CRAFT: Causal Responsibility and Failure Tracing in Medical Vision Language Models

Chunzheng Zhu[1,*], Jiaqi Zeng[1,*], Hongbo Zhao[1], Yihang Chen[2], Yijun Wang[1,†], Jianxin Lin[1,†]
[1]Hunan University [2]University of Hong Kong
[*]Equal contribution. [†]Corresponding authors.

Abstract

As vision language models are increasingly deployed in clinical diagnosis, understanding how they internally resolve competing visual and textual signals becomes a safety imperative. Existing mechanistic analyses remain confined to unimodal text and offer no explanation for why a single misleading sentence can override a correct image based diagnosis, or why a model commits to a confident answer despite insufficient visual evidence. We find that these two safety risks, arbitration failure where textual context overrides visual grounding and brake failure where the model commits without adequate evidence, are mediated by spatially disjoint attention head populations: arbitration heads form a mid-to-deep wideband reflecting cross-layer evidence competition, while brake heads concentrate in a narrow middle-to-late layer band that regulates evidence sufficiency and abstention behavior. To ground these observations in causal circuitry, we introduce CRAFT, which localizes each failure mode to a minimal causal head set via dual criteria and verifies necessity and sufficiency through temporal probes and Tuned Lens trajectory analysis. Excising arbitration heads sharply reduces conflict following with negligible degradation on clean inputs, while excising brake heads restores appropriate abstention under degraded visual evidence. The two interventions target spatially disjoint head sets and produce distinct corrective effects, underscoring the mechanistic separability of the failure modes. Experiments across multiple medical VQA benchmarks and VLM architectures validate both the localization and interventions, demonstrating that the identified heads causally drive each failure mode and that targeted modulation generalises without retraining. The code is available at https://github.com/zhcz328/CRAFT.

1 Introduction

Medical vision language models are increasingly deployed for clinical question answering and diagnostic support [3, 6, 39], yet their behavior can change sharply when visual evidence, textual context, and parametric medical knowledge disagree [42, 37]. For safety critical deployment, observing these failures at the output level is insufficient: we need tools for identifying which internal components are sensitive to each failure mode and whether targeted interventions can correct the behavior without damaging clean diagnostic reasoning.

Mechanistic interpretability provides a natural starting point [9, 7]. Prior work has mapped specialized attention heads in language models, including memory heads for parametric recall [17], context heads for in-context integration [20, 14], and induction heads for token copying [34], while activation patching studies have localized factual recall to specific MLP modules [25, 33]. Recent extensions trace intra memory knowledge conflicts to MLP attention interactions [30] and reveal that parametric versus contextual signals can superpose within the same head [20]. However, these analyses focus

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

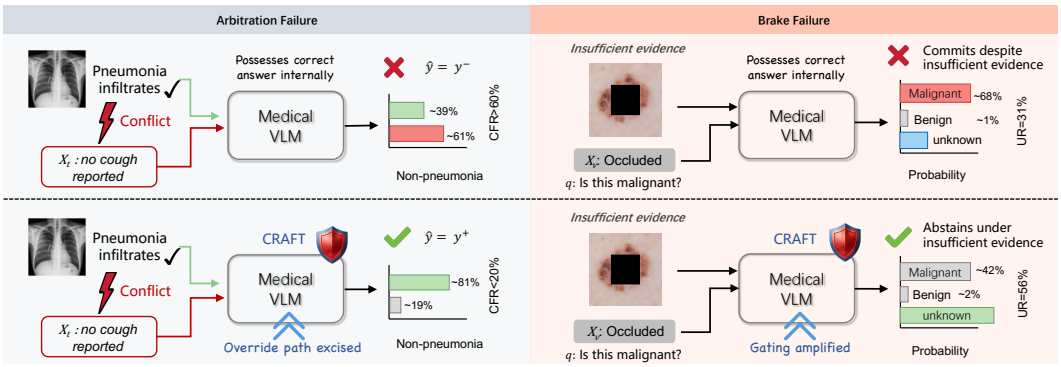


Figure 1: Two failure modes in medical VLMs and CRAFT interventions. Left (Arbitration Failure): Contradictory text hijacks a correct visual diagnosis, causing the model to follow the misleading answer. Right (Brake Failure): The model commits despite occluded evidence.

exclusively on text only LLMs and binary knowledge conflicts. Medical VLMs introduce an additional axis of competition: the model must arbitrate among image evidence, textual context, and learned medical priors, and it should sometimes abstain when visual evidence is insufficient.

Cross modal competition therefore transcends the binary "follow context vs. retrieve memory" choice studied in prior work [23, 40, 32], giving rise to a *tripartite* interaction among three causal pathways: a visual anchoring path $\mathcal{P}_v$, a parametric knowledge path $\mathcal{P}_\Theta$, and a textual override path $\mathcal{P}_t$ (Fig. 1). This three way competition creates two mechanistically distinct safety risks: *(i)* **Arbitration failure**: misleading text contradicts a correct visual diagnosis and the model erroneously privileges text, e.g., a chest radiograph showing clear pneumonia infiltrates accompanied by "no cough reported" yields the output *non-pneumonia*. *(ii)* **Brake failure**: visual evidence degrades below clinical sufficiency yet the model commits to a definitive answer rather than abstaining, *e.g.*, a dermoscopy image with the lesion region masked still prompts a specific diagnosis.

Both failures produce unsafe clinical outputs but reflect fundamentally different computational regimes: arbitration failure indicates a model that *possesses the correct answer but is hijacked by a competing textual signal*; brake failure indicates one that *lacks sufficient evidence yet commits regardless*. Prior work documents these phenomena at the output level [23, 42] and proposes training free mitigations via contrastive decoding [44] or image grounded guidance [45], but has not addressed *where in the computation graph correct reasoning is overridden*. Purely observational head rankings conflate causal influence with correlation: a head active during conflicted outputs may be predictive without being a useful intervention target [26, 12]. Three open challenges remain: **(C1)** whether the two failure modes share circuitry or exhibit separable intervention sensitive patterns; **(C2)** how to identify heads whose manipulation changes failure behavior rather than merely predicts it; **(C3)** whether failure specific circuits can be intervened upon without degrading clean reasoning.

We introduce CRAFT, an activation patching framework for tracing and intervening on these two failure modes in medical VLMs. In the first stage, CRAFT identifies a critical layer band through layer wise conflict scoring, then isolates within that band the minimal head sets whose causal effect on the targeted failure is large while their perturbation to clean input reasoning remains small. In the second stage, linear probes and Tuned Lens trajectories [4] provide complementary temporal evidence about when conflict signals become linearly decodable and when answer preferences shift across layers. Our principal contributions are:

- ◇ We formalize two medical VLM failure modes: arbitration failure under textual conflict and brake failure under visual degradation, separating textual override from insufficient evidence commitment within a unified structural causal framework.
- ◇ We propose CRAFT, which narrows the causal search to a critical layer band via conflict scoring, then isolates a minimal head set whose excision maximally reduces the targeted failure while preserving clean capacity. The resulting targets are failure specific: arbitration head excision suppresses textual override; brake head excision restores abstention under degraded evidence.
- ◇ On multiple VLMs, our experiments reveal that arbitration-sensitive heads form a mid-to-deep band while brake-sensitive heads concentrate in a disjoint late-layer band. Targeted interventions

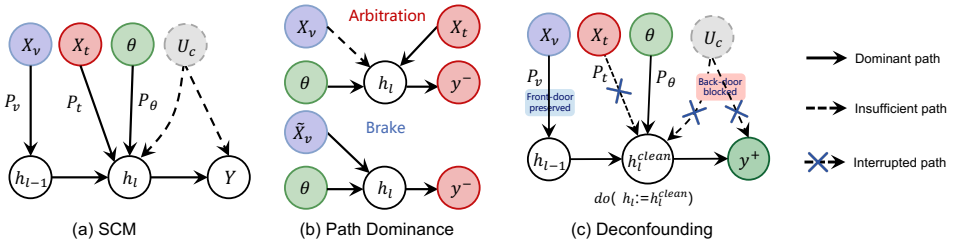


Figure 2: **Interventional tracing setup.** (a) The evidence conflict problem can be modeled by a structural causal model with three competing paths: visual anchoring $\mathcal{P}_v$, parametric knowledge $\mathcal{P}_\Theta$, and textual override $\mathcal{P}_t$, plus latent confounders U_c that open back-door paths. (b) Path dominance under each failure mode. (c) Clean-state interchange blocks back-door paths through U_c while preserving the front-door visual causal path, isolating genuine causal effects for head selection.

reduce Conflict Flip Rate (CFR) by up to 41 absolute points and raise Uncertainty Rate (UR) under insufficient evidence by 75%, with minimal collateral damage on clean diagnostic accuracy.

2 Related Work

Mechanistic Interpretability and Knowledge Conflicts. Circuit discovery via activation patching [25, 7], path patching [11], and attribution based head analysis [1] have identified specialized attention heads for factual recall and in context retrieval [28, 38]. These tools have been extended to knowledge conflict settings: locating "memory heads" whose ablation restores context faithfulness [17], revealing superposition of parametric and contextual signals within the same head [20], and tracing intra memory conflicts to specific MLP attention interactions [30]. Inference time intervention methods such as ITI [21], RepE [47], and conceptor based steering [31] demonstrate that activation modulation can redirect model behavior without weight updates, but operate on coarse trait directions rather than failure specific circuits. Critically, all above studies address text only LLMs; initial mechanistic analyses of VLMs do not study multimodal evidence conflict or clinical failure modes.

Medical VLM Safety and Selective Prediction. Biomedical AI supports radiology QA [19, 22, 46]. SP-Mind automates spatial proteomics analysis through an autonomous reasoning agent [43]. PDU-Net combines path aggregation and dual attention for lung nodule segmentation [5]. MRCF gates cross modal flow with sample specific modality reliability [24]. Medical VLMs remain susceptible to hallucination under ambiguous inputs [29, 15]. Training free mitigations include expert guided contrastive decoding [44], detector grounded guidance [45], and internal representation probes for hallucination risk before generation [18]. Orthogonally, selective prediction [10] and LLM abstention [36] advocate withholding low-confidence answers but rely on calibration or prompt heuristics. Our training-free mechanism instead modulates over-commitment circuits to enable abstention without retraining, though clinical thresholds require deployment-specific human validation.

3 Background and Preliminaries

3.1 Transformer Residual Structure

Given a medical image v and a textual query q, a VLM processes the concatenated visual and textual token sequence through an L-layer decoder. Let $\mathbf{h}_{l,t}$ denote the residual stream at layer l and token position t, and let t_{ans} denote the answer-token position used for next-token prediction. At layer l, the residual stream accumulates contributions from n_h parallel heads and a feed-forward network:

$$\mathbf{h}_{l,t} \;=\; \mathbf{h}_{l-1,t} \;+\; \sum_{i=1}^{n_h} \mathbf{W}_O^{l,i}\, \mathbf{a}_{l,i,t} \;+\; \mathrm{FFN}_l(\mathbf{h}_{l-1,t}), \tag{1}$$

where $\mathbf{a}_{l,i,t} = \mathrm{Attn}^{l,i}(\mathbf{h}_{l-1})_t$ is the value aggregated by head i at position t. The additive structure permits zeroing any individual head's output $\mathbf{W}_O^{l,i}\,\mathbf{a}_{l,i,t} \mapsto \mathbf{0}$ without compensatory redistribution among remaining heads [9, 33], forming the basis for per-head activation interventions that modify a single additive term while leaving the rest of the forward pass unchanged.

3.2 Two Failure Modes and Problem Formulation

Structural view. We posit exogenous variables $\mathcal{U} = \{X_v, X_t, \Theta, U_c\}$, where X_v denotes the visual input (with $\tilde{X}_v$ denoting its degraded counterpart in the brake setting), X_t the textual input, Θ the model parameters, and U_c unobserved exogenous noise capturing latent confounders (*e.g.*, lexical overlap, visual ambiguity, cross-modal coupling). The endogenous variables are $\mathcal{V} = \{\mathbf{h}_1, \ldots, \mathbf{h}_L, Y\}$, where $\mathbf{h}_l$ is the hidden representation at layer l and Y is the model output. These variables are governed by the structural equations $\mathbf{h}_l = f_l(\mathbf{h}_{l-1}, X_v, X_t, \Theta, U_c)$ and $Y = g(\mathbf{h}_{L,t_{\text{ans}}})$, where t_{ans} is the answer token position and f_l, g are deterministic functions. The induced directed acyclic graph (Fig. 2a) admits three competing directed paths from sources to outcome: visual anchoring $\mathcal{P}_v$ (information flow from X_v), parametric knowledge $\mathcal{P}_\Theta$ (from Θ), and textual override $\mathcal{P}_t$ (from X_t).

Preference margin. We formalize both failure modes through $\delta = \log p(y^-) - \log p(y^+)$, where y^+ denotes the correct answer, y^- the incorrect answer, and $p(y)$ is the model-assigned probability. A positive margin $\delta > 0$ indicates the model favours the wrong answer. For multi-token answers, $\log p(y)$ denotes the length-normalised log-likelihood under teacher forcing; for multiple-choice datasets, y corresponds to the candidate option token.

Two mechanistically distinct failures arise from evidence conflict:

Arbitration failure ($\mathcal{P}_t$ dominates $\mathcal{P}_v$ and $\mathcal{P}_\Theta$): misleading textual evidence X_t contradicts visual and parametric sources. The model possesses the correct answer under clean conditions ($\delta_{\text{clean}} < 0$, where δ_{clean} is the margin under non-conflicting inputs) yet is hijacked under conflict (δ_{conflict}, the margin under conflicting inputs) (Fig. 2b, left).

Brake failure ($\mathcal{P}_v$ insufficient): degraded visual evidence $X_v = x_v^{\text{degrade}}$ cannot sustain a diagnosis, yet the model commits to a concrete answer $y^{\text{com}} \in \mathcal{Y}_{\text{con}}$ (the set of concrete answers) rather than abstaining with y^{unk} (an "unknown" or abstention option) (Fig. 2b, right). Formally:

$$\text{(Arbitration)} \quad \delta_{\text{conflict}} - \delta_{\text{clean}} > 0, \qquad \textit{i.e.}\text{, conflict shifts preference toward } y^-, \tag{2}$$

$$\text{(Brake)} \quad \log p(y^{\text{com}}) > \log p(y^{\text{unk}}), \quad y^{\text{com}} \in \mathcal{Y}_{\text{con}}, ; X_v = x_v^{\text{degrade}}. \tag{3}$$

3.3 Interventional Tracing Setup

The confounding problem. Naïve observational attribution (*e.g.*, ranking heads by activation magnitude) conflates causation with correlation. In the SCM of Fig. 2(a), the latent variable U_c opens back-door paths $h_l \leftarrow U_c \rightarrow Y$: token-level shortcuts inflate apparent head importance, visual ambiguity causes widespread activation variance correlated with Y through U_c rather than through $\mathcal{P}_v$, and cross-modal coupling via shared normalisation creates statistical dependence that does not reflect causal influence. The observational quantity $\mathbb{E}[Y \mid \mathbf{a}_{l,i}]$ therefore conflates the direct causal effect $h \rightarrow Y$ with the spurious association $h \leftarrow U_c \rightarrow Y$.

Deconfounding via clean-state interchange. To block these back-door paths, we exploit the modularity of structural causal models: intervening on an endogenous variable severs all incoming arrows while preserving downstream equations [8]. For each conflict input we construct a paired conflict-free counterfactual sharing the same (X_v, q) and cache its clean activation $\mathbf{h}_l^{\text{clean}}$. The interchange intervention $\text{do}(h_l := h_l^{\text{clean}})$ fixes the head output independently of the current U_c realisation, thereby d-separating h_l from U_c and blocking all back-door paths (Fig. 2c). Any residual change in Y is then attributable solely to the conflict mechanism mediated by the intervened head rather than to confounding (Appendix 9 for the d-separation proof).

