000
Compound	Anatomy teacher	0.9047	0.9906	0.9416	0.3113
Compound	Anatomy teacher + R-GTAC	0.9453	0.9953	0.9670	0.0000

Table 5: External VLM-style baselines on the hardest compound split. Stronger VLM features alone do not guarantee shortcut safety; R-GTAC removes wrong same-template errors.

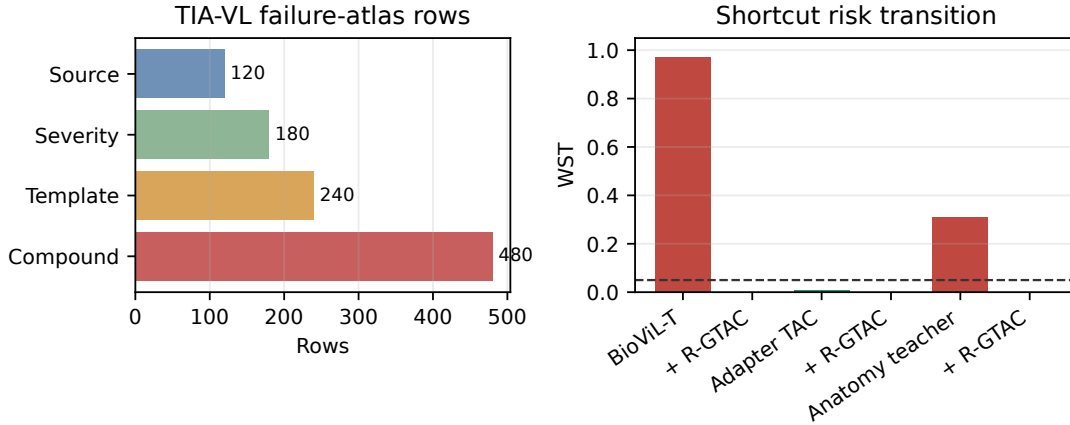


Figure 2: Failure-atlas view of template-induced ambiguity. Left: TIA-VL contains explicit source, severity, template-family, and compound failure rows. Right: high-capacity VLM-style features can carry high wrong-same-template risk, while R-GTAC transitions them to safer calibrated ranking by discounting template-dominated evidence.

Method	Same-template wrong	Other wrong	Total error
BioViL-T frozen	94.5	3.0	97.5
BioViL-T + R-GTAC	0.0	51.2	51.2
VLM adapter	21.4	25.3	46.7
VLM adapter + R-GTAC	0.0	22.6	22.6
Anatomy teacher	5.8	3.8	9.5
Anatomy teacher + R-GTAC	0.0	5.4	5.4

Table 6: Top-1 error distribution beyond WST across 2560 query exposures. Values are reported in percentage (%). R-GTAC completely removes wrong same-template errors while substantially lowering total error rates, though some residual errors shift to other negative candidates.

5.6 Counterfactual Rationale Deferral

Counterfactual rationale tests ask whether confidence falls when rationale text conflicts with image/template evidence. We summarize this behavior as

$$C_{cf} = r(q, c; R_{\text{conflict}}) - r(q, c; R_{\text{normal}}),$$

where R_{normal} is anatomy-consistent and R_{conflict} contradicts anatomy or template evidence. Deferral

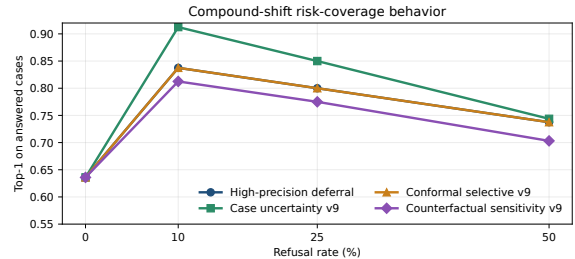


Figure 3: Risk-coverage behavior on the compound split. Refusing the highest-risk cases improves answered-case Top-1, but different readiness policies disagree under compound shift. This is the empirical basis for treating confidence calibration as a separate unsolved challenge rather than assuming retrieval safety implies reliable refusal.

ral is strong on severity and source splits, with v7 counterfactual-only AUROC 0.9602 and 0.9696 respectively. The compound split is harder: v9 counterfactual sensitivity AUROC is 0.7434 (v7 counterfactual-only AUROC 0.7438), and v6 stress-

Split	Policy	AUROC	ECE	Full Top-1	Low-risk 25% Top-1
Severity	High-precision deferral	0.9676	0.1336	0.6385	0.8667
Source	High-precision deferral	0.9762	0.1080	0.6219	0.9187
Template	High-precision deferral	0.9652	0.1351	0.6328	0.8438
Compound	High-precision deferral	0.7698	0.1186	0.6359	0.8000
Compound	Case uncertainty v9	0.7458	0.2366	0.6359	0.8500
Compound	Conformal selective v9	0.7698	0.1585	0.6359	0.8000
Compound	Counterfactual sensitivity v9	0.7434	0.1209	0.6359	0.7750

Table 7: Selective answering. Readiness works well on non-compound shifts. On the compound split, v9 improves performance on the low-risk subset.

Expert focus	<i>n</i>	Align.
Bone-anchor dominant	56	0.547
Mixed anatomy focus	13	0.431
Valve/device ROI	12	0.373
Soft-tissue landmark	6	0.644
Weak/non-anatomy trace	1	0.238

Table 8: Private doctor-trace expert alignment. The dominant category is bone-anchor reasoning, matching the clinical motivation that stable structures such as clavicle, spine, and vertebrae guide projection-image interpretation.

gate deferral AUROC is 0.7245. This is the main language-centered experiment: the rationale is an intervention, and readiness should change when the stated clinical reason conflicts with the image. If changing the rationale leaves readiness unchanged, the system is using language as decoration. If readiness changes in the expected direction, the model’s confidence is at least behaviorally coupled to the stated clinical reason. This reframes explanation evaluation as a behavioral test rather than a surface-form plausibility score.

5.7 Private Doctor-Trace Expert Alignment

We use private fluoroscopy annotations only as an expert-rationale alignment probe. Table 8 summarizes the expert-focus distribution. Among 88 trace-derived images, 56 are bone-anchor dominant, with a mean alignment 0.547; the remaining groups are mixed anatomy focus (13), ROI or valve/device focus (12), soft-tissue or landmark focus (6), and weak/non-anatomy trace (1). The robust observation is distributional: most traces point to stable anatomy, supporting R-GTAC’s use of the clavicle, spine, and vertebrae as rationale-bearing structures. We do not use these traces to claim solved private retrieval, outcome prediction, or deployment readiness. The small soft-tissue landmark group has a

higher mean alignment score, but it contains only six traces and should not be overinterpreted. The more substantiated conclusion is that available expert traces frequently mark stable anchors, consistent with the first-principles claim that anatomy is needed to discount projection-template appearance.

6 Error Analysis

The main failure pattern is structural rather than simply weak image-text capacity. Even anatomy-teacher features can match the wrong same-template candidate: in $I = A + T + G + Q + R$, raw VLM similarity often lets T and protocol-correlated G dominate A and R . The failure atlas indexes this behavior over source, template family, severity, method, and fold. Its 1020 rows include 480 compound-shift rows, 240 template-family rows, 180 severity rows, and 120 source rows. Figure 2 shows large wrong same-template risk for several VLM-style feature rows and its collapse after R-GTAC, separating shortcut safety from ordinary Top-1. This is why the failure atlas is more than a visualization of mistakes. It separates two cases that conventional retrieval metrics conflate: raw or frozen VLM features can prefer shared template appearance, while stronger anatomy features can retrieve the correct anatomy more often but still carry unsafe same-template errors. High Top-1 is therefore not sufficient evidence of clinical safety.

The compound split exposes the harder confidence problem. R-GTAC can set WST to 0.0000 by explicitly penalizing same-template evidence, but v9 case uncertainty still has compound AUROC 0.7458 and ECE 0.2366. This non-linear breakdown in confidence calibration suggests that future medical VLMs need anatomy-preserving geometry invariance and anatomy-token-grounded pre-training, not only larger contrastive encoders or scalar confidence heads. In short, R-GTAC

removes a known ranking shortcut; it does not yet learn a complete anatomy-preserving geometry model. A useful way to read the errors is through the variables in $I = A + T + G + Q + R$. Source and template-family shifts primarily stress T ; severity and crop stress G ; counterfactual rationales stress R ; private traces test whether A and R match what experts mark. R-GTAC is strongest when the failure is dominated by T and enough A or R evidence remains to support the positive. It is weakest when T and G move together so that the same pair appears both protocol-novel and geometry-novel.

7 Discussion

Why benchmark this failure? Template shortcuts are easy to miss because they can improve apparent similarity. TIA-VL makes them measurable with same-template hard negatives, counterfactual rationales, and compound protocol splits. Without these counterfactual conditions, fluent text can make acquisition-style matching appear clinically grounded.