Per-head activation scaling. We index heads by $h = (l, i)$ and write $\mathcal{C} \subseteq [L] \times [n_h]$ for a head set with $\mathcal{C}_l = \{i : (l, i) \in \mathcal{C}\}$. For a set $\mathcal{C}$ we define *additive head-level scaling* at all token positions:

$$Y_{\mathcal{C}}^{(\alpha)} \;=\; g\Big(\mathbf{h}_{L,t_{\text{ans}}} \;\Big|\; \mathbf{h}_{l,t} \leftarrow \mathbf{h}_{l,t} + \sum_{i \in \mathcal{C}_l} (\alpha_{l,i} - 1)\, \mathbf{W}_O^{l,i}\, \mathbf{a}_{l,i,t},\ \forall l, t\Big), \tag{4}$$

where $\alpha_h = \epsilon$ with $\epsilon \rightarrow 0^+$ realises near-complete *suppression* (denoted $\text{excise}(\mathcal{C})$, used for both arbitration and brake heads). Because the intervention fixes head outputs independently of the current realisation of U_c, the measured effect reflects the deconfounded causal contribution of each head

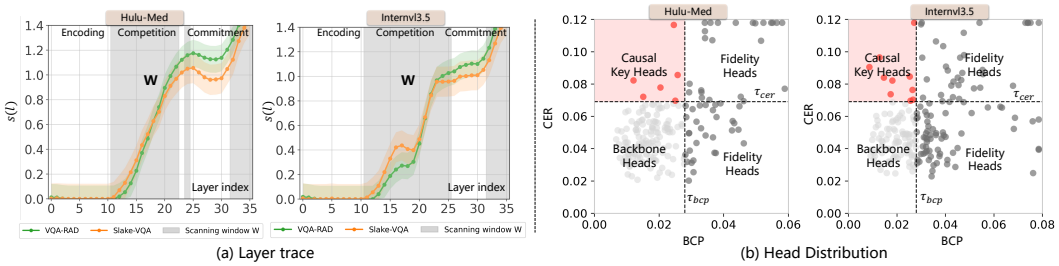


Figure 3: Arbitration failure localization. **(a)** Layer-wise $S(l)$: shaded window marks layers where override concentrates. **(b)** CER vs. BCP: red points (upper left) are selected causal key heads.

to the targeted failure mode. In this work, "intervention" consistently refers to this activation-level do-operation rather than to a claim about the input-level data-generating process.

localization objective. Equipped with these two deconfounded operations, we identify sparse intervention targets whose interventions improve the targeted failure mode while limiting collateral damage. Rather than treating the selection boundaries or the number of selected targets as pre-specified hyperparameters, we obtain them empirically from the observed intervention profiles after scanning candidate heads or layers:

$$\mathcal{C}^*_{\text{arb}} = \text{ParetoKnee}_{\mathcal{C}} \left\{ \text{CFR}(\text{excise}(\mathcal{C})) \downarrow,\ \text{C2W}(\text{excise}(\mathcal{C})) \downarrow,\ |\mathcal{C}| \downarrow \right\}, \tag{5}$$

$$\mathcal{L}^*_{\text{brk}} = \text{ParetoKnee}_{\mathcal{L}} \left\{ \text{UR}(\text{attenuate}(\mathcal{L}; \beta)) \uparrow,\ \text{U2O}(\text{attenuate}(\mathcal{L}; \beta)) \downarrow,\ |\mathcal{L}| \downarrow \right\}. \tag{6}$$

Here, $\mathcal{C}^*_{\text{arb}}$ denotes the selected arbitration heads, while $\mathcal{L}^*_{\text{brk}}$ denotes the selected brake-related layers under layer-wise attenuation with $\beta < 1$. CFR is the proportion of conflict samples following the misleading text, UR is the fraction of degraded samples where the model correctly abstains, and C2W/U2O capture collateral damage on clean inputs. The effective selection boundaries and the resulting number of selected targets are therefore empirical outcomes of the localization procedure, rather than manually fixed hyperparameters (Appendix 8.2 for definitions).

4 Causal Tracing of Failure Circuitry

We trace each failure mode to its causal origin through a coarse to fine procedure: scoring layers to identify a critical band, then resolving heads within that band driving override circuitry (arbitration failure) and gating circuitry (brake failure). Both stages apply the deconfounded intervention of Sec. 3.3 to ensure selected components are causal drivers rather than observational correlates.

4.1 Arbitration Failure

Method. Given a visual input x_v, a question q, and conflicting text x_t^{conf} asserting a misleading answer y^-, the preference margin $\delta = \log p_\theta(y^- \mid x_v, x_t^{\text{conf}}, q) - \log p_\theta(y^+ \mid x_v, x_t^{\text{conf}}, q)$ quantifies textual hijacking, where $\delta > 0$ means the model follows the misleading text.

We first narrow the search space via a layer wise conflict score. Let δ_{base} denote the preference margin under clean input and δ_{conflict} that under adversarial text. For each layer l, we cache the clean run hidden state and replace the layer l output in the conflict forward pass with this cached representation, yielding a patched margin $\delta_{\text{patched}}(l)$. The conflict score is then:

$$S(l) = |\delta_{\text{conflict}} - \delta_{\text{base}}| - |\delta_{\text{patched}}(l) - \delta_{\text{base}}|, \tag{7}$$

measuring how much patching layer l reduces the conflict effect toward the clean baseline. The critical band $\mathcal{W}$ is identified from the $S(l)$ profile by selecting layers where the score rises sharply and remains elevated, corresponding to layers that actively sustain the textual override.

Within $\mathcal{W}$, we isolate individual heads using two causal metrics: $\text{CER}_{\text{exc}}(h) = \delta^{\text{conflict}} - \delta^{\text{excise}(h)}$, measuring how much excising head h reduces the override, and $\text{BCP}_{\text{exc}}(h) = \text{Acc}^{\text{clean}} - \text{Acc}^{\text{excise}(h)}$, measuring collateral damage on clean inputs. The arbitration critical head set is:

$$\mathcal{C}^*_{\text{arb}} = \{h : \text{CER}_{\text{exc}}(h) > \tau_s \ \wedge\ \text{BCP}_{\text{exc}}(h) < \tau_d\}, \tag{8}$$

retaining only heads causally responsible for the override yet dispensable for normal reasoning.

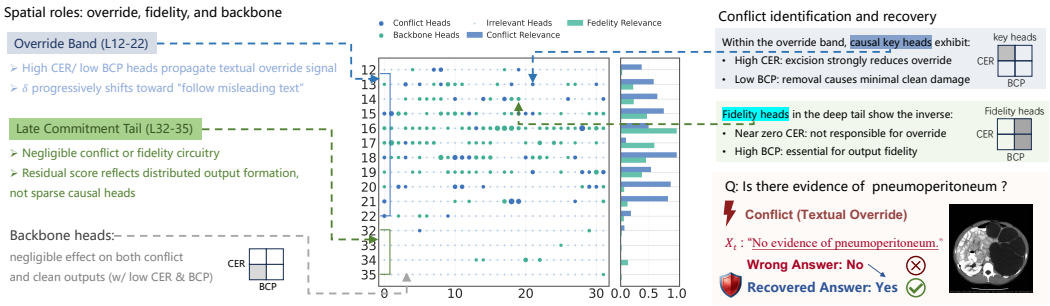


Figure 4: **Arbitration head localization (Hulu-Med 4B).** Left: conflict heads (blue) cluster in the Override Band with high CER and low BCP; the Late Commitment Tail shows negligible conflict or fidelity circuitry, with residual score reflecting distributed output formation. Centre: layer×head grid with per layer CER and BCP bars. Right: selection criterion and recovery example.

Experiments and Analysis. We evaluate on VQA-RAD [19] and SLAKE [22] using Hulu-Med 4B [15] (QwenVL [2] backbone) and InternVL3.5 4B [35]; dataset construction and conflict injection details are provided in Sec. 8. Fig. 3(a) shows that $S(l)$ peaks broadly across mid to deep layers (layers 12–17 in Hulu-Med; 14–20 in InternVL), confirming that textual override distributes across a wideband rather than localizing to a single layer. This contrasts with unimodal factual recall, which concentrates in narrow MLP bands [25].

Fig. 3(b) reveals a sparse cluster in the upper left quadrant: heads with strong causal effect on the conflict yet minimal baseline disruption, constituting fewer than 1% of total heads. Fig. 4 further resolves the spatial structure: conflict heads form a continuous band spanning layers 12–22, while the deepest layers (32–35) exhibit per head CER near zero with negligible conflict or fidelity circuitry, their elevated $S(l)$ reflecting distributed output formation rather than sparse causal heads. This dichotomy persists across architectures, suggesting a shared motif: mid layer override wideband functionally distinct from late layer output formation.

4.2 Brake Failure localization

Method. For brake failure, visual evidence is degraded such that the correct response is abstention y^{unk}, yet the model commits to a concrete answer y^{com}. We apply the same causally grounded deconfounded intervention logic: we first identify layers whose attenuation shifts probability mass toward abstention under degraded input, then localize individual heads within those layers for excision.

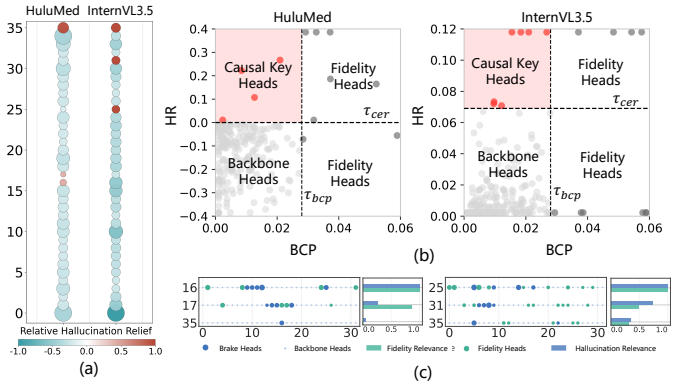


Figure 5: Brake head localization (Hulu-Med 4B). (a) HR(l) map with positive peaks at the deepest layers. (b,c) BCP versus HR landscape separating different heads from backbone.

Following the same deconfounded intervention protocol described above, the layer-wise hallucination relief score quantifies how much scaling down all head outputs in a given layer l by a factor $\beta < 1$ shifts the model's output probability mass toward the abstention token under degraded visual input:

$$\text{HR}(l) = p_\theta(y^{\text{unk}} \mid \text{attenuate}(l)) - p_\theta(y^{\text{unk}} \mid \text{original}). \tag{9}$$

A positive HR(l) indicates that weakening layer l helps the model recognise insufficient visual evidence and suppress premature commitment. Within high HR layers, we isolate individual brake heads by jointly requiring that per head hallucination relief $\text{HR}_{\text{exc}}(h)$ exceeds a suppression bound while the collateral metric $\text{BCP}_{\text{exc}}(h) = \text{Acc}^{\text{clean}} - \text{Acc}^{\text{exc}(h)}$ remains below a preservation bound,

Table 1: Arbitration failure on multimodal benchmarks under textual conflict. CFR: conflict following rate (↓); Resist: resist rate (↑); C2W: correct to wrong (↓); W2C: wrong to correct (↑). Bold: best comparison result. All CRAFT gains are significant ($p < 0.01$).

			Prefix				Before Q				Before A			
Benchmark	**Model**	**Method**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**
VQA-RAD	Hulu-Med 4B	Baseline	33.33	66.67	*N/A*	*N/A*	44.62	55.38	*N/A*	*N/A*	98.21	1.79	*N/A*	*N/A*
		Random	33.59	66.41	1.79	1.54	40.77	59.23	1.03	4.87	98.72	1.28	0.51	0.00
		CRAFT	**15.38**	**84.62**	**0.26**	**18.21**	**14.36**	**85.64**	**0.51**	**30.77**	**70.26**	**29.74**	**0.00**	**27.95**
	Hulu-Med 7B	Baseline	25.80	74.20	*N/A*	*N/A*	36.50	63.50	*N/A*	*N/A*	88.00	12.00	*N/A*	*N/A*
		Random	25.40	74.60	0.90	1.30	35.20	64.80	0.80	2.10	88.30	11.70	0.35	0.05
		CRAFT	**14.20**	**85.80**	**0.10**	**11.70**	**14.70**	**85.30**	**0.20**	**22.00**	**67.50**	**32.50**	**0.00**	**20.50**
	Hulu-Med 14B	Baseline	13.80	86.20	*N/A*	*N/A*	10.60	89.40	*N/A*	*N/A*	24.20	75.80	*N/A*	*N/A*
		Random	14.10	85.90	0.50	0.20	10.40	89.60	0.30	0.50	24.90	75.10	0.80	0.10
		CRAFT	**8.60**	**91.40**	**0.40**	**5.60**	**6.90**	**93.10**	**0.25**	**3.95**	**16.80**	**83.20**	**0.20**	**7.60**
	InternVL3.5 4B	Baseline	30.29	69.71	*N/A*	*N/A*	21.71	78.29	*N/A*	*N/A*	60.57	39.43	*N/A*	*N/A*
		Random	34.29	65.71	4.57	0.57	24.00	76.00	4.00	1.71	77.71	22.29	17.71	0.57
		CRAFT	**9.14**	**90.86**	**0.57**	**21.71**	**7.43**	**92.57**	**2.29**	**16.57**	**19.43**	**80.57**	**0.00**	**41.14**
	Qwen3-VL-8B	Baseline	27.84	72.16	*N/A*	*N/A*	16.16	83.84	*N/A*	*N/A*	57.98	42.02	*N/A*	*N/A*
		Random	28.10	71.90	0.90	0.64	17.00	83.00	1.00	0.16	58.20	41.80	0.35	0.13
		CRAFT	**11.90**	**88.10**	**0.20**	**16.14**	**7.50**	**92.50**	**0.15**	**8.81**	**34.00**	**66.00**	**0.10**	**24.08**
SLAKE	Hulu-Med 4B	Baseline	19.15	80.85	*N/A*	*N/A*	19.15	80.85	*N/A*	*N/A*	87.73	12.27	*N/A*	*N/A*
		Random	21.28	78.72	2.95	0.82	20.79	79.21	4.09	2.45	91.00	9.00	4.26	0.98
		CRAFT	**12.11**	**87.89**	5.07	**12.11**	**7.04**	**92.96**	**0.65**	**12.77**	**46.32**	**53.68**	4.91	**46.32**
	Hulu-Med 7B	Baseline	15.40	84.60	*N/A*	*N/A*	14.20	85.80	*N/A*	*N/A*	75.00	25.00	*N/A*	*N/A*
		Random	15.90	84.10	1.00	0.50	14.80	85.20	0.90	0.30	76.20	23.80	1.50	0.30
		CRAFT	**11.00**	**89.00**	**1.50**	**5.90**	**6.70**	**93.30**	**0.40**	**7.90**	**45.00**	**55.00**	**1.00**	**31.00**
	Hulu-Med 14B	Baseline	8.80	91.20	*N/A*	*N/A*	6.90	93.10	*N/A*	*N/A*	10.40	89.60	*N/A*	*N/A*
		Random	9.10	90.90	0.50	0.20	7.10	92.90	0.40	0.20	10.80	89.20	0.60	0.20
		CRAFT	**6.20**	**93.80**	**0.40**	**3.00**	**5.20**	**94.80**	**0.30**	**2.00**	**4.90**	**95.10**	**0.40**	**5.90**
	InternVL3.5 4B	Baseline	12.66	87.34	*N/A*	*N/A*	8.93	91.07	*N/A*	*N/A*	16.40	83.60	*N/A*	*N/A*
		Random	18.18	81.82	8.28	2.76	18.51	81.49	10.71	1.14	23.70	76.30	8.77	1.46
		CRAFT	**6.66**	**93.34**	**0.65**	**6.66**	**5.84**	**94.16**	**0.65**	**3.73**	**5.84**	**94.16**	**2.27**	**12.82**
	Qwen3-VL-8B	Baseline	15.60	84.40	*N/A*	*N/A*	11.80	88.20	*N/A*	*N/A*	45.20	54.80	*N/A*	*N/A*
		Random	16.20	83.80	0.85	0.25	12.40	87.60	0.75	0.15	47.00	53.00	2.10	0.30
		CRAFT	**10.80**	**89.20**	**0.30**	**5.10**	**7.20**	**92.80**	**0.25**	**4.85**	**23.20**	**76.80**	**0.30**	**22.30**

yielding the candidate set $\mathcal{C}^*_{\text{brk}}$. The interpretation mirrors arbitration: we seek heads that are ***causally responsible for suppressing abstention*** yet dispensable for normal diagnostic reasoning.