What does R-GTAC solve? R-GTAC addresses retrieval safety: wrong same-template errors can be eliminated across the main protocol splits and strong VLM-style features. This does not imply full anatomical understanding; it shows that a dangerous shortcut can be diagnosed and corrected at the ranking layer. It is a safety layer over existing representations, not a replacement for anatomy-aware pretraining or prospective validation.

What remains open? Confidence calibration under compound shift remains open. When T and G shift together, a scalar readiness head sees a non-linear mixture of template novelty, crop/zoom geometry, quality, and rationale conflict. TIA-VL therefore tests whether a medical VLM’s rationale is actually connected to confidence. The low-risk subset can improve even when global risk ranking remains imperfect, so retrieval safety and refusal calibration must be reported separately.

Why is this an NLP/VLM contribution? The benchmark is grounded in radiology, but the central question is multimodal model behavior: can a VLM distinguish a clinically meaningful rationale from a protocol-correlated shortcut, lower confidence when language conflicts with the image, and defer when a fluent explanation is unsafe? TIA-VL complements report retrieval and explanation-quality benchmarks by testing whether rationales

are causally connected to confidence. The counterfactual rationale is natural language used as an intervention; the measured behavior is whether confidence responds.

8 Conclusion

We introduced TIA-VL, a benchmark for template-induced ambiguity, and R-GTAC, a plug-in anti-shortcut calibration layer. Strong medical VLM features can still rely on acquisition-template shortcuts, while R-GTAC removes wrong same-template errors across the tested protocol splits. The main lesson is conservative: stronger encoders improve clinical signal, but retrieval safety and refusal calibration must still be measured separately. Medical VLMs should be tested not only for answers and rationales, but for when those rationales are unsafe. TIA-VL asks whether apparent rationale support survives anti-shortcut counterfactuals, and whether the system knows when to defer under compound protocol shift.

Limitations

TIA-VL is a chest projection-image benchmark and does not establish generalization to abdomen, limb, or non-radiographic imaging. It does not solve CT-to-fluoroscopy registration, autonomous navigation, or clinical deployment. Private traces are expert-rationale signals rather than dense segmentations or outcome labels, and anatomy-teacher rows should be read as oracle/reference rows. Readiness scores are not calibrated clinical probabilities; under compound shift, v9 case uncertainty still has ECE 0.2366. Prospective blind physician review remains necessary before any clinical-use claim. The current work is therefore a benchmark and mechanism study, not a deployable diagnostic system.

Ethical Considerations

This work studies medical AI safety under controlled public benchmark conditions and retrospective private expert traces. The traces are used only for expert-rationale alignment, not clinical outcome claims. Private data are handled as de-identified internal research artifacts and are reported only through aggregate trace summaries; private image pixels, case identifiers, and source clinical records are not part of the public benchmark release. Deployment would require institutional review, privacy review, prospective validation, and physician-in-the-loop evaluation. The benchmark should not

be used to rank hospitals, clinicians, or patient populations. Its intended use is model evaluation: exposing shortcut behavior before any downstream clinical study or deployment decision. Appendix A consolidates artifact terms, anonymization boundaries, compute and package details, hyperparameters, expert-trace ethics context, and AI-assistant-use disclosure.

Acknowledgments

This work was supported by the NSFC (62436007), the Zhejiang NSF (LRG25F020001), the Key R&D Program in Zhejiang Province (2025C01030), and the Ministry of Culture and Tourism (No. 2023DMKLB002), Ningbo Key Research and Development Program (2025Z178); Ningbo Top Medical and Health Research Program (2025021121).

References

- Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15016–15027.
- Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jena Jitsev, and 4 others. 2024. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.
- Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chihying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.
- Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llavamed: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakkka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pages 353–367. PMLR.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, and 14 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Springer.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21372–21383.

Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, and Mehdi Moradi. 2021. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*.

Shuchang Ye, Mingyuan Meng, Hao Wang, Usman Naseem, and Jinman Kim. 2026. Radgenome-anatomy: A large-scale anatomy-labeled chest radiograph dataset via physically grounded volumetric projection. *arXiv preprint arXiv:2605.17368*.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. 2023. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, and 4 others. 2024. A generalist vision-language foundation model for diverse biomedical tasks. *Nature medicine*, 30(11):3129–3141.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR.