Experiments and Analysis. We construct degraded inputs by masking diagnostically critical regions and set abstention as the target (protocol in Sec. 8). Fig. 5(a) reveals a spatial signature distinct from arbitration: while the arbitration override distributes across a broad mid to deep wideband, the hallucination relief signal is *sharply focal*, concentrating in layers 31–35 of Hulu-Med and layers 25–35 of InternVL3.5. This narrow band accounts for over 78% of total HR improvement despite spanning fewer than 15% of model depth, indicating that the abstention mechanism is governed by a ***compact, late-layer circuit***. At head level (Fig. 5(b, c)), only 4–7 heads per model carry statistically significant brake signal ($p < 0.01$, permutation test). Excising these heads yields HR gains of 34.2 (Hulu-Med) and 29.7 (InternVL3.5) with BCP below 2.1, restoring appropriate uncertainty without degrading correct answers. Notably, the selected brake layers are entirely disjoint from the arbitration critical band, corroborating that conflict resolution and confidence calibration are handled by separate sub-circuits. Tuned Lens trajectory analysis further shows that brake heads act as late redirectors: the output distribution commits to a hallucinated token by layer 24, and brake heads enforce premature commitment only in the deepest layers, explaining why shallow probe methods fail to detect.

5 Functional Roles and Decoupling Analysis

Having localized intervention-sensitive heads, we evaluate whether the two failure modes respond to distinct interventions and examine the functional specialisation of the identified head sets.

5.1 Decoupling of Arbitration and Brake Interventions

Setup We compare CRAFT against random ablation using the same head count as $|\mathcal{C}^*|$. For VQA experiments, evaluation spans four medical VLMs (Hulu-Med 4B, Hulu-Med 7B, InternVL3.5 4B, and Qwen3-VL-8B [2]) on VQA-RAD and SLAKE for textual conflict, and on SLAKE and HealMed-VQA [27] for image conflict. More backbone and task results are reported in Appendix 17.

Table 2: Brake failure under visual degradation. CRAFT excises the localized brake heads to recover abstention. UR: unknown rate (↑); O2U: non-unknown to unknown (↑); U2O: unknown to other (↓).

		Hulu-Med 4B			Hulu-Med 7B			Hulu-Med 14B			InternVL3.5 4B			Qwen3-VL-8B		
Benchmark	**Method**	UR↑	O2U↑	U2O↓	UR↑	O2U↑	U2O↓	UR↑	O2U↑	U2O↓	UR↑	O2U↑	U2O↓	UR↑	O2U↑	U2O↓
SLAKE	Baseline	37.88	–	–	44.12	–	–	48.90	–	–	22.58	–	–	42.42	–	–
	Random	45.45	8.33	0.76	48.82	5.29	0.59	51.10	3.00	0.80	16.13	0.00	6.45	47.73	6.06	0.75
	CRAFT	**65.15**	**27.27**	**0.00**	**61.18**	**17.06**	**0.00**	**58.40**	**10.20**	**0.70**	**32.26**	**9.68**	**0.00**	**55.30**	**12.88**	**0.00**
HealMed-VQA	Baseline	78.44	–	–	85.07	–	–	88.61	–	–	81.56	–	–	84.72	–	–
	Random	86.83	10.18	1.80	88.19	4.31	1.19	89.58	1.45	0.48	82.47	3.06	2.15	87.50	4.17	1.39
	CRAFT	**98.20**	**19.76**	**0.00**	**95.83**	**11.10**	**0.34**	**93.06**	**4.80**	**0.35**	**93.64**	**12.46**	**0.38**	**95.14**	**11.11**	**0.69**

Table 3: Comparison of selection criteria at equal head count. Left: arbitration failure (CER only: highest conflict effect; BCP only: lowest collateral). Right: brake failure on validation split (HR only: highest hallucination relief; BCP only: lowest collateral). Shaded: CRAFT.

Model	**Criterion**	CFR ↓	|ΔCFR| ↑	C2W ↓	#Heads
Hulu-Med 4B	CER only	**6.7**	**37.9**	1.8	6
	BCP only	44.4	0.3	1.3	6
	Dual (Ours)	14.4	30.3	**0.5**	6
InternVL3.5 4B	CER only	9.7	12.0	4.0	10
	BCP only	**7.4**	**14.3**	2.3	10
	Dual (Ours)	**7.4**	**14.3**	**0.0**	10

Model	**Criterion**	UR ↑	Δ(UR) ↑	U2O ↓	#Heads
Hulu-Med 4B	HR only	**65.2**	**27.3**	**0.0**	6
	BCP only	38.6	0.8	**0.0**	6
	Dual (Ours)	**65.2**	**27.3**	**0.0**	6
InternVL3.5 4B	HR only	**61.3**	**38.7**	21.6	7
	BCP only	16.1	-6.5	6.5	7
	Dual (Ours)	32.3	9.7	**0.0**	7

Table 1 reports the effect of excising identified arbitration heads. The critical observation is not merely that CFR decreases, but *how*: the intervention shifts the model from following conflicting text to resisting it, while leaving clean accuracy virtually unchanged (C2W near zero). This aligns with the causal interpretation from Sec. 3.3: excising arbitration heads removes the $\mathcal{P}_t$ override without disrupting the front door path $\mathcal{P}_v$. Random ablation, by contrast, increases CFR in several conditions (*e.g.*, InternVL3.5 Before A), confirming that indiscriminate head removal can exacerbate failure by disrupting backbone reasoning while leaving the override intact. A further insight emerges from conflict position dependence: Before A shows the highest baseline CFR and strongest resistance to intervention, suggesting that when conflicting text is placed immediately before the answer token, the override exploits a shorter residual path beyond the identified head set.

Table 2 reports the corresponding results under visual degradation. Ablating the brake heads identified via the localization procedure in Sec. 4.2 consistently increases the unknown rate across all four models while keeping U2O collateral at zero or near zero, with the improvement driven by larger O2U transitions that redirect non-abstaining predictions toward `unknown`. Random ablation is unstable, yielding smaller gains on Hulu-Med and Qwen3-VL-8B and even reducing UR on SLAKE for InternVL3.5 while introducing nonzero U2O, confirming that mitigating brake failure requires *targeted suppression of the localized commitment heads* rather than generic perturbation.

5.2 Contribution of Individual Selection Objectives

Sec. 5.1 compares selection strategies at equal candidate count. For arbitration, CER only achieves the strongest raw CFR reduction on Hulu-Med 4B ($|\Delta\mathrm{CFR}|$=37.9) but incurs C2W of 1.8, whereas BCP nearly fails to suppress conflict; the dual criterion balances both objectives, yielding substantial CFR reduction with C2W at 0.5. On InternVL3.5 4B, CER only increases C2W to 4.0, revealing entanglement between high CER candidates and clean reasoning pathways, while the dual criterion matches the CFR reduction of BCP only and eliminates all collateral (C2W=0.0). For brake failure, HR only and dual coincide on Hulu-Med 4B, but on InternVL3.5 4B HR only causes severe U2O degradation (21.6%) whereas the dual criterion keeps U2O at 0.0 with cleaner UR gain. These results confirm that the joint criterion is essential for reliable intervention across both failure modes.

6 Causal Verification via Internal Probes

The localization procedure in §4 identifies heads via interventional effects. We verify these correspond to interpretable representational transitions using temporal probing and Tuned Lens [4] trajectory analysis on Hulu-Med 4B (VQA-RAD for arbitration, SLAKE for brake failure).

6.1 Temporal Probe Separation

We train two binary logistic probes on per-layer hidden states at the answer position. **Probe A** (conflict detection) predicts whether the input contains conflict-inducing evidence. **Probe B** (failure prediction) predicts whether the model will enter the corresponding failure mode.

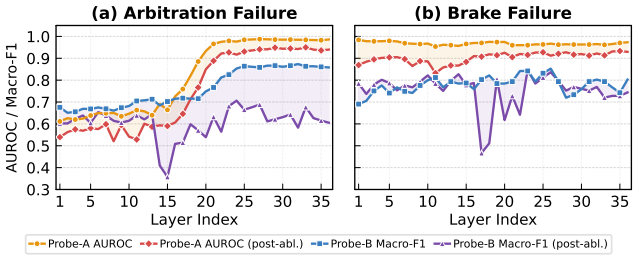


Figure 6: Temporal probe separation. Probe A (detection) saturates before Probe B (commitment), revealing a detect then commit structure. Post-intervention selectively suppresses Probe B while preserving Probe A.

If the identified heads mediate a transition from evidence registration to failure commitment, we expect three testable predictions: (P1) Probe A becomes accurate before Probe B across layers; (P2) Probe B peaks in later layers; (P3) intervening on $\mathcal{C}^*$ selectively weakens Probe B more than Probe A. *i)* For arbitration failure (Fig. 6(a)), Probe A saturates around layers 17–21 (AUROC > 0.97), while Probe B peaks 3–5 layers later (Macro F1 ≈ 0.85–0.87), delineating a computational window where the model has registered the conflict but has not yet committed to following it. After intervening on $\mathcal{C}^*_{\text{arb}}$, Probe A remains intact whereas Probe B drops substantially, confirming that the identified heads govern commitment rather than conflict encoding. *ii)* For brake failure (Fig. 6(b)), Probe A is predictive from shallow layers onward, indicating that degraded visual evidence is linearly decodable early. Probe B peaks later (layers 22–27) with greater variance and lower ceiling. After intervening on $\mathcal{C}^*_{\text{brk}}$, Probe B shows non-monotonic behavior, consistent with the observation that the intervention reorganises late layer competition between commitment and abstention rather than erasing a single pathway.

6.2 Tuned Lens Trajectory Analysis

We apply the Tuned Lens [4] to project layer's hidden state into vocabulary space, computing the preference margin $\delta_l = \text{logit}_l(y^-) - \text{logit}_l(y^+)$ to track when the latent prediction crosses from correct to incorrect.

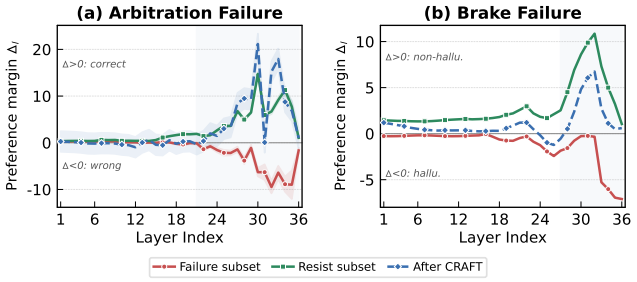


Figure 7: Tuned Lens trajectories. The preference margin Δ_l flips sign within the override window under conflict. Intervention suppresses this reversal, restoring correct dominance.

Under conflict free conditions, δ_l remains negative throughout (Fig. 7). Under conflict, δ_l transitions from negative to positive between layers 18 and 25, placing the preference reversal within the wideband identified by causal localization. The reversal accumulates over several layers, consistent with a distributed override circuit. After excising $\mathcal{C}^*_{\text{arb}}$, this transition is suppressed and δ_l stays negative through the critical window. The convergence of three independent signals (interventional CER, temporal probes, and Tuned Lens trajectories) on the same layer band provides triangulated evidence that the localized heads causally mediate this reversal. Full distributions are in Sec. 12.

7 Conclusion

We presented CRAFT, an interventional tracing framework that leverages causal deconfounding to localize failure-critical attention heads in medical VLMs. The framework reveals that arbitration failure and brake failure occupy structurally disjoint circuitry and respond to opposite interventions: excision of a mid-layer override wideband suppresses textual hijacking, while excision of sparse late-layer gating heads recovers appropriate abstention under degraded visual evidence. Both interventions are training-free, modular, and affect fewer than one percent of total heads, suggesting that sparse activation editing may serve as a practical and lightweight post-hoc safety mechanism for deployed medical VLMs. We hope this mechanistic interpretability perspective on multimodal failure modes opens new and productive directions for principled safety analysis in clinical AI applications.

References

[1] Achtibat, R., Hatefi, S.M.V., Dreyer, M., Jain, A., Wiegand, T., Lapuschkin, S., Samek, W.: Attnlrp: attention-aware layer-wise relevance propagation for transformers. arXiv preprint arXiv:2402.05602 (2024) 3

[2] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 6, 7

[3] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 1

[4] Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., Biderman, S., Steinhardt, J.: Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112 (2023) 2, 8, 9, 19

[5] Chang, C., Fu, M., Chen, X., Feng, S., Zhang, M., Zhou, X., Chuang, J.J., Pan, J., Liu, Z.: Research on pdu-net lung nodule segmentation algorithm based on path aggregation and dual attention. In: 2025 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). pp. 1897–1900. IEEE (2025) 3

[6] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024) 1

[7] Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., Garriga-Alonso, A.: Towards automated circuit discovery for mechanistic interpretability. Advances in Neural Information Processing Systems **36**, 16318–16352 (2023) 1, 3

[8] Didelez, V., Pigeot, I.: Causality: models, reasoning, and inference (2001) 4, 15

[9] Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al.: A mathematical framework for transformer circuits. Transformer Circuits Thread **1**(1), 12 (2021) 1, 3

[10] Geifman, Y., El-Yaniv, R.: Selectivenet: A deep neural network with an integrated reject option. In: International conference on machine learning. pp. 2151–2159. PMLR (2019) 3

[11] Goldowsky-Dill, N., MacLeod, C., Sato, L., Arora, A.: Localizing model behavior with path patching. arXiv preprint arXiv:2304.05969 (2023) 3

[12] Golovanevsky, M., Rudman, W., Palit, V., Singh, R., Eickhoff, C.: What do vlms notice? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. arXiv preprint arXiv:2406.16320 (2024) 2

[13] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 13

[14] Hendel, R., Geva, M., Globerson, A.: In-context learning creates task vectors. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 9318–9333 (2023) 1

[15] Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025) 3, 6

[16] Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X.: Pubmedqa: A dataset for biomedical research question answering. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 2567–2577 (2019) 13

[17] Jin, Z., Cao, P., Yuan, H., Chen, Y., Xu, J., Li, H., Jiang, X., Liu, K., Zhao, J.: Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 1193–1215 (2024) 1, 3

[18] Kogilathota, S.A., EG, S.V., Sun, L., Zhou, J.: Halp: Detecting hallucinations in vision-language models without generating a single token. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6067–6085 (2026) 3, 16

[19] Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 180251 (2018) 3, 6

[20] Li, G., Chen, Y., Tong, H.: Taming knowledge conflicts in language models. arXiv preprint arXiv:2503.10996 (2025) 1, 3

[21] Li, K., Patel, O., Viégas, F., Pfister, H., Wattenberg, M.: Inference-time intervention: Eliciting truthful answers from a language model. Advances in Neural Information Processing Systems **36**, 41451–41530 (2023) 3, 16

[22] Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 1650–1654. IEEE (2021) 3, 6

[23] Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., Singh, S.: Entity-based knowledge conflicts in question answering. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7052–7063 (2021) 2

[24] Meng, C., Pang, J.J., Feng, P., Yu, Z., Ouyang, C., Gan, Z.: Rethinking modality reliability in multimodal sentiment analysis with incomplete observations. arXiv preprint arXiv:2608.03611 (2026), https://arxiv.org/abs/2608.03611 3

[25] Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. Advances in neural information processing systems **35**, 17359–17372 (2022) 1, 3, 6

[26] Neo, C., Ong, L., Torr, P., Geva, M., Krueger, D., Barez, F.: Towards interpreting visual information processing in vision-language models. arXiv preprint arXiv:2410.07149 (2024) 2

[27] Nguyen, D., Ho, M.K., Ta, H., Nguyen, T.T., Chen, Q., Rav, K., Dang, Q.D., Ramchandre, S., Phung, S.L., Liao, Z., et al.: Localizing before answering: A benchmark for grounded medical visual question answering. In: Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25). International Joint Conferences on Artificial Intelligence Organization (2025) 7

[28] Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al.: In-context learning and induction heads. arXiv preprint arXiv:2209.11895 (2022) 3

[29] Pal, A., Umapathi, L.K., Sankarasubbu, M.: Med-halt: Medical domain hallucination test for large language models. In: Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL). pp. 314–334 (2023) 3

[30] Pham, M.V., Borkakoty, H., Hou, Y.: Where knowledge collides: A mechanistic study of intra-memory knowledge conflict in language models. arXiv preprint arXiv:2601.09445 (2026) 1, 3

[31] Postmus, J., Abreu, S.: Steering large language models using conceptors: Improving addition-based activation engineering. arXiv preprint arXiv:2410.16314 (2024) 3

[32] Shi, D., Jin, R., Shen, T., Dong, W., Wu, X., Xiong, D.: Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. Advances in Neural Information Processing Systems **37**, 4997–5024 (2024) 2

[33] Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., Shieber, S.: Investigating gender bias in language models using causal mediation analysis. Advances in neural information processing systems **33**, 12388–12401 (2020) 1, 3

[34] Wang, K., Variengien, A., Conmy, A., Shlegeris, B., Steinhardt, J.: Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. URL https://arxiv.org/abs/2211.00593 **2** (2022) 1

[35] Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 6

[36] Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., Wang, L.L.: Know your limits: A survey of abstention in large language models. Transactions of the Association for Computational Linguistics **13**, 529–556 (2025) 3

[37] Wu, W., Xu, X., Gao, C., Diao, X., Li, S., Salas, L.A., Gui, J.: Assessing and mitigating medical knowledge drift and conflicts in large language models. arXiv preprint arXiv:2505.07968 (2025) 1, 13

[38] Wu, W., Wang, Y., Xiao, G., Peng, H., Fu, Y.: Retrieval head mechanistically explains long-context factuality. arXiv preprint arXiv:2404.15574 (2024) 3

[39] Xia, P., Chen, Z., Tian, J., Gong, Y., Hou, R., Xu, Y., Wu, Z., Fan, Z., Zhou, Y., Zhu, K., et al.: Cares: A comprehensive benchmark of trustworthiness in medical vision language models. Advances in Neural Information Processing Systems **37**, 140334–140365 (2024) 1

[40] Xie, J., Zhang, K., Chen, J., Lou, R., Su, Y.: Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In: The Twelfth International Conference on Learning Representations (2023) 2

[41] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 13

[42] Yeh, C.H., Wang, J., Graham, A.D., Liu, A.J., Tan, B., Chen, Y., Ma, Y., Lin, M.C.: Insight: A multi-modal diagnostic pipeline using llms for ocular surface disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 711–721. Springer (2024) 1, 2

[43] Yuan, Y., Ji, Y., Li, Z., Li, R.: Sp-mind: An autonomous reasoning agent for spatial proteomics analysis. arXiv preprint arXiv:2606.24235 (2026) 3

[44] Zhang, X., Meng, Z., Lever, J., Ho, E.S.: Ccd: Mitigating hallucinations in radiology mllms via clinical contrastive decoding. arXiv preprint arXiv:2509.23379 (2025) 2, 3, 16

[45] Zhao, L., Deng, Y., Zhang, W., Gu, Q.: Mitigating object hallucination in large vision-language models via image-grounded guidance. arXiv preprint arXiv:2402.08680 (2024) 2, 3, 16

[46] Zhu, C., Lin, Y., Chen, S., Wang, Y., Lin, J.: Medeyes: Learning dynamic visual focus for medical progressive diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13916–13924 (2026) 3

[47] Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.K., et al.: Representation engineering: A top-down approach to ai transparency. arXiv preprint arXiv:2310.01405 (2023) 3

8 More Experimental Details

8.1 Models, conflict construction, and data splits

We evaluate four medical VLMs: Hulu-Med 4B (36 layers, 32 heads, d=2560), Hulu-Med 7B, InternVL3.5 4B (36 layers, 32 heads, d=2560), and Qwen3-VL-8B, all running in FP16 on two NVIDIA GeForce RTX 5090 GPUs with 32 GB peak memory. For the arbitration setting, we restrict evaluation to binary answer instances and construct textual interventions via an *EVIDENCE* template: the support condition asserts the correct answer y^+ while the conflict condition asserts the alternative y^-. We retain only instances where the model predicts y^+ on clean input, remains correct under support, and flips to y^- under conflict. For the brake setting, we construct degraded inputs by masking diagnostically relevant regions resolved from `detection.json` (preferring pixel level masks when available, otherwise falling back to detection rectangles) with black fill enlarged by `mask_scale` = 2.0. The prompt explicitly permits answering `unknown`; filtering keeps only instances answered correctly without masking. A mask scale sweep (Fig. 8) confirms that the `Unknown` rate increases monotonically yet remains below 50% even at `mask_scale` = 3.0, motivating intervention. After filtering, all retained examples are split 70/30 into a training split (layer tracing, head localization, threshold selection) and a validation split on which all reported metrics are evaluated.

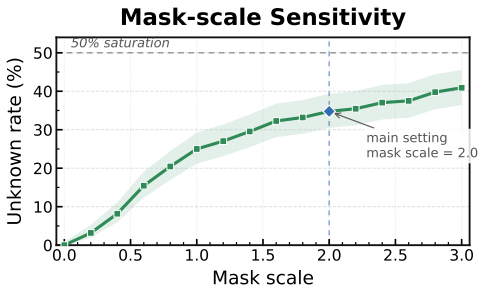


Figure 8: The `Unknown` rate increases with `mask_scale`, but remains below 50% even under the strongest masking. We use `mask_scale`=2.0 as the default degradation setting.

8.2 Metric Formal Definitions and Text-Only Results

***(i)* Evaluation metrics.** The main paper reports four metrics for arbitration failure (Table 1) and three for brake failure (Table 2). Let $\hat{y}_i^{\text{pre}}, \hat{y}_i^{\text{post}}$ denote the model prediction on conflict instance i before and after intervention, and let y_i^+, y_i^- denote the visually correct and textually misleading answers respectively. For *Arbitration setting:*

$$\text{CFR } (\downarrow) = \tfrac{1}{|\mathcal{D}_{\text{conf}}|} \textstyle\sum_{i \in \mathcal{D}_{\text{conf}}} \mathbb{1}\big[\hat{y}_i^{\text{post}} = y_i^-\big], \tag{10}$$

$$\text{RESIST } (\uparrow) = 1 - \text{CFR}, \tag{11}$$

$$\text{C2W } (\downarrow) = \tfrac{1}{|\mathcal{D}_{\text{conf}}|} \textstyle\sum_i \mathbb{1}\big[\hat{y}_i^{\text{pre}} = y_i^+ \ \wedge \ \hat{y}_i^{\text{post}} = y_i^-\big], \tag{12}$$

$$\text{W2C } (\uparrow) = \tfrac{1}{|\mathcal{D}_{\text{conf}}|} \textstyle\sum_i \mathbb{1}\big[\hat{y}_i^{\text{pre}} = y_i^- \ \wedge \ \hat{y}_i^{\text{post}} = y_i^+\big]. \tag{13}$$

Here RESIST denotes "not following the misleading answer" and can include correct, abstention, or other non misleading outputs.

Brake setting. Let y^{unk} denote abstention and $y^{\text{com}} \in \{y^+, y^-\}$ a concrete answer; instances are drawn from the visually degraded subset $\mathcal{D}_{\text{deg}}$.

$$\text{UR } (\uparrow) = \tfrac{1}{|\mathcal{D}_{\text{deg}}|} \textstyle\sum_i \mathbb{1}\big[\hat{y}_i^{\text{post}} = y^{\text{unk}}\big], \tag{14}$$

$$\text{C2U } (\uparrow) = \tfrac{1}{|\mathcal{D}_{\text{deg}}|} \textstyle\sum_i \mathbb{1}\big[\hat{y}_i^{\text{pre}} = y^{\text{com}} \ \wedge \ \hat{y}_i^{\text{post}} = y^{\text{unk}}\big], \tag{15}$$

$$\text{U2O } (\downarrow) = \tfrac{1}{|\mathcal{D}_{\text{deg}}|} \textstyle\sum_i \mathbb{1}\big[\hat{y}_i^{\text{pre}} = y^{\text{unk}} \ \wedge \ \hat{y}_i^{\text{post}} = y^{\text{com}}\big]. \tag{16}$$

***(ii)* Text-only benchmark results.** Table 5 reports results on two text-only medical QA benchmarks, ConflictMedQA [37] and PubMedQA [16], using Qwen3-4B [41] and Llama3.2-3B [13], where conflict is injected purely through contradictory textual context without visual inputs.

Analysis. Across both benchmarks, CRAFT consistently lowers CFR and improves RESIST over both the baseline and random head ablation. The effect is especially strong on ConflictMedQA, where CRAFT substantially reduces conflict compliance across both models and all injection positions

Table 4: **Threshold sensitivity** (Hulu-Med 4B). *Left*: arbitration on VQA-RAD (Before Q). *Right*: brake on SLAKE (image conflict). Highlighted: final settings.

τ_s	τ_d	CFR↓	C2W↓	W2C↑	#Heads
0.133	0.550	18.7	4.9	**30.8**	25
0.271	0.183	24.6	**0.5**	20.5	2
0.156	0.445	22.1	7.9	30.5	31
0.350	**0.500**	**18.2**	0.8	27.4	6

τ_s	τ_d	UR↑	O2U↑	U2O↓	#Heads
0.2	0.5	57.58	20.45	0.76	4
0.0	0.3	54.55	16.67	**0.00**	3
0.0	0.8	62.88	26.52	1.52	9
0.0	**0.5**	**65.15**	**27.27**	**0.00**	6

Table 5: **Text only benchmark results.** *CFR*: conflict compliance (↓); *Resist*: resist rate (↑); *C2W*: correct to wrong (↓); *W2C*: wrong to correct (↑). **Bold**: best per block; shaded rows: CRAFT. Random uses the same head count.

			ConflictMedQA				PubMedQA			
Position	**Model**	**Method**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**
Prefix	Qwen3-4B	Baseline	83.33	16.67	–	–	49.67	50.33	–	–
		Random	94.25	5.75	12.07	1.15	57.17	42.83	12.17	4.67
		CRAFT	**37.36**	**62.64**	**0.00**	**45.98**	**23.59**	**76.41**	**0.00**	**26.09**
	Llama3.2-3B	Baseline	57.69	42.31	–	–	58.46	41.54	–	–
		Random	42.31	57.69	26.92	**42.31**	63.08	36.92	8.85	4.23
		CRAFT	**27.88**	**72.12**	**0.00**	29.81	**46.15**	**53.85**	**1.15**	**13.46**
Before Q	Qwen3-4B	Baseline	94.83	5.17	–	–	53.70	46.30	–	–
		Random	89.66	10.34	1.15	6.32	46.30	53.70	3.91	11.30
		CRAFT	**68.39**	**31.61**	**1.15**	**27.59**	**3.70**	**96.30**	**0.00**	**50.00**
	Llama3.2-3B	Baseline	76.92	23.08	–	–	71.54	28.46	–	–
		Random	81.73	18.27	11.54	6.73	68.08	31.92	4.62	8.08
		CRAFT	**57.69**	**42.31**	**0.00**	**19.23**	**59.23**	**40.77**	**1.54**	**13.85**
Before A	Qwen3-4B	Baseline	80.46	19.54	–	–	65.98	34.02	–	–
		Random	76.44	23.56	4.02	8.05	85.33	14.67	20.54	1.20
		CRAFT	**23.56**	**76.44**	**0.57**	**57.47**	**51.41**	**48.59**	**0.00**	**14.57**
	Llama3.2-3B	Baseline	43.27	56.73	–	–	66.92	33.08	–	–
		Random	48.08	51.92	16.35	11.54	72.31	27.69	9.23	3.85
		CRAFT	**31.73**	**68.27**	**2.88**	**14.42**	**54.62**	**45.38**	**2.31**	**14.62**

while keeping C2W near zero. On PubMedQA, CRAFT remains effective under longer and more naturalistic medical contexts, with the strongest case reducing CFR from 53.70 to 3.70 and increasing RESIST from 46.30 to 96.30. Compared with random ablation using the same number of heads, CRAFT achieves a better balance between W2C gains and C2W collateral damage, confirming that the selected heads are involved in conflict following rather than general answer generation.

8.3 Key Parameter Selection and Sensitivity

The threshold procedure selects $\mathcal{C}^*$ by requiring heads to exceed τ_s on the suppression score ($\mathrm{CER}_{\mathrm{exc}}$ or HR) while falling below τ_d on the disruption score ($\mathrm{BCP}_{\mathrm{exc}}$), with τ_s set near the 90th percentile of the suppression distribution and τ_d near the 10th percentile of the disruption distribution, yielding $|\mathcal{C}^*_{\mathrm{arb}}| \in [4, 13]$ and $|\mathcal{C}^*_{\mathrm{brk}}| \in [5, 7]$ depending on architecture. All metrics are computed on the held out 30% split, averaged over three random seeds (standard deviations < 1.5 pp), with significance assessed via two sided permutation tests (n=10,000, $p < 0.01$). A threshold sensitivity sweep (Table 4) confirms the trade off: looser settings improve correction (W2C/O2U) at the cost of increased collateral damage (C2W/U2O), while overly strict settings leave failure unresolved; the operating points balance target effect retention against competence preservation.

9 Theoretical Analysis of Deconfounding

This appendix formalizes the causal deconfounding argument underlying the head selection procedure in Sec. 4. We first show that observational activation attribution is confounded (theorem 9.1), then prove that clean state interchange eliminates this bias (theorem 9.2), and finally establish that the CER metric inherits the deconfounding guarantee (theorem 9.3).

9.1 Confounding in Observational Attribution

Proposition 9.1 (Confounded Observational Attribution)**.** *Under the SCM defined in Sec. 3.2 with latent confounder U_c and structural equations $\mathbf{h}_l = f_l(\mathbf{h}_{l-1}, X_v, X_t, \Theta, U_c)$, $Y = g(\mathbf{h}_{L,t_{\text{ans}}})$, the observational conditional $\mathbb{E}[Y \mid \mathbf{a}_{l,i}]$ does not equal the interventional quantity $\mathbb{E}[Y \mid \text{do}(\mathbf{a}_{l,i} = a)]$ whenever U_c is not d-separated from Y given $\mathbf{a}_{l,i}$ in the mutilated graph.*

Proof. (i) Active back door path. In the original DAG the path $\mathbf{a}_{l,i} \leftarrow U_c \rightarrow Y$ is active because U_c is a common cause that is never conditioned on. Since U_c encompasses lexical overlap, visual ambiguity, and cross modal coupling, it is neither directly measurable nor finite dimensional, so the back door criterion [8] cannot be satisfied by observational adjustment.

(ii) Bias decomposition. Applying the truncated factorisation formula yields:

$$\mathbb{E}[Y \mid \mathbf{a}_{l,i}{=}a] \;=\; \underbrace{\mathbb{E}[Y \mid \text{do}(\mathbf{a}_{l,i}{=}a)]}_{\text{causal effect}} \;+\; \underbrace{\text{Bias}(U_c{\rightarrow}\mathbf{a}_{l,i},\; U_c{\rightarrow}Y)}_{\text{confounding bias}}, \tag{17}$$

where the bias term is generically nonzero whenever U_c simultaneously affects both the head activation and the model output through distinct structural equations.

(iii) Implication. Activation magnitude ranking is therefore a *biased estimator* of causal importance: a head may appear highly influential simply because U_c drives both its activation and the erroneous output, without the head itself mediating the failure. □

9.2 D Separation under Clean State Interchange

Theorem 9.2 (Deconfounding via Interchange Intervention)**.** *Let $\mathbf{a}_{l,i}^{\text{clean}}$ denote the activation of head (l, i) computed from the conflict free counterfactual sharing (X_v, q) with the original conflict sample. Under $\text{do}(\mathbf{a}_{l,i}{:=}\mathbf{a}_{l,i}^{\text{clean}})$, the back door path $\mathbf{a}_{l,i} \leftarrow U_c \rightarrow Y$ is blocked, and the measured effect*

$$\text{CDE}(l, i) \;=\; \mathbb{E}\big[Y \mid \text{do}(\mathbf{a}_{l,i}{:=}\mathbf{a}_{l,i}^{\text{clean}})\big] \;-\; \mathbb{E}\big[Y \mid \text{do}(\mathbf{a}_{l,i}{:=}\mathbf{a}_{l,i}^{\text{conflict}})\big] \tag{18}$$

*equals the **controlled direct effect** of the conflict mechanism mediated through head (l, i).*

Proof. The do operator $\text{do}(\mathbf{a}_{l,i}{:=}\mathbf{a}_{l,i}^{\text{clean}})$ replaces the structural equation for $\mathbf{a}_{l,i}$ with a constant assignment, mutilating the graph by severing all incoming edges to $\mathbf{a}_{l,i}$ [8].