A Responsible Research and Reproducibility Details

Artifact licenses and terms of use. TIA-VL is released as derived benchmark metadata, score summaries, generated tables, and diagnostic figures. It does not redistribute restricted source datasets, private image pixels, private case identifiers, or source clinical records. Source artifact records tracked by the project include the following terms: SCR/JSRT is used through its Zenodo release under CC BY 4.0; Chest ImaGenome is governed by credentialed PhysioNet access, CITI Data or Specimens Only Research training, and a signed data-use agreement; RadGenome-Anatomy is listed under CC BY 4.0; CheXmask-U is listed under Apache-2.0; and ARCADE is listed under CC0-1.0. Users of any released benchmark rows remain responsible for satisfying the upstream dataset and model terms, especially for credentialed resources such as Chest ImaGenome and for any BioViL-T-style model checkpoints or features.

Privacy and anonymization. Public datasets are used in the form made available by their original maintainers and are handled according to their release terms. Private expert traces are treated as de-identified internal research artifacts and are used only for aggregate rationale-alignment analysis. The public release excludes private image pixels, patient names, accession numbers, dates, source clinical records, free-text private reports, and any direct case identifiers. The paper reports only aggregate trace counts and non-identifying alignment summaries.

Consent, ethics review, and expert-trace context. The private-trace component is retrospective and observational: no new prospective clinical intervention, crowdsourcing task, or paid participant-recruitment study is reported. The traces are used as expert rationale marks, not as outcome labels or private retrieval benchmarks. Any applicable approval, exemption, or consent-waiver determination is treated as institution-specific documentation. Independent of that local review category, the public benchmark is designed not to expose private patient data.

Compute, hardware, model size, and software. Heavy image preprocessing, encoder feature extraction, and score-matrix generation were run on a single NVIDIA A100 server. The reported v6/v9 R-GTAC and readiness analyses are post-hoc score-level experiments over exported score matrices; they do not retrain the underlying image encoder. Exact exploratory GPU-hours were not fully logged, so the current artifact should be read as reporting hardware class and pipeline stage rather than a complete carbon or GPU-hour audit. The local paper/artifact environment used Python 3.8.5 with NumPy 1.18.5, PyTorch 1.8.0, torchvision 0.9.0, Pillow 8.0.1, matplotlib 3.3.2, and scikit-learn 0.23.2. The lightweight EmbCNN variants contain 110,144 parameters for the one-channel model and 110,944 parameters for the two-channel model. The ResNet-50 backbone used for BioViL-T-style feature extraction contains about 25.6M parameters before task-specific projection or wrapper layers. The score-level R-GTAC gates and readiness policies are lightweight post-hoc calibrators rather than newly trained large VLMs.

Hyperparameters and thresholds. Core TIA-Bench defaults use five folds, seed 2026, image size 224, embedding dimension 128,

30 epochs, batch size 64, learning rate 10^{-3} , contrastive temperature 0.1, hard-negative boost 0.5, and triplet margin 0.2. Additional default loss and training settings are SWAD start fraction 0.5, Fishr weight 0.5, CNC weight 0.7, view-adversarial weight 0.2 with adversarial lambda 1.0, residual lambda 0.05, factor-template weight 0.2, factor-orthogonality weight 0.05, teacher and canonicalizer training for 12 epochs, teacher distillation and rank weights of 0.5, and rank temperatures of 0.07. For the v6/v9 score-level release experiments, the held-out split families are source, template family, severity, and compound; each setting uses five folds and a stress-task limit of 12. The searched calibration grids are $\alpha \in \{-0.2, -0.1, -0.05, 0, 0.05, 0.1, 0.2, 0.3, 0.5\}$ and $\beta \in \{0, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.8, 1.2\}$. Policies are selected on training or stress folds with a wrong-same-template target of 0.05. Low-risk 25% results use the lowest-risk quartile under the selected readiness score, and answer/defer thresholds are selected before evaluating the held-out fold. The v9 readiness families are base high-precision reference, case uncertainty, conformal selective risk, counterfactual sensitivity, combo readiness, and combo-conservative readiness.

AI-assistant-use disclosure. AI assistants were used for code navigation, checklist drafting, wording suggestions, and manuscript editing support. The authors remain responsible for verifying the final scientific claims, numbers, citations, code, and wording. AI tools did not determine the research claims, did not inspect private data independently, and do not qualify for authorship.

B Benchmark Construction, Confound Diagnostics, and Confidence Intervals

Table 9 provides the complete dataset scale and candidate construction details for TIA-VL. The benchmark comprises 160 original cases producing 640 variants across 4 template categories, evenly drawn from RadGenome (80 cases / 320 variants) and SCR-JSRT (80 cases / 320 variants). Each retrieval query is evaluated against a 128-candidate pool containing 3 anatomy-preserving positive views (transformed views of the same original case) and 50 same-template hard negatives from different cases, evaluated across 5 folds.