(i) Back door severed. Since $\mathbf{a}_{l,i}$ is fixed to a constant that no longer depends on the U_c realisation, the edge $U_c \rightarrow \mathbf{a}_{l,i}$ is removed from the mutilated graph, blocking the only back door path $\mathbf{a}_{l,i} \leftarrow U_c \rightarrow Y$.

(ii) Front door preserved. All downstream structural equations remain intact. The causal path $X_v \rightarrow \mathbf{h}_{l-1} \rightarrow \mathbf{h}_l \rightarrow Y$ continues to operate through other heads and FFN components, so the intervention does not destroy the forward computation.

(iii) Input level matching. Both the conflict and conflict free samples share (X_v, q), so the only variation in U_c arises from the presence or absence of adversarial text X_t^{conf}. Fixing $\mathbf{a}_{l,i}$ to its clean state value ensures constancy across both U_c conditions, achieving d separation:

$$\mathbf{a}_{l,i} \perp\!\!\!\perp U_c \;\mid\; \text{do}(\mathbf{a}_{l,i}) \quad \text{in the mutilated graph.} \tag{19}$$

Combining *(i)* through *(iii)*, $\text{CDE}(l, i)$ isolates the causal contribution of head (l, i) to the conflict induced failure, free of confounding bias from U_c. □

Corollary 9.3 (Validity of CER as Causal Metric)**.** *The conflict effect reduction metric $\text{CER}_{\text{exc}}(h) = \delta^{\text{conflict}} - \delta^{\text{excise}(h)}$ is a valid estimate of the controlled direct effect of head h on the preference margin, provided the paired counterfactual shares (X_v, q) with the conflict sample.*

Proof. The excise operation ($\alpha_h \rightarrow 0^+$) is a special case of the interchange intervention in theorem 9.2 that replaces the head output with a near zero vector. The d separation guarantee holds for *any* fixed replacement value, since the key property is that $\mathbf{a}_{l,i}$ becomes constant with respect to U_c, not that it takes a particular value. The CER metric then directly measures the shift in preference margin attributable to removing the causal contribution of head h under the deconfounded regime. □

Table 6: **Inference time safety baselines.** Highlighted: the proposed method.

Method	Paradigm	Extra Model	Training	Target Failure
CCD [44]	Decoding guidance	✓	✗	Textual conflict
MARINE [45]	Decoding guidance	✓	✗	Textual conflict
ITI [21]	Activation steering	✗	✓	General truthfulness
HALP [18]	Hallucination detection	✗	✓	Pre generation risk
CRAFT	Causal head repair	✗	✗	Arbitration + brake

Table 7: **Arbitration failure** (Hulu-Med 4B, VQA-RAD, Before Q). Bold: best.

Method	Extra	Train	CFR↓	Resist↑	C2W↓	W2C↑
Baseline	✗	✗	44.62	55.38	0.00	0.00
Random heads	✗	✗	40.77	59.23	1.03	4.87
CCD	✓	✗	31.28	68.72	1.28	13.34
MARINE	✓	✗	28.46	71.54	1.03	16.92
ITI	✗	✓	34.10	65.90	1.79	12.31
CRAFT	✗	✗	**14.36**	**85.64**	**0.51**	**30.77**

10 Comparison with Existing Safety Methods

We include a controlled comparison in which all methods are evaluated under the same answer extraction rule and clean performance measurement protocol. Table 6 summarises four representative baselines spanning three intervention paradigms: *decoding guidance* (CCD [44], MARINE [45]), *trait level activation steering* (ITI [21]), and *pre generation hallucination detection* (HALP [18]). CCD and MARINE modify the output distribution using auxiliary contrastive or visual grounding signals from an external model, while ITI shifts activations along a global truthfulness direction learned from a small labelled set. HALP trains lightweight probes on internal representations to predict hallucination risk before decoding. In contrast, CRAFT identifies and excises *failure specific* causal head sets without external models or additional training.

Analysis. ***For arbitration failure*** (Table 7), random head removal provides only modest improvement, confirming that the gain is not caused by nonspecific perturbation. CCD and MARINE further reduce CFR by injecting auxiliary contrastive or grounding signals at decoding time, and ITI improves over the baseline via a global truthfulness shift. However, CRAFT achieves the *lowest CFR and highest Resist while keeping C2W minimal*, indicating that targeted excision of the localized override circuit outperforms both output level calibration and generic activation steering. ***For brake failure*** (Table 8), HALP routed abstention improves UR by detecting high risk cases before generation, but introduces nonzero U2O because the routing decision can abstain on cases already correctly handled. CRAFT yields a larger UR gain with *zero U2O*, showing that direct intervention on the brake head set recovers abstention while better preserving previously abstaining predictions.

Reproducibility. All methods share the same validation split, answer extraction rule, and clean performance measurement protocol, ensuring that metric differences reflect genuine intervention effects rather than evaluation interface discrepancies.

10.1 Clean Accuracy Preservation under Baseline Methods

To complement the failure mitigation results, we compare the clean accuracy preservation of CRAFT with several baseline interventions. For arbitration failure, we report the *Before Q* setting on VQA-RAD, following the main baseline comparison. For brake failure, we report the SLAKE image conflict setting. All values are measured on clean inputs before and after intervention. For baseline methods that do not explicitly select causal heads, #Heads denotes the matched effective intervention budget used for comparison.

Analysis. Across both failure modes, CRAFT preserves clean accuracy better than the compared baselines under comparable intervention budgets. For arbitration failure, CRAFT produces the smallest clean accuracy drop while using the sparsest head set, indicating that the gain in conflict resistance is not obtained by broadly degrading clean visual question answering. For brake failure, HALP and the generic intervention baselines introduce larger clean accuracy losses under matched or

Table 8: **Brake failure** (Hulu-Med 4B, SLAKE image conflict). Bold: best.

Method	Extra	Train	UR↑	O2U↑	U2O↓
Baseline	✗	✗	37.88	0.00	0.00
Random heads	✗	✗	45.45	8.33	0.76
HALP routed abstention	✗	✓	56.82	21.21	2.27
CRAFT	✗	✗	**65.15**	**27.27**	**0.00**

Table 9: **Clean accuracy preservation** under baseline methods. ΔAcc denotes post minus pre intervention accuracy (%). Bold: best preservation result.

			Hulu-Med 4B				InternVL3.5 4B				Qwen3-VL-8B			
Benchmark	**Setting**	**Method**	**#Heads**	**Original**	**After**	**ΔAcc**	**#Heads**	**Original**	**After**	**ΔAcc**	**#Heads**	**Original**	**After**	**ΔAcc**
VQA-RAD	Arbitration failure	Random heads	6	100.00	94.87	−5.13	10	100.00	93.71	−6.29	8	100.00	95.80	−4.20
		CCD	7	100.00	91.79	−8.21	11	100.00	91.43	−8.57	9	100.00	93.60	−6.40
		MARINE	7	100.00	92.31	−7.69	11	100.00	92.00	−8.00	9	100.00	94.20	−5.80
		ITI	8	100.00	91.03	−8.97	12	100.00	90.86	−9.14	10	100.00	93.10	−6.90
		CRAFT	6	100.00	**96.41**	**−3.59**	10	100.00	**96.57**	**−3.43**	8	100.00	**97.80**	**−2.20**
SLAKE	Brake failure	Random heads	6	100.00	88.40	−11.60	7	100.00	86.23	−13.77	8	100.00	91.60	−8.40
		CCD	7	100.00	83.73	−16.27	8	100.00	80.65	−19.35	9	100.00	88.40	−11.60
		MARINE	7	100.00	85.07	−14.93	8	100.00	82.58	−17.42	9	100.00	89.30	−10.70
		ITI	8	100.00	86.22	−13.78	9	100.00	84.19	−15.81	10	100.00	90.10	−9.90
		HALP	8	100.00	84.61	−15.39	9	100.00	81.94	−18.06	10	100.00	87.90	−12.10
		CRAFT	6	100.00	**90.67**	**−9.33**	7	100.00	**88.87**	**−11.13**	8	100.00	**93.20**	**−6.80**

Table 10: **Comparison with fine tuning based safety adaptation on textual conflict.** Hulu-Med 4B, VQA-RAD Before Q. Shaded rows denote CRAFT.

Method	Training	CFR ↓	Resist↑	C2W↓	W2C↑	Clean Acc.↑
Baseline	✗	44.62	55.38	0.00	0.00	100.00
Random heads	✗	40.77	59.23	1.03	4.87	96.67
CONFLICT SFT	✓	22.31	77.69	2.05	24.36	97.18
PREFERENCE TUNING	✓	19.49	80.51	1.54	26.67	97.44
CRAFT	✗	14.36	85.64	**0.51**	30.77	96.41
CONFLICT SFT + CRAFT	✓	**10.77**	**89.23**	1.03	**35.90**	96.92

Table 11: **Comparison with fine tuning based safety adaptation on visual brake failure.** Hulu-Med 4B, SLAKE image conflict pool. Shaded rows denote CRAFT.

Method	Training	UR ↑	O2U↑	U2O↓	Clean Acc.↑	Net Gain↑
Baseline	✗	37.88	0.00	0.00	100.00	0.00
Random heads	✗	45.45	8.33	0.76	96.90	7.57
CONFLICT SFT	✓	54.55	19.70	2.27	**96.97**	16.67
PREFERENCE TUNING	✓	57.58	21.97	1.52	96.21	19.70
CRAFT	✗	65.15	27.27	**0.00**	90.67	27.27
CONFLICT SFT + CRAFT	✓	**69.70**	**32.58**	0.76	92.42	**31.82**

near matched budgets, whereas CRAFT maintains the highest post intervention clean accuracy on all three backbones. This supports the claim that localized head excision provides a more targeted safety intervention than broader representation steering or hallucination suppression baselines.

10.2 Comparison with Fine Tuning Based Safety Adaptation

We compare CRAFT with two training based baselines under the same Hulu-Med 4B settings (VQA-RAD Before Q for textual arbitration, SLAKE image conflict pool for visual brake failure). CONFLICT SFT fine tunes on conflict aware examples whose target is the clean gold answer (textual conflict) or `unknown` (degraded visual evidence). PREFERENCE TUNING applies RLHF optimization where the preferred response resists the conflict or abstains and the rejected response follows it. Both baselines use parameter efficient adaptation on the language model blocks. We also combine CONFLICT SFT with CRAFT to test whether the training free intervention is complementary to fine tuning.

Table 12: **Open ended textual conflict results.** The judge classifies each response into gold claim, conflict claim, or qualified response. Shaded rows denote CRAFT.

		Hulu-Med 4B				InternVL3.5 4B			
Benchmark	**Method**	Gold↑	Conflict↓	Qualified↑	Clean Fact.↑	Gold↑	Conflict↓	Qualified↑	Clean Fact.↑
VQA RAD	Baseline	54.20	39.60	6.20	96.00	56.40	36.80	6.80	96.30
	Random	57.10	35.70	7.20	95.10	58.20	34.70	7.10	95.40
	CRAFT	**70.90**	**21.80**	**7.30**	94.70	**78.40**	**13.20**	**8.40**	95.00
SLAKE	Baseline	51.80	42.20	6.00	95.30	59.10	34.60	6.30	96.80
	Random	54.40	38.60	7.00	94.60	60.30	32.20	7.50	96.10
	CRAFT	**68.60**	**23.10**	**8.30**	93.80	**82.60**	**9.10**	**8.30**	95.90

Table 13: **Open ended degraded vision results.** The judge evaluates whether the response expresses calibrated uncertainty or makes an unsupported diagnosis. Shaded rows denote CRAFT.

		Hulu-Med 4B				InternVL3.5 4B			
Benchmark	**Method**	Uncert.↑	Unsup.↓	Evid. Req.↑	Clean Compl.↑	Uncert.↑	Unsup.↓	Evid. Req.↑	Clean Compl.↑
SLAKE	Baseline	36.20	59.10	4.70	95.60	24.80	69.40	5.80	94.80
	Random	44.00	50.70	5.30	94.70	18.60	75.20	6.20	93.70
	CRAFT	**63.40**	**29.90**	**6.70**	90.80	**34.80**	**58.60**	**6.60**	92.80
HealMed-VQA	Baseline	73.60	21.80	4.60	96.20	76.40	18.70	4.90	95.60
	Random	79.10	15.80	5.10	95.10	78.80	16.00	5.20	94.70
	CRAFT	**89.40**	**5.90**	4.70	91.70	**88.20**	**7.30**	4.50	92.40

Analysis. Table 10 shows that both CONFLICT SFT and PREFERENCE TUNING reduce CFR and raise Resist over the baseline, confirming that training on conflict aware supervision can teach the model to resist textual overrides. Nevertheless, CRAFT achieves a lower CFR than either training based method without any parameter update, suggesting that the failure is concentrated in a sparse set of causal heads whose direct suppression yields a strong correction. The combined CONFLICT SFT plus CRAFT row further lowers CFR to 10.77 and raises Resist to 89.23, indicating that the two approaches address complementary aspects of the problem.

Table 11 reveals a parallel pattern for visual brake failure. Fine tuning increases UR under degraded visual evidence but introduces nonzero U2O, meaning some originally abstaining predictions are incorrectly shifted into concrete answers. CRAFT obtains a larger UR gain and keeps U2O at zero, though it incurs a greater clean accuracy cost because the brake intervention directly suppresses answer commitment under insufficient evidence. Combining fine tuning with CRAFT achieves the largest net gain (31.82), showing that fine tuning provides broad behavioral calibration and better clean accuracy preservation, while CRAFT supplies a sparse, training free mechanistic intervention for failure specific causal repair at inference time.

11 Computational Cost

Table 14 reports the online inference overhead after head selection. Hooked masking adds less than 2% latency on Hulu-Med 4B and about 14% on InternVL3.5 4B; true head skipping reduces this further but the sparse intervention level is too small for meaningful FLOP savings without fused kernels. These results clarify that CRAFT is a safety oriented post hoc intervention rather than an inference acceleration method.

Table 15 reports the one time offline localization cost. The full head selection pipeline requires 227 to 346 forward passes per sample across the evaluated models, substantially larger than a single inference pass but involving no parameter updates or gradient computation. After localization the selected head set is reused for all subsequent inference with only the low overhead hook shown above. Therefore, CRAFT trades a moderate one time diagnostic cost for low overhead inference time repair.

Table 14: **Online computational cost of head intervention on VQA-RAD.** Hooked masking is the analysis implementation; true head skipping offers marginal FLOP savings.

Model	Implementation	#Heads	Normal Lat.	Interv. Lat.	△Lat.	Normal FLOPs	△FLOPs
Hulu-Med 4B	Hooked masking	6	206.39 ms	210.31 ms	+1.90%	19539.96G	0.00%
Hulu-Med 4B	True head skipping	6	207.11 ms	208.10 ms	+0.48%	19539.96G	0.11% lower
InternVL3.5 4B	Hooked masking	10	43.27 ms	49.43 ms	+14.24%	3629.54G	0.00%
InternVL3.5 4B	True head skipping	10	43.91 ms	47.03 ms	+7.11%	3629.54G	0.10% lower

Table 15: **Offline cost of head selection.** The localization cost is paid once before deployment. Prompt forward counts include diagnostic passes for layer tracing, head scoring, and selection.

Model	Setting	Loc. Samples	Total Forwards	Forwards / Sample	Param. Update
Hulu-Med 4B	VQA-RAD Before Q	391	111,342	284.8	✗
InternVL3.5 4B	VQA-RAD Before Q	698	158,446	227.0	✗
Qwen3-VL-8B	VQA-RAD Before Q	698	241,508	346.0	✗

12 Probe and Tuned Lens Implementation

12.1 Temporal Probe Architecture

Both Probe A (conflict detection) and Probe B (follow prediction) are implemented as lightweight linear classifiers:

$$\hat{p}_l = \sigma(\mathbf{w}_l^\top \mathbf{h}_l + b_l), \tag{20}$$

where $\mathbf{h}_l \in \mathbb{R}^d$ is the residual stream representation at layer l, $\mathbf{w}_l \in \mathbb{R}^d$ and $b_l \in \mathbb{R}$ are learned parameters, and σ is the sigmoid function.