Table 10 details the confound diagnostics of the benchmark construction. In chest projection

Benchmark Statistic	Specification
Original patient cases	160
Generated image variants	640
Sources (RadGenome / SCR-JSRT)	80 / 80 cases
Template categories	4 (160 variants each)
Candidate pool size (N)	128 per query
Anatomy-preserving positives (c^+)	3 views
Same-template hard negatives (c^-)	50 candidates
Evaluation protocols / Folds	4 splits / 5 folds

Table 9: Benchmark statistics and candidate pool construction.

Diagnostic Axis	Construction Observation	Implication
Template $\times$ Source	80 RadGenome / 80 SCR-JSRT each	Source-balanced
Source $\times$ Severity	240 standard / 80 severe each	Not source-specific
Template $\times$ Severity	Severe cases in square-zoom	Protocol-linked
Template $\times$ Orientation	Orientation fixed per template	Bundled
Template $\times$ Crop	Crop family fixed per template	Bundled
Template $\times$ View	View role tied to template	Bundled

Table 10: Confound diagnostics for benchmark construction.

imaging, template, crop family, orientation, view role, and severity are naturally bundled or protocol-linked rather than fully independent visual factors. TIA-VL explicitly measures shortcut behavior under these realistic protocol couplings, while ensuring that data sources remain balanced across template categories.

Table 11 reports 5-fold confidence intervals on the hardest compound split. Reporting fold-level confidence intervals provides variance information, confirming that R-GTAC reliably removes wrong same-template retrieval errors across different validation folds.

C Evidence Scope and Artifact Provenance

Table 12 maps the paper’s main evidence types to local exported artifacts. We include them for three reasons. First, the main claim depends on a failure mode rather than a single leaderboard number, so the reader should be able to check which artifacts support retrieval safety, refusal behavior, and expert-trace alignment. Second, several experiments are negative or partially positive: they show that stronger VLM features and stronger deferral policies still leave open problems. Third, the benchmark is intended as a release package, so provenance matters as much as the final metric.

Method	Top-1 mean $\pm$ 95% CI	Conditional WST mean $\pm$ 95% CI
BioViL-T frozen	0.0250 $\pm$ 0.0190	0.9693 $\pm$ 0.0165
BioViL-T + R-GTAC	0.4844 $\pm$ 0.0278	0.0000 $\pm$ 0.0000
VLM adapter	0.5328 $\pm$ 0.0817	0.4612 $\pm$ 0.0781
VLM adapter + R-GTAC	0.7781 $\pm$ 0.0316	0.0000 $\pm$ 0.0000
Anatomy teacher	0.9047 $\pm$ 0.0333	0.5594 $\pm$ 0.2340
Anatomy teacher + R-GTAC	0.9453 $\pm$ 0.0295	0.0000 $\pm$ 0.0000

Table 11: Fold-level confidence intervals on the compound split.

Evidence Domain	Exported Artifact	Diagnostic Purpose
Retrieval safety	Release leaderboard; VLM compound rows	Verifies shortcut elimination across source, template, and compound shifts.
Gate mechanism	Gate ablation summary; generated rows	Isolates explicit anti-template penalty from auxiliary OOD/dominance features.
Selective answering	Readiness policy summary; generated rows	Evaluates the gap between retrieval correction and refusal calibration.
Failure atlas	Failure-atlas CSV (1020 rows)	Categorizes shortcut error patterns by split, method family, and condition.
Expert traces	Private-trace summary; generated rows	Confirms clinical reasoning alignment with stable anatomical anchors.

Table 12: Appendix provenance mapping the manuscript’s evidence domains to local exported artifacts.

D Validity and Leakage Audit

Table 13 states the validity checks that matter most for interpreting $WST = 0.0000$. The concern is legitimate: if the correction directly used the evaluation template label, then eliminating wrong same-template errors would be uninteresting. The reported setting avoids that interpretation in three ways. First, the benchmark split is defined before method scoring, with held-out source, template-family, severity, and compound conditions. Second, WST is computed after ranking as an evaluation statistic; it is not a training label for the reported leaderboard rows. Third, the ablation table keeps a bias-only penalty, fixed G-TAC, full learned gate, and feature-removal variants separate, so the reader can see that shortcut removal is not hidden inside a single opaque system.