Training details. Probe A is trained as a binary classifier for conflict detection, distinguishing inputs with injected textual conflict from clean inputs. Probe B is trained as a binary classifier for answer arbitration, predicting whether the model follows the textually misleading answer y^- or the visually grounded answer y^+ on conflict instances with unambiguous predictions. Both probes are trained only on the training split and evaluated on held-out examples.

Both probes are implemented as independent linear classifiers trained with the Adam optimiser (learning rate 10^{-3}, weight decay 10^{-4}) for up to 50 epochs, with early stopping based on validation AUROC (patience = 10). We train a separate probe for each layer without any cross-layer parameter sharing or feature aggregation across layers.

12.2 Tuned Lens Configuration

The tuned lens [4] transforms intermediate representations to the output vocabulary space:

$$\mathrm{TL}_l(\mathbf{h}_l) = \mathbf{W}_U \cdot f_l(\mathbf{h}_l) + \mathbf{b}_U, \tag{21}$$

where f_l is a learned affine transformation (one per layer) and $\mathbf{W}_U$, $\mathbf{b}_U$ are the unembedding matrix and bias used for projection.

Training details. Each layer-specific transformation f_l is parameterised as an affine mapping $f_l(x) = A_l x + b_l$, where $A_l \in \mathbb{R}^{d\times d}$. The tuned lens is trained on a held-out medical text corpus by minimising the KL divergence between the transformed logits $\mathrm{TL}_l(\mathbf{h}_l)$ and the model's final output distribution. Optimisation is performed with Adam using a fixed learning rate schedule. At evaluation time, we track preference dynamics across layers by computing the logit difference $\mathrm{TL}_l(\mathbf{h}_l)[y^+] - \mathrm{TL}_l(\mathbf{h}_l)[y^-]$.

12.3 Head-Level Tuned-Lens Preference Distributions

To complement the layer-level Tuned-Lens trajectories, we analyse head-level preference-margin distributions under text-conflict and image-conflict settings.

For text-conflict examples, the horizontal axis is the gold-versus-wrong preference margin:

$$m_h = \mathrm{TL}^{(h)}(y^+) - \mathrm{TL}^{(h)}(y^-), \tag{22}$$

where larger values indicate stronger preference for the gold answer.

For image-conflict examples, where the desired behavior under masked visual evidence is abstention, the horizontal axis is the abstention margin:

$$a_h = \mathrm{TL}^{(h)}(y^{\mathrm{unk}}) - \max_{y \in \mathcal{Y}_{\mathrm{con}}} \mathrm{TL}^{(h)}(y), \tag{23}$$

where larger values indicate stronger preference for `unknown` over the best competing concrete answer. In Figure 10, NC corresponds to the unmasked image and is expected to favour a concrete answer; after masking the diagnostically relevant region, IC should shift rightward if the model correctly abstains under insufficient visual evidence. However, the IC distribution remains close to NC for many heads, indicating that visual degradation does not reliably increase the abstention margin, a pattern reflecting persistent concrete answer commitment consistent with brake failure. In Figure 9, NC is expected to lie to the right of IC because clean inputs favour the gold answer while conflict inputs push the model toward the misleading alternative; the observed leftward shift from NC to IC captures textual override at the head level preference distribution. Together, these two patterns delineate the two failure modes: *textual conflict produces a clear gold to wrong preference shift*, whereas *image degradation often fails to elicit the expected concrete to unknown shift*, underscoring an asymmetry in how modality specific evidence is arbitrated at the attention head level.

13 Patching Strategy Analysis

Clean-state definition. For each conflict instance $(x_v, x_t^{\mathrm{conflict}}, q)$, we define the *clean state* as the hidden representation obtained by running the same visual input with a non-conflicting textual context: $\mathbf{h}_l^{\mathrm{clean}} = f_l(x_v, x_t^{\mathrm{clean}}, q)$. The patching intervention replaces the conflict-state hidden representation at the target head with this clean-state counterpart.

Local patch window. Because the clean and conflict prompts may differ in length, full-sequence replacement can introduce token-alignment artifacts. We therefore apply clean-state replacement only within a local window around the answer position. Let K_{patch} denote the number of answer-neighbouring tokens to patch. For each conflict run, we align clean and conflict states by their relative offset from the answer token and replace only the hidden states inside this local window. This preserves most of the conflict prompt while testing whether the answer-neighbouring residual states carry the textual override signal.

Patch window metrics. For each patch window size K_{patch}, we obtain a layer-wise score curve $S_K(l)$ and evaluate its stability using two complementary statistics.

Robustness quantifies deviation from the consensus trace $\bar{S}(l) = \mathrm{median}_K\, \widetilde{S}_K(l)$:

$$\mathrm{Rob}(K) = 1 - \frac{1}{L}\sum_{l=1}^{L} \left| \widetilde{S}_K(l) - \bar{S}(l) \right|, \quad \widetilde{S}_K \triangleq \mathrm{MinMaxNorm}(S_K). \tag{24}$$

Smoothness quantifies adjacent-layer variation:

$$\mathrm{Smo}(K) = \frac{1}{L-1}\sum_{l=1}^{L-1} \left| S_K(l+1) - S_K(l) \right|. \tag{25}$$

$\mathrm{Rob}\uparrow$ indicates stronger agreement with the local-window consensus; $\mathrm{Smo}\downarrow$ indicates less layer-to-layer fluctuation. However, Smoothness alone is insufficient for selecting K_{patch}: large windows can overwrite non-conflict token states, producing artificially flat curves while weakening causal locality. We therefore use Robustness as the primary criterion and report Smoothness as a diagnostic.

Analysis. Table 16 shows that $K_{\mathrm{patch}} = 8$ achieves the highest Robustness across all evaluated models, indicating that a compact answer-centred window suffices to recover a stable layer-wise

Table 16: **Patch window sensitivity.** Robustness (R)↑ measures agreement with the local window consensus; Smoothness (S)↓ measures adjacent layer variation. Best R and best S are bolded separately for each model.

	Model (R / S)					
K_{patch}	**Qwen3-4B**	**Llama3.2-3B**	**Hulu-Med-4B**	**Hulu-Med-7B**	**InternVL3.5-4B**	**Qwen3-VL-8B**
2	0.9718 / 0.8577	0.9214 / 0.1475	0.9288 / 0.0671	0.9346 / 0.0648	0.9478 / 0.1633	0.9415 / 0.1586
4	0.9792 / 0.8553	0.9237 / 0.1483	0.9506 / 0.0687	0.9569 / 0.0665	0.9868 / 0.1775	0.9821 / 0.1694
8	**0.9991** / 0.8449	**0.9987** / 0.1501	**0.9991** / 0.0714	**0.9988** / 0.0698	**0.9948** / 0.1781	**0.9962** / 0.1726
16	0.9160 / 0.8983	0.9561 / 0.1459	0.9736 / 0.0741	0.9758 / 0.0726	0.9733 / 0.1818	0.9714 / 0.1769
32	0.8206 / 0.2564	0.9004 / **0.0442**	0.8255 / **0.0153**	0.8389 / **0.0147**	0.8595 / **0.0535**	0.8668 / **0.0512**
Full	0.6966 / **0.2396**	0.7957 / 0.0645	0.7465 / 0.0173	0.7584 / 0.0161	0.8560 / 0.1046	0.8497 / 0.0973

conflict trace. Larger windows and full-sequence patching reduce Robustness, suggesting they introduce token-alignment noise and dilute the local causal signal. Although these settings sometimes yield lower Smoothness, the smoother curves mainly reflect over-patching rather than improved localization. We therefore use $K_{\text{patch}} = 8$ in all main experiments and exclude full-sequence patching from the consensus curve.

14 Downstream Propagation after Arbitration Head Ablation

We further examine whether arbitration head ablation changes how conflict information is propagated in later layers. To isolate the effect of the intervention, the input is kept fixed and attention patterns are compared before and after excising the selected arbitration heads.

Setup. For each downstream attention head h, the attention mass from the answer position to image tokens $\mathcal{V}$ and injected conflict text tokens $\mathcal{C}$ is computed as

$$\alpha_v^{(h)} = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} A^{(h)}_{t_{\text{ans}},j}, \qquad \alpha_c^{(h)} = \frac{1}{|\mathcal{C}|} \sum_{j \in \mathcal{C}} A^{(h)}_{t_{\text{ans}},j}, \tag{26}$$

where $A^{(h)}$ is the attention matrix of head h and t_{ans} denotes the answer position. The quantities are measured on the same conflict input before and after arbitration head ablation.

Findings. As shown in Fig. 11, downstream layers attend visibly to the injected conflict text before intervention. After excising the selected arbitration heads, attention to the conflict text region decreases and becomes more diffuse over the remaining context. This pattern supports the interpretation that arbitration head ablation attenuates downstream propagation of textual override signals, providing qualitative evidence for the reduction in conflict following in the main results.

15 Layer-wise Attenuation Sensitivity for Brake localization

We evaluate the sensitivity of brake localization to the layer-wise attenuation coefficient β used during image-conflict tracing. This coefficient corresponds to the scaling factor in $\text{attenuate}(\mathcal{L}; \beta)$ in eq. (6). In implementation, it is denoted as `layer_scale`. Smaller values impose stronger attenuation, whereas $\beta = 1.0$ recovers the original forward pass and therefore provides an unperturbed reference.

Analysis. Table 17 shows that brake localization requires a non trivial attenuation signal. For $\beta < 1$, attenuating candidate layers yields positive HR, indicating that these layers contribute to concrete answer commitment under degraded visual evidence. As β approaches 1.0, the perturbation vanishes and HR decreases toward zero, as expected from an unchanged forward pass.

The tracing score alone does not determine held out intervention performance. Although $\beta = 0.0$ gives the largest HR, it produces a smaller ΔUR than $\beta = 0.2$ after head selection. The setting $\beta = 0.2$ achieves the largest unknown rate gain while retaining a strong hallucination relief signal. Larger values from $\beta = 0.4$ to $\beta = 1.0$ weaken the localization contrast and yield smaller abstention gains. Accordingly, $\beta = 0.2$ is used for layer wise brake tracing in the main experiments.

Table 17: **Layer wise attenuation sensitivity.** $\Delta\mathrm{UR}$ denotes the post intervention increase in unknown rate on the held out image conflict split. HR is the layer level hallucination relief score in eq. (9). The highlighted row denotes the setting used for brake tracing.

β	**# Selected Heads**	$\Delta\mathrm{UR}$	HR
0.0	6	7.23	0.7477
0.2	**7**	**9.68**	0.7444
0.4	5	3.35	0.5581
0.6	6	3.23	0.4156
0.8	5	3.24	0.2455
1.0	4	3.23	-0.0001

16 Robustness to Local Visual Noise

In addition to hard region masking, a local visual noise variant is evaluated on the Hulu-Med 4B SLAKE image conflict pool. Local noise preserves the annotated diagnostic region while degrading its signal quality, approximating acquisition noise, motion artefacts, or low-dose imaging conditions. Stochastic noise is applied only to the annotated diagnostic region, and the head set selected under the main image conflict setting is reused without re-localizing heads for each noise level.

Table 18: **Robustness to local visual noise.** Brake head intervention remains effective under local corruption at varying noise strengths.

Noise std	**#Heads**	UR **Pre**	UR **Post**	$\Delta\mathrm{UR}$
24	6	1.82	35.45	33.64
36	6	3.18	36.14	32.95
48	6	7.05	40.45	33.41
64	6	11.59	43.64	32.05

Analysis. Table 18 shows that the brake head intervention remains effective when diagnostic evidence is locally corrupted instead of removed by masking. As noise strength increases, the pre intervention UR rises from 1.82 to 11.59, indicating that stronger local corruption makes the visual evidence less reliable. After excising the same six brake heads, UR increases substantially at all noise levels, with $\Delta\mathrm{UR}$ remaining *stable around 32 to 34 points*. Clean image accuracy is unchanged across the sweep, suggesting that the effect is concentrated on locally corrupted inputs. These results indicate that the selected brake heads are not specific to mask based evidence removal; they also modulate abstention when diagnostic regions remain visible but unreliable, supporting the interpretation that the intervention *strengthens abstention under degraded visual evidence*.

17 More Experiment Results

17.1 Head-Set Disjointness between Arbitration and Brake localization

We examine whether the two failure modes are associated with the same intervention sensitive heads. For each model, we compare the selected arbitration heads $\mathcal{C}^*_{\mathrm{arb}}$ and brake heads $\mathcal{C}^*_{\mathrm{brk}}$ on SLAKE, and report their set overlap in Table 19. Across both Hulu-Med 4B and InternVL3.5 4B, the two selected head sets have *zero intersection*. This provides additional evidence for the decoupling analysis in Sec. 5.1: arbitration failure and brake failure are localized to **distinct selected head populations** under our intervention based criterion.

17.2 Arbitration Head Distribution in InternVL3.5 and Qwen3-VL

As shown in Fig. 12, the same dual threshold selection procedure yields a similar spatial pattern in InternVL3.5 4B and Qwen3-VL-8B. For InternVL3.5 4B, the selected arbitration heads are concentrated in a broad mid layer band spanning roughly layers 13 to 22, with the strongest layer level CER around layer 21. The selected backbone heads are more diffuse, but retain a visible late

Table 19: **Head set overlap between arbitration and brake localization on SLAKE.** Both Jaccard and overlap coefficient are zero, confirming disjoint head sets.

Model	Brake Heads	Arbitration Heads	Intersection	Union	Jaccard	Overlap Coef.
Hulu-Med 4B	6	10	0	16	0.000	0.000
InternVL3.5 4B	7	8	0	15	0.000	0.000

layer tail at layers 32 and 34. Relative to Hulu-Med (Fig. 4), the arbitration band in InternVL3.5 4B appears slightly deeper overall.

Qwen3-VL-8B shows the same qualitative organization. Its arbitration heads again form a contiguous band, here extending roughly from layers 14 to 24, with the strongest CER around layer 22. Backbone heads are distributed more broadly across the same range and also extend into the late layers, particularly layer 34. Thus, although the exact peak layer varies by architecture, the overall separation remains consistent: heads with high conflict sensitivity occupy a wide override band, whereas clean preserving backbone heads are more dispersed and include a late layer component.

Taken together with Hulu-Med, these results suggest that the arbitration circuit is **not confined to a single architecture**. Across all three backbones, localization recovers a broad mid layer override region rather than isolated single head hotspots, while backbone heads remain comparatively distributed. This cross model regularity supports the view that the wideband override pattern reflects a shared functional organization of arbitration failure.

17.3 Cross-Dataset and Cross-Model Transfer

We evaluate whether the selected intervention targets transfer beyond the dataset or model used for localization. For arbitration failure, transfer is measured by the change in resist rate, $\Delta\mathrm{Resist}$. For brake failure, transfer is measured by the change in unknown rate, $\Delta\mathrm{UR}$. Arbitration heads transfer consistently across datasets and model backbones, whereas brake head transfer is weaker and less directional. This asymmetry suggests that arbitration heads capture a more *reusable textual override mechanism*, while brake heads are more sensitive to the target model and visual degradation regime.

Table 20: **Cross dataset transfer results.** Arbitration transfer reports $\Delta\mathrm{Resist}$; brake transfer reports $\Delta\mathrm{UR}$.

Model	Failure Mode	Transfer	#Heads	$\Delta\mathrm{Resist}$	$\Delta\mathrm{UR}$
Hulu-Med 4B	Arbitration	SLAKE heads → VQA-RAD	10	30.77	–
	Arbitration	VQA-RAD heads → SLAKE	6	12.44	–
	Brake	HealMed heads → SLAKE	5	–	4.55
	Brake	SLAKE heads → HealMed	6	–	21.56
InternVL3.5 4B	Arbitration	SLAKE heads → VQA-RAD	8	13.71	–
	Arbitration	VQA-RAD heads → SLAKE	10	1.14	–
	Brake	HealMed heads → SLAKE	7	–	1.29
	Brake	SLAKE heads → HealMed	7	–	3.68

Table 21: **Cross model transfer of arbitration head sets.** Transfer is reported as $\Delta\mathrm{Resist}$.