This audit also clarifies the role of the release package. TIA-VL is not a single curated hard-negative list optimized for R-GTAC. It contains 143 leaderboard rows, 1020 failure-atlas rows, 48 split-metric rows, and private trace-derived alignment examples. The multiple row types make it possible to inspect whether a method improves accuracy, removes the shortcut, improves selective answering, or merely moves errors to a different condition.

E Additional Retrieval and Gate Ablations

The ablation results in Table 3 support a narrow mechanism claim: explicit anti-template penalty is the essential retrieval-safety component, while richer features (protocol-OD, composite-OD, and template dominance) remain primarily valuable for selective answering and refusal calibration. Once the ranking margin is dominated by a same-template negative, a direct penalty reverses the error if sufficient anatomical signal remains, consistent with the margin view in Section 4.

The failure atlas complements this analysis by indexing the exact structural errors made across splits. Table 14 shows the split composition across 1020 failure rows. The compound block is intentionally the largest (480 rows) because it stresses source, template family, and severity simultaneously—the exact regime where models risk being correct for the wrong procedural reason. The atlas is therefore not a post-hoc visualization used to select the best method; it is a diagnostic table that separates ordinary retrieval failure from a shortcut-driven wrong same-template failure.

Audit Concern	Methodological Safeguard	Integrity Risk Mitigated
Label leakage	Target identity is withheld; pairwise scores are computed prior to evaluation.	Prevents label lookup or oracle leakage during ranking.
Template memorization	Evaluation holds out unseen sources, template families, and compound shifts.	Ensures generalization under distribution shift rather than memorization.
Optimization artifact	WST serves strictly as a post-hoc diagnostic, not an optimization objective.	Avoids metric gaming or masking shortcut-driven errors.
Component opacity	Evaluates isolated heuristic penalties, learned gates, and feature ablations.	Disentangles explicit template penalty from auxiliary OOD features.
Private data misuse	Expert traces are used solely as alignment probes, not as task labels or benchmarks.	Prevents unsubstantiated clinical deployment claims.

Table 13: Validity and leakage audit summary of evaluation safeguards.

Failure-atlas block	Rows
Compound shift	480
Template-family shift	240
Severity shift	180
Source shift	120
Total	1020

Table 14: Failure-atlas composition. Compound shift is emphasized because it is the hardest setting for both retrieval and deferral.

F Expanded VLM Baseline Interpretation

The VLM baseline experiments in Table 5 rule out the alternative hypothesis that template shortcuts naturally disappear with larger capacity: even strong anatomy-supervised features retain substantial same-template error rates before R-GTAC calibration.

Table 15 explicitly defines the intended scientific role and interpretation of each evaluated baseline family, ensuring reference/oracle rows are not conflated with deployable claims.

For closed-source generalist VLMs, we define a standardized prompt-based stress-test protocol rather than reporting unverified API outputs, preserving submission reproducibility.

G Selective Answering and Deferral Details

The deferral experiments were designed after the retrieval experiments because shortcut removal does not imply calibrated confidence. Table 16 collects representative compound-shift rows from the v7, v8, and v9 experiment series. The pattern is deliberately mixed. High-precision deferral and conformal-style reference policies can identify many unsafe cases, but compound shift remains

less reliable than source, template-family, or severity shift alone. The v8 two-stage policy adds an initial extreme-OOD stage, but it does not solve the compound setting. The v9 case-uncertainty policy improves the low-risk 25% subset to 0.8500 Top-1, while AUROC remains 0.7458.

This is why the paper avoids claiming that R-GTAC is a complete medical decision system. The method can remove the measured wrong same-template retrieval failure, but refusal calibration still has a non-linear failure frontier when source, template, severity, crop, and rationale conflict appear together. The appendix results strengthen the main claim precisely because they are not uniformly positive: they show where the method works, where it is only partially effective, and what remains open for future medical VLM evaluation.

H Public TIA Hard-Negative Examples

Figure 4 makes the benchmark comparison concrete. Each row keeps the same retrieval question: the system should choose the anatomy-preserving positive over a different-case hard negative. What changes is the shortcut dimension that makes the negative tempting: shared template, shared orientation, shared crop family, severe zoom/geometry, or source-domain split. The examples use only public benchmark variants; private expert traces are not shown.