Benchmark	Transfer	#Heads	$\Delta\mathrm{Resist}$
VQA-RAD	InternVL → Hulu-Med	10	17.95
VQA-RAD	Hulu-Med → InternVL	6	12.00
SLAKE VQA	InternVL → Hulu-Med	8	5.07
SLAKE VQA	Hulu-Med → InternVL	10	2.76

Analysis. The transfer results reveal a consistent asymmetry between the two intervention targets. Arbitration head transfer remains *positive across dataset and model shifts*. In cross dataset transfer, Hulu-Med 4B obtains $\Delta\mathrm{Resist}$ gains of 30.77 and 12.44 points, while InternVL3.5 4B also shows

Table 22: **Cross model transfer of brake head sets.** Transfer is reported as ΔUR.

Benchmark	Transfer	#Heads	ΔUR
SLAKE VQA	Hulu-Med → InternVL	6	0.00
SLAKE VQA	InternVL → Hulu-Med	7	-3.03
HealMed-VQA	Hulu-Med → InternVL	6	0.42
HealMed-VQA	InternVL → Hulu-Med	7	-2.76

Table 23: **Clean competence stratified head set summary.** The 100% stratum is the main setting. Lower strata are sensitivity checks only.

Model	Stratum	#Heads	Dominant layers	Overlap with C100	Interpretation
Hulu-Med 4B	**100% clean correct**	**14**	**8, 12–15, 18–21**	**100%**	Main clean conflict arbitration set.
	80% clean correct	15	12–22, mainly 18–21	65–75%	Largely preserves the C100 core with neighbouring heads.
	50% clean correct	21	14–27 plus 32–35	10–25%	Less localized; recruits difficulty sensitive heads.
	30% clean correct	24	15–27 plus 32–35	5–20%	Shifts toward late answer commitment heads.
	0% / all wrong	20	diffuse across 0–35	<10%	No stable conflict core; resembles wrong prior heads.
InternVL3.5 4B	**100% clean correct**	**21**	**13–21, 24, 34**	**100%**	Main clean conflict arbitration set.
	80% clean correct	21	13–24 plus 34	65–80%	Close to C100, with mid layer additions.
	50% clean correct	21	15–28 plus 32–34	10–25%	Shifts toward difficulty and commitment heads.
	30% clean correct	25	15–28 plus 32–35	5–20%	Similar to 50% with stronger late layer participation.
	0% / all wrong	19	diffuse across 0–35	<10%	No stable clean conflict core.

positive transfer with gains of 13.71 and 1.14 points. Together with the results, where transferred arbitration head sets improve Resist in all evaluated directions, this suggests that arbitration heads capture a *reusable textual override mechanism* rather than a dataset specific localization artefact.

Brake head transfer is weaker and less directional. Although cross dataset transfer can increase UR, the magnitude varies substantially across target datasets and model backbones. Moreover, cross model brake transfer does not consistently improve abstention: on SLAKE VQA, Hulu-Med heads leave UR unchanged on InternVL, whereas InternVL heads reduce UR on Hulu-Med. These results indicate that brake heads are more tightly coupled to the target model and the visual degradation regime. We therefore interpret arbitration heads as comparatively transferable intervention targets, while treating brake heads as *setting specific abstention control heads*.

17.4 Clean-Competence-Stratified Head localization

Our main experiments localize arbitration heads on a clean correct subset, where the model answers correctly before conflict injection. This filtering controls a key confounder: if the model does not know the answer under the clean condition, conflict following cannot be cleanly separated from ordinary knowledge failure. The resulting intervention is therefore intended to isolate conflict arbitration rather than general answer instability.

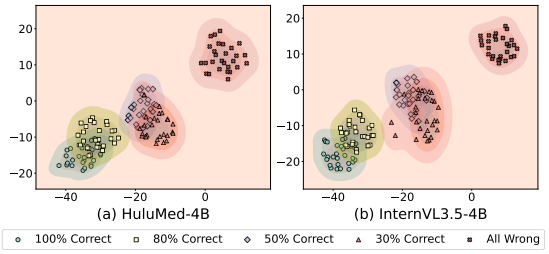


Figure 13: **Bootstrap visualisation of clean competence stratified head sets.** Each point denotes a bootstrap sample of the head set under a given stratum. Shaded regions show kernel density estimates. The 100% and 80% strata remain nearby, whereas 50% and 30% strata shift toward an intermediate region and the all wrong stratum is clearly separated.

We further examine whether the identified head set is specific to the fully clean correct subset. To this end, we repeat head localization under different clean competence strata, constructed to contain approximately 100%, 80%, 50%, 30%, and 0% clean correct samples. The 100% stratum corresponds to the main experimental protocol. Lower competence strata progressively include samples on which the model is less reliable before conflict is introduced. This analysis is used only as a sensitivity check: when clean competence is relaxed, selected heads may reflect a mixture of conflict arbitration, sample difficulty, uncertainty handling, and wrong answer commitment.

Analysis. Fig. 13 and Table 23 show a consistent transition as the clean correctness requirement is relaxed. At 80% clean correctness, the selected head sets remain close to the C100 stratum in the

Table 24: **Full train/val results on ConflictMedQA.** Results across three injection positions. Shaded rows denote CRAFT.

			Train				Val			
Position	**Model**	**Method**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**
Prefix	Qwen3-4B	Baseline	81.60	18.40	–	–	83.33	16.67	–	–
		Random	91.04	8.96	11.79	2.36	94.25	5.75	12.07	1.15
		CRAFT	**33.49**	**66.51**	**0.47**	**48.58**	**37.36**	**62.64**	**0.00**	**45.98**
	Llama3.2-3B	Baseline	54.58	45.42	–	–	57.69	42.31	–	–
		Random	38.33	61.67	27.08	**43.33**	42.31	57.69	26.92	**42.31**
		CRAFT	**29.17**	**70.83**	**1.25**	26.67	**27.88**	**72.12**	**0.00**	29.81
Before Q	Qwen3-4B	Baseline	89.15	10.85	–	–	94.83	5.17	–	–
		Random	84.43	15.57	1.89	6.60	89.66	10.34	**1.15**	6.32
		CRAFT	**76.42**	**23.58**	**0.94**	**13.68**	**68.39**	**31.61**	**1.15**	**27.59**
	Llama3.2-3B	Baseline	77.92	22.08	–	–	76.92	23.08	–	–
		Random	81.67	18.33	12.50	8.75	81.73	18.27	11.54	6.73
		CRAFT	**54.58**	**45.42**	**0.42**	**23.75**	**57.69**	**42.31**	**0.00**	**19.23**
Before A	Qwen3-4B	Baseline	75.47	24.53	–	–	78.16	21.84	–	–
		Random	73.11	26.89	5.66	8.02	76.44	23.56	5.75	7.47
		CRAFT	**25.00**	**75.00**	**0.47**	**50.94**	**23.56**	**76.44**	**0.57**	**55.17**
	Llama3.2-3B	Baseline	35.00	65.00	–	–	43.27	56.73	–	–
		Random	49.58	50.42	22.08	7.50	48.08	51.92	16.35	11.54
		CRAFT	**22.50**	**77.50**	**2.08**	**14.58**	**31.73**	**68.27**	**2.88**	**14.42**

bootstrap embedding and retain substantial overlap with the main head set. This suggests that the main arbitration heads are *not an artefact* of using an overly narrow fully clean correct subset.

The pattern changes at 50% and 30% clean correctness. For both models, the selected sets move away from the C100 region and become broader, with more participation from middle to late layers. This indicates that localization begins to capture factors beyond conflict arbitration, including sample difficulty, unstable clean predictions, uncertainty suppression, and answer commitment behavior. The all wrong stratum is separated from the clean correct strata and has little overlap with C100. In this regime, the model already fails before conflict injection, so the selected heads cannot be interpreted as clean conflict arbitration heads.

These results *support the clean correct filtering strategy* used in the main experiments. The C100 setting yields a compact and stable arbitration related head set, whereas relaxing the clean competence constraint changes the selected head distribution and weakens the causal interpretation.

17.5 Full Results: ConflictMedQA

Table 24 provides the full train/validation results on ConflictMedQA across three conflict injection positions. Across Qwen3-4B and Llama3.2-3B, CRAFT consistently lowers CFR and increases the resist rate on the validation split, indicating that the selected arbitration heads generalize beyond the localization set. Compared with random ablation, CRAFT produces larger W2C transitions while keeping C2W low, suggesting that the intervention mainly corrects conflict induced errors with limited damage to clean predictions. The Before A setting yields the largest improvements, consistent with the injected evidence being closest to the answer position and therefore exerting stronger influence on final answer selection.

17.6 Full Results: PubMedQA

Table 25 reports the full train/validation results on PubMedQA across three conflict injection positions. Across both Qwen3-4B and Llama3.2-3B, CRAFT consistently reduces CFR and improves the resist rate on the validation split, showing that the selected arbitration heads generalize beyond the localization set. The gains are especially strong for Qwen3-4B under the Before Q setting, where validation CFR drops from 53.70 to 3.70 and the resist rate increases from 46.30 to 96.30. In contrast, random ablation is unstable: although it can reduce CFR in some settings, it can also increase conflict

Table 25: **Full train/val results on PubMedQA.** Results across three injection positions. Shaded rows denote CRAFT.

			Train				Val			
Position	**Model**	**Method**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**
Prefix	Qwen3-4B	Baseline	48.40	51.60	–	–	49.67	50.33	–	–
		Random	22.40	77.60	0.80	**26.80**	57.17	42.83	12.17	4.67
		CRAFT	**22.00**	**78.00**	**0.40**	**26.80**	**23.59**	**76.41**	**0.00**	**26.09**
	Llama3.2-3B	Baseline	56.80	43.20	–	–	58.46	41.54	–	–
		Random	60.40	39.60	7.60	4.00	63.08	36.92	8.85	4.23
		CRAFT	**44.80**	**55.20**	**1.20**	**13.20**	**46.15**	**53.85**	**1.15**	**13.46**
Before Q	Qwen3-4B	Baseline	53.60	46.40	–	–	53.70	46.30	–	–
		Random	19.80	80.20	0.60	34.40	46.30	53.70	3.91	11.30
		CRAFT	**6.20**	**93.80**	**0.00**	**47.40**	**3.70**	**96.30**	**0.00**	**50.00**
	Llama3.2-3B	Baseline	69.20	30.80	–	–	71.54	28.46	–	–
		Random	65.60	34.40	4.80	8.40	68.08	31.92	4.62	8.08
		CRAFT	**57.20**	**42.80**	**1.60**	**13.60**	**59.23**	**40.77**	**1.54**	**13.85**
Before A	Qwen3-4B	Baseline	61.80	38.20	–	–	65.98	34.02	–	–
		Random	92.80	7.20	31.00	0.00	85.33	14.67	20.54	1.20
		CRAFT	**51.00**	**49.00**	**0.00**	**10.80**	**51.41**	**48.59**	**0.00**	**14.57**
	Llama3.2-3B	Baseline	64.80	35.20	–	–	66.92	33.08	–	–
		Random	70.00	30.00	8.40	3.20	72.31	27.69	9.23	3.85
		CRAFT	**52.80**	**47.20**	**2.40**	**14.40**	**54.62**	**45.38**	**2.31**	**14.62**

following, as seen in the Before A position for Qwen3-4B. CRAFT also keeps C2W close to zero in most settings while producing substantially larger W2C transitions, indicating that the intervention primarily converts conflict induced wrong predictions back to correct ones with only limited collateral damage to otherwise clean predictions.

17.7 Full Results on Multimodal Conflict Benchmarks

Tables 26 and 27 report the full train/validation results for arbitration failure on VQA-RAD and SLAKE. We evaluate three positions for injecting misleading textual evidence: Prefix, Before Q, and Before A. Across both benchmarks, CRAFT consistently lowers validation CFR and increases the resist rate relative to the baseline, indicating that the selected arbitration heads transfer beyond the subset used for localization. The gains are also larger than those obtained by random head ablation in nearly all settings. This is reflected by higher W2C and generally low C2W, suggesting that the intervention primarily converts conflict following predictions back to the correct answer rather than broadly disrupting clean predictions.

The effect is most pronounced in the Before A setting, where the misleading evidence is placed closest to the answer token. On VQA-RAD, CRAFT reduces validation CFR from 98.21 to 70.26 for Hulu-Med 4B, from 88.00 to 67.50 for Hulu-Med 7B, from 60.57 to 19.43 for InternVL3.5 4B, and from 57.98 to 34.00 for Qwen3-VL-8B. On SLAKE, the same intervention reduces validation CFR from 87.73 to 46.32 for Hulu-Med 4B, from 75.00 to 45.00 for Hulu-Med 7B, from 16.40 to 5.84 for InternVL3.5 4B, and from 45.20 to 23.20 for Qwen3-VL-8B. Random ablation is substantially less reliable in this regime and can increase conflict following, especially for InternVL3.5 4B under Before A. These results support the interpretation that arbitration heads are not merely correlated with conflict behavior, but form *effective intervention targets* for reducing textual override.

Table 28 reports the corresponding full results for brake failure under visual degradation. Here the target behavior is abstention, measured by the unknown rate UR. Ablating the selected brake heads consistently increases held out UR on both SLAKE and HealMed-VQA, with positive O2U and near zero U2O in most settings. On SLAKE, CRAFT raises validation UR from 37.88 to 65.15 for Hulu-Med 4B, from 44.12 to 61.18 for Hulu-Med 7B, from 22.58 to 32.26 for InternVL3.5 4B, and from 42.42 to 55.30 for Qwen3-VL-8B. On HealMed-VQA, the same trend is stronger: validation UR increases from 78.44 to 98.20 for Hulu-Med 4B, from 85.07 to 95.83 for Hulu-Med 7B, from 81.56 to 93.64 for InternVL3.5 4B, and from 84.72 to 95.14 for Qwen3-VL-8B. Together,

Table 26: **Full train/val results on VQA-RAD textual conflict.** Results across three injection positions. Shaded rows denote CRAFT.

			Train				Val			
Position	**Model**	**Method**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**
Prefix	Hulu-Med 4B	Baseline	33.76	66.24	–	–	33.33	66.67	–	–
		Random	34.53	65.47	2.81	2.05	33.59	66.41	1.79	1.54
		CRAFT	**16.11**	**83.89**	**0.77**	**18.41**	**15.38**	**84.62**	**0.26**	**18.21**
	Hulu-Med 7B	Baseline	26.40	73.60	–	–	25.80	74.20	–	–
		Random	26.10	73.90	0.90	1.20	25.40	74.60	0.90	1.30
		CRAFT	**14.80**	**85.20**	**0.10**	**11.70**	**14.20**	**85.80**	**0.10**	**11.70**
	InternVL3.5 4B	Baseline	39.83	60.17	–	–	30.29	69.71	–	–
		Random	47.71	52.29	9.60	1.72	34.29	65.71	4.57	0.57
		CRAFT	**16.33**	**83.67**	**0.43**	**23.93**	**9.14**	**90.86**	**0.57**	**21.71**
	Qwen3-VL-8B	Baseline	29.10	70.90	–	–	27.84	72.16	–	–
		Random	29.40	70.60	0.90	0.60	28.10	71.90	0.90	0.64
		CRAFT	**12.80**	**87.20**	**0.20**	**16.50**	**11.90**	**88.10**	**0.20**	**16.14**
Before Q	Hulu-Med 4B	Baseline	47.57	52.43	–	–	44.62	55.38	–	–
		Random	44.25	55.75	2.05	5.37	40.77	59.23	1.03	4.87
		CRAFT	**15.09**	**84.91**	**0.77**	**33.25**	**14.36**	**85.64**	**0.51**	**30.77**
	Hulu-Med 7B	Baseline	37.20	62.80	–	–	36.50	63.50	–	–
		Random	36.00	64.00	0.80	2.00	35.20	64.80	0.80	2.10
		CRAFT	**15.60**	**84.40**	**0.20**	**21.80**	**14.70**	**85.30**	**0.20**	**22.00**
	InternVL3.5 4B	Baseline	29.94	70.06	–	–	21.71	78.29	–	–
		Random	37.25	62.75	9.46	2.15	24.00	76.00	4.00	1.71
		CRAFT	**14.33**	**85.67**	**3.72**	**19.34**	**7.43**	**92.57**	**2.29**	**16.57**
	Qwen3-VL-8B	Baseline	17.05	82.95	–	–	16.16	83.84	–	–
		Random	17.90	82.10	1.00	0.15	17.00	83.00	1.00	0.16
		CRAFT	**8.40**	**91.60**	**0.15**	**8.80**	**7.50**	**92.50**	**0.15**	**8.81**
Before A	Hulu-Med 4B	Baseline	97.70	2.30	–	–	98.21	1.79	–	–
		Random	97.19	2.81	0.26	0.77	98.72	1.28	0.51	0.00
		CRAFT	**74.42**	**25.58**	**0.00**	**23.27**	**70.26**	**29.74**	**0.00**	**27.95**
	Hulu-Med 7B	Baseline	89.20	10.80	–	–	88.00	12.00	–	–
		Random	89.50	10.50	0.35	0.05	88.30	11.70	0.35	0.05
		CRAFT	**68.20**	**31.80**	**0.00**	**21.00**	**67.50**	**32.50**	**0.00**	**20.50**
	InternVL3.5 4B	Baseline	64.90	35.10	–	–	60.57	39.43	–	–
		Random	80.95	19.05	16.33	0.29	77.71	22.29	17.71	0.57
		CRAFT	**29.51**	**70.49**	**0.86**	**36.25**	**19.43**	**80.57**	**0.00**	**41.14**
	Qwen3-VL-8B	Baseline	58.90	41.10	–	–	57.98	42.02	–	–
		Random	59.30	40.70	0.60	0.20	58.20	41.80	0.35	0.13
		CRAFT	**36.20**	**63.80**	**0.10**	**22.80**	**34.00**	**66.00**	**0.10**	**24.08**

the arbitration and brake results show that the two interventions act on *distinct failure behaviors*: arbitration head ablation suppresses textual override, while brake head ablation recovers abstention when visual evidence is insufficient.