I Qualitative Failure Case

Figure 5 gives a public visual case distilled from the failure-atlas and counterfactual-readiness experiments, and Table 17 abstracts the same logic. The goal is not to introduce a new metric, but to make the failure mode easy to inspect. The same query can look safe under a fluent VLM-style similarity score, fail under protocol-like visual agreement, recover when anatomy/rationale support is added,

Baseline Family	Benchmark Role	Diagnostic Takeaway
BioViL-T frozen	Frozen zero-shot image-text representations	Tests off-the-shelf shortcut resistance; fails severely under compound shift.
VLM adapter	Standard fine-tuned feature adaptation	Evaluates standard tuning; gains retrieval accuracy but leaves high shortcut risk.
VLM adapter TAC	Template-conditioned visual adapter	Tests early template awareness; reduces bias but cannot substitute score calibration.
Anatomy teacher	Reference oracle with dense anatomical supervision	Benchmarks anatomical signal headroom; confirms calibration is still required.
VLM GTAC fusion	High-capacity multi-feature fusion reference	Demonstrates that maximizing retrieval metrics alone is insufficient for safety.

Table 15: Roles and diagnostic interpretations of expanded VLM-style baselines.

Series	Compound policy	AUROC	AUPRC	Full Top-1	Low-risk 25% Top-1	High-risk 25% unsafe
v7	High-precision deferral	0.7819	0.9856	0.6094	0.7656	0.9609
v7	Template-conflict refusal	0.6357	0.9400	0.6979	0.9167	0.8750
v8	Two-stage deferral	0.7488	0.9564	0.6359	0.7875	0.9125
v9	Base high-precision reference	0.7698	0.9785	0.6359	0.8000	0.9437
v9	Case uncertainty	0.7458	0.9334	0.6359	0.8500	0.8812
v9	Counterfactual sensitivity	0.7434	0.9583	0.6359	0.7750	0.9250

Table 16: Representative compound-shift deferral rows across experiment series. The low-risk subset can improve before global risk ranking is solved.

and then become unsafe again when the stated rationale is counterfactually inconsistent. This is the behavior that motivates the paper’s evaluation design.

The concrete case in Figure 5 is a held-out SCR/JSRT query under the tight-landscape, high-severity compound condition where fixed visual scoring fails and the rationale-supported score corrects the ranking. The logged row records a stronger same-template shortcut score for the wrong candidate than for the anatomy-positive target under fixed visual scoring, but a stronger rationale score for the anatomy-supported candidate. This is the intended use of the atlas: it shows the transition from template-familiar ranking to anatomy-supported ranking without making a clinical diagnosis claim.

Diagnostic Stage	Observed Model Behavior	Mechanistic Interpretation
Raw VLM scoring	Top candidate shares acquisition template with query despite mismatched anatomy.	Template-induced shortcut: procedural familiarity is mistakenly treated as clinical signal.
R-GTAC calibration	Anatomy/rationale support and template penalty re-rank the true positive to top-1.	Validates that calibration exposes underlying anatomical signals over spurious styles.
Counterfactual probe	Conflicting rationale deliberately injected, causing readiness risk to rise sharply.	Evaluates whether textual explanation acts as a causal intervention rather than superficial text.
Selective answering	The system reliably defers when ranking or rationale fails under the counterfactual probe.	Disentangles ranking-level correction from deployment-grade refusal calibration.

Table 17: Four-stage diagnostic breakdown of the qualitative failure case in Figure 5.

Public TIA-VL examples: anatomy match versus protocol-style hard negative				
Each row keeps the retrieval question fixed and changes which shortcut dimension the hard negative stresses.				
Level	Query	Anatomy positive	Shortcut hard negative	Shortcut tested
L1 same template	 q: RadGenome subset landscape_pan moderate	 c+: RadGenome subset same case / portrait_zoom moderate	 c-: RadGenome subset different case / landscape_pan moderate	Hard negative shares the full template appearance with the query. same template 1, same crop 1, same orientation 1, same source 1
L2 same orientation	 q: RadGenome subset landscape_pan moderate	 c+: RadGenome subset same case / square_shift extreme	 c-: RadGenome subset different case / tight_landscape high	Both images are landscape views, but the negative is a different case. same template 0, same crop 0, same orientation 1, same source 1
L3 same crop family	 q: RadGenome subset portrait_zoom moderate	 c+: RadGenome subset same case / landscape_pan moderate	 c-: RadGenome subset different case / portrait_zoom moderate	The crop family matches the query; in this release it is tied to template family. same template 1, same crop 1, same orientation 1, same source 1
L4 severe zoom geometry	 q: RadGenome subset landscape_pan moderate	 c+: RadGenome subset same case / portrait_zoom moderate	 c-: RadGenome subset different case / square_shift extreme	A severe crop/shift can make a wrong case look deceptively aligned. same template 0, same crop 0, same orientation 0, same source 1
L5 source split	 q: RadGenome subset landscape_pan moderate	 c+: RadGenome subset same case / portrait_zoom moderate	 c-: SCRJSRT different case / landscape_pan moderate	The hard negative crosses the public source domain while preserving protocol appearance. same template 1, same crop 1, same orientation 1, same source 0

Figure 4: Public TIA-VL examples for the five hard-negative dimensions. Each row shows a query, an anatomy-positive candidate from the same case, and a shortcut hard negative from a different case. The five levels are a cumulative inspection ladder, so dimensions can co-occur; for example, source-split negatives can also preserve template appearance. In this release, crop family is generated by template family, so L3 should be read as a crop/template shortcut rather than an independent visual factor.

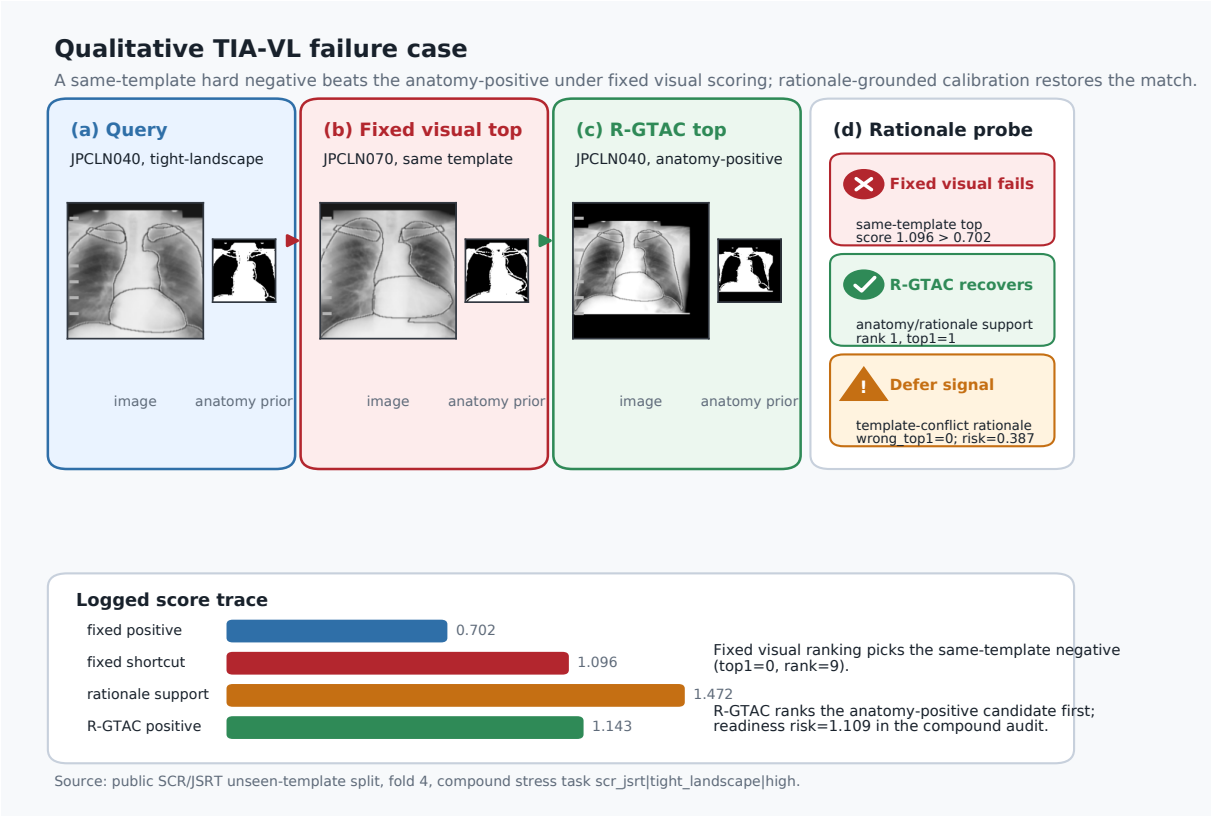


Figure 5: Public qualitative failure case from the SCR/JSRT unseen-template split. The query and the fixed-visual top result share the tight-landscape template but correspond to different cases, producing a same-template hard-negative error. R-GTAC instead ranks the anatomy-positive candidate first by using anatomy/rationale support. A template-conflicting rationale then breaks the recovered match and produces a deferral signal, illustrating why retrieval correction and counterfactual readiness are evaluated separately.