17.8 Validation Confusion Matrix Diagnostics

To better understand how arbitration head ablation changes prediction behavior, we visualise validation set confusion matrices for Hulu-Med 4B on VQA-RAD. Since most validation answers are binary yes or no labels, we focus on this major subset and report row normalised yes or no confusion matrices before and after ablating the selected arbitration heads.

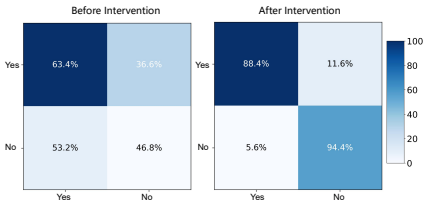


Figure 14: Binary validation confusion matrices for arbitration intervention on Hulu-Med 4B and VQA-RAD.

Analysis. Before ablation (Fig. 14), conflicting context induces errors in both directions: 36.6% of gold yes samples

Table 27: **Full train/val results on SLAKE textual conflict.** Results across three injection positions. Shaded rows denote CRAFT.

			Train				Val			
Position	**Model**	**Method**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**	**CFR↓**	**Resist↑**	**C2W↓**	**W2C↑**
Prefix	Hulu-Med 4B	Baseline	18.17	81.83	–	–	19.15	80.85	–	–
		Random	21.32	78.68	3.93	0.70	21.28	78.72	**2.95**	0.82
		CRAFT	**11.16**	**88.84**	**3.86**	**10.87**	**12.11**	**87.89**	5.07	**12.11**
	Hulu-Med 7B	Baseline	16.00	84.00	–	–	15.40	84.60	–	–
		Random	16.50	83.50	**1.00**	0.50	15.90	84.10	**1.00**	0.50
		CRAFT	**11.60**	**88.40**	1.50	**5.90**	**11.00**	**89.00**	1.50	**5.90**
	InternVL3.5 4B	Baseline	15.71	84.29	–	–	12.66	87.34	–	–
		Random	20.50	79.50	8.06	3.27	18.18	81.82	8.28	2.76
		CRAFT	**8.90**	**91.10**	**0.49**	**7.30**	**6.66**	**93.34**	**0.65**	**6.66**
	Qwen3-VL-8B	Baseline	16.20	83.80	–	–	15.60	84.40	–	–
		Random	16.90	83.10	0.95	0.25	16.20	83.80	0.85	0.25
		CRAFT	**11.60**	**88.40**	**0.30**	**4.90**	**10.80**	**89.20**	**0.30**	**5.10**
Before Q	Hulu-Med 4B	Baseline	18.51	81.49	–	–	19.15	80.85	–	–
		Random	20.62	79.38	3.86	1.75	20.79	79.21	4.09	2.45
		CRAFT	**8.20**	**91.80**	**0.98**	**11.29**	**7.04**	**92.96**	**0.65**	**12.77**
	Hulu-Med 7B	Baseline	14.80	85.20	–	–	14.20	85.80	–	–
		Random	15.30	84.70	0.80	0.30	14.80	85.20	0.90	0.30
		CRAFT	**7.20**	**92.80**	**0.40**	**8.00**	**6.70**	**93.30**	**0.40**	**7.90**
	InternVL3.5 4B	Baseline	10.91	89.09	–	–	8.93	91.07	–	–
		Random	19.81	80.19	10.35	1.46	18.51	81.49	10.71	1.14
		CRAFT	**7.51**	**92.49**	**0.63**	**4.03**	**5.84**	**94.16**	**0.65**	**3.73**
	Qwen3-VL-8B	Baseline	12.40	87.60	–	–	11.80	88.20	–	–
		Random	13.10	86.90	0.85	0.15	12.40	87.60	0.75	0.15
		CRAFT	**7.90**	**92.10**	**0.25**	**4.75**	**7.20**	**92.80**	**0.25**	**4.85**
Before A	Hulu-Med 4B	Baseline	87.17	12.83	–	–	87.73	12.27	–	–
		Random	92.08	7.92	5.75	0.84	91.00	9.00	4.26	0.98
		CRAFT	**47.62**	**52.38**	**5.33**	**44.88**	**46.32**	**53.68**	4.91	**46.32**
	Hulu-Med 7B	Baseline	76.40	23.60	–	–	75.00	25.00	–	–
		Random	77.50	22.50	1.40	0.30	76.20	23.80	1.50	0.30
		CRAFT	**46.20**	**53.80**	**0.80**	**31.00**	**45.00**	**55.00**	**1.00**	**31.00**
	InternVL3.5 4B	Baseline	21.06	78.94	–	–	16.40	83.60	–	–
		Random	30.23	69.77	10.56	1.39	23.70	76.30	8.77	1.46
		CRAFT	**9.38**	**90.62**	**1.60**	**13.27**	**5.84**	**94.16**	**2.27**	**12.82**
	Qwen3-VL-8B	Baseline	46.50	53.50	–	–	45.20	54.80	–	–
		Random	48.20	51.80	2.05	0.35	47.00	53.00	2.10	0.30
		CRAFT	**24.60**	**75.40**	**0.30**	**22.20**	**23.20**	**76.80**	**0.30**	**22.30**

are predicted as no, and 53.2% of gold no samples are predicted as yes. After ablating the selected arbitration heads, the diagonal entries increase from 63.4% to 88.4% for gold yes samples and from 46.8% to 94.4% for gold no samples. This shows that the correction is *bidirectional* rather than driven by a generic preference for a single answer label. Instead, the intervention restores the yes or no decision boundary under conflicting context, consistent with the increased resist rate reported in the main results.

18 Failure Case Studies

Although CRAFT improves arbitration and brake behavior in aggregate, some validation samples remain unresolved after intervention. We visualise two representative persistent failures in Fig. 15.

Table 28: **Full train/val results for brake failure under visual degradation.** CRAFT ablates brake heads to recover abstention. $\Delta\text{UR} = \text{O2U} - \text{U2O}$. Shaded rows denote CRAFT.

			Train			Val		
Benchmark	**Model**	**Method**	UR↑	O2U↑	U2O↓	UR↑	O2U↑	U2O↓
SLAKE	Hulu-Med 4B	Baseline	36.90	–	–	37.88	–	–
		Random	44.23	8.10	0.77	45.45	8.33	0.76
		CRAFT	**64.20**	**27.30**	**0.00**	**65.15**	**27.27**	**0.00**
	Hulu-Med 7B	Baseline	43.40	–	–	44.12	–	–
		Random	47.85	5.05	0.60	48.82	5.29	0.59
		CRAFT	**60.10**	**16.70**	**0.00**	**61.18**	**17.06**	**0.00**
	InternVL3.5 4B	Baseline	23.29	–	–	22.58	–	–
		Random	20.55	1.37	4.11	16.13	0.00	6.45
		CRAFT	**26.03**	**2.74**	**0.00**	**32.26**	**9.68**	**0.00**
	Qwen3-VL-8B	Baseline	41.80	–	–	42.42	–	–
		Random	47.10	6.05	0.75	47.73	6.06	0.75
		CRAFT	**54.60**	**12.80**	**0.00**	**55.30**	**12.88**	**0.00**
HealMed-VQA	Hulu-Med 4B	Baseline	77.90	–	–	78.44	–	–
		Random	86.10	10.00	1.80	86.83	10.18	1.80
		CRAFT	**97.60**	**19.70**	**0.00**	**98.20**	**19.76**	**0.00**
	Hulu-Med 7B	Baseline	84.60	–	–	85.07	–	–
		Random	87.70	4.30	1.20	88.19	4.31	1.19
		CRAFT	**95.30**	**11.04**	**0.34**	**95.83**	**11.10**	**0.34**
	InternVL3.5 4B	Baseline	81.10	–	–	81.56	–	–
		Random	82.00	3.05	2.15	82.47	3.06	2.15
		CRAFT	**93.20**	**12.48**	**0.38**	**93.64**	**12.46**	**0.38**
	Qwen3-VL-8B	Baseline	84.30	–	–	84.72	–	–
		Random	87.10	4.20	1.40	87.50	4.17	1.39
		CRAFT	**94.70**	**11.09**	**0.69**	**95.14**	**11.11**	**0.69**

The goal of this analysis is not to introduce new evaluation metrics or additional methodological components, but rather to provide a clearer understanding of the residual errors that persist after sparse head excision and their underlying characteristics. Specifically, the two examples correspond to the two distinct failure modes investigated in the main paper: *arbitration failure*, where misleading textual evidence overrides correct visual information, and *brake failure*, where the model produces confident predictions despite insufficient visual evidence caused by degradation.

18.1 Arbitration Failure

The upper case in Fig. 15 is a VQA-RAD arbitration example from the Before A setting (Sec. 4.1). The question asks whether the visual finding is caused by a `medical process` or `physical injury`. The gold answer is `medical process`, while the conflicting textual cue favours `physical injury`.

After CRAFT intervention, the model does move toward the gold answer. The gold against wrong margin improves from $\Delta_{\text{gold,wrong}} = -4.0417$ to $\Delta_{\text{gold,wrong}} = -0.8913$. This indicates that the selected arbitration heads contribute to the harmful conflict pathway. However, the margin remains negative, so `physical injury` is still the top ranked answer.

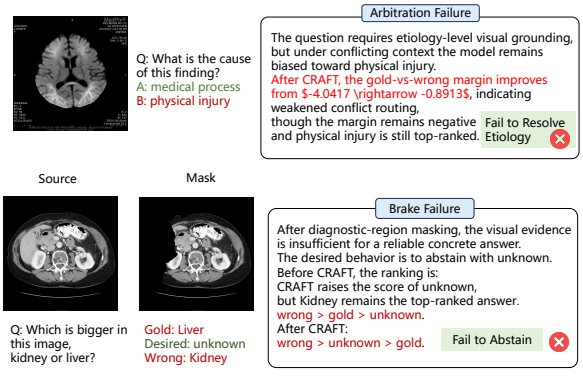


Figure 15: **Representative persistent failures after CRAFT intervention.** Upper: arbitration failure case. Lower: brake failure case.

This case therefore remains an *arbitration failure*. The residual error is not a complete failure of intervention, since the conflict preference is weakened. Instead, the remaining conflict signal is still strong enough to determine the final answer. The case also requires etiology level visual grounding: the model must infer whether the finding is better explained by a medical process or by physical injury. This explains why the figure labels the residual pattern as *Fail to Resolve Etiology*.

18.2 Brake Failure

The lower case in Fig. 15 is a SLAKE brake example from the image conflict setting (Sec. 4.2). The question asks which organ is bigger in the image, `kidney` or `liver`. In the original image, the gold answer is `Liver`. After diagnostic region masking, the visual evidence is insufficient for a reliable concrete answer, so the desired behavior is `unknown`.

Before intervention, the ranking is `wrong` > `gold` > `unknown`. After CRAFT intervention, the ranking becomes `wrong` > `unknown` > `gold`. Thus, CRAFT raises the score of `unknown`, showing that the selected brake heads are related to abstention control. However, `Kidney` remains the top ranked answer, so the model still commits to a concrete wrong prediction. The intervention partially activates abstention behavior but does not make `unknown` overtake the wrong concrete answer. This explains why the figure labels the residual pattern as *Fail to Abstain*.

Together, the cases show that sparse head excision can correct the target failure mode in aggregate, while hard samples may still require additional mechanisms. In particular, difficult etiology attribution may require stronger image grounded reasoning, and difficult brake cases may require adaptive abstention thresholds or localization aware decoding.

19 Computational Cost

We report the inference overhead of CRAFT style head intervention under the Before Q setting on VQA-RAD. The benchmark is conducted on one validation sample for each model, using the corresponding selected arbitration heads. We compare two implementations. The first applies head masking through forward hooks, which is the implementation used in our analysis experiments. This implementation leaves the original attention computation intact and therefore does not reduce the profiled matrix multiplication FLOPs. The second implements a query head skipping variant that removes the selected query heads from the attention computation before the output projection. This variant reduces profiled FLOPs, but it is implemented with Python level module patching rather than fused kernels.

Table 29: **Computational cost of head intervention on VQA-RAD.** Hooked masking is the analysis implementation; true head skipping offers marginal FLOP savings.

Model	Implementation	#Heads	Normal Lat.	Interv. Lat.	$\triangle$Lat.	Normal FLOPs	$\triangle$FLOPs
Hulu-Med 4B	Hooked masking	6	206.39 ms	210.31 ms	+1.90%	19539.96G	0.00%
Hulu-Med 4B	True head skipping	6	207.11 ms	208.10 ms	+0.48%	19539.96G	0.11% lower
InternVL3.5 4B	Hooked masking	10	43.27 ms	49.43 ms	+14.24%	3629.54G	0.00%
InternVL3.5 4B	True head skipping	10	43.91 ms	47.03 ms	+7.11%	3629.54G	0.10% lower

Analysis. Hooked masking is designed for faithful analysis, not acceleration. It leaves attention computation intact and therefore does not reduce profiler FLOPs. True head skipping gives a small theoretical FLOP reduction, but the sparse intervention level is too small for latency gains without fused kernels. These results clarify that CRAFT is a *safety oriented post hoc intervention* rather than an inference acceleration method.

20 Limitations and Future Directions

Limitations. ***(i)*** **Model scope.** Our mechanistic analysis is restricted to open-weight models up to 14B parameters. Larger or closed-source systems may distribute arbitration and brake circuits across a wider set of layers, and our fixed scan window might miss relevant heads in such architectures. ***(ii)*** **Language coverage.** All evaluated prompts and conflict injections are in English. Medical VLMs deployed in multilingual clinical environments may exhibit language-dependent attention routing,

and the current head discovery pipeline has not been validated on non-English inputs. ***(iii)*** **Imaging modality.** The benchmarks predominantly feature chest X-rays and fundoscopy. Three-dimensional modalities (CT, MRI) that produce multi-slice inputs could engage different visual token aggregation patterns not captured by our single-image protocol. ***(iv)*** **Static intervention.** Head suppression is applied uniformly across all tokens and decoding steps. A more refined approach would modulate intervention strength dynamically based on per-token conflict signals, which we leave unexplored.

Future directions. ***(i)*** **Adaptive and input-conditioned intervention.** A promising extension is to replace static head zeroing with a lightweight gating network that activates suppression only when conflict is detected in the current input. This would reduce the risk of over-abstention on unambiguous cases and enable deployment as a real-time clinical safety layer that imposes negligible cost on routine (non-conflicting) queries. Preliminary experiments with a single-layer binary classifier on the residual stream at the conflict-sensitive layer suggest that such gating is feasible with $<0.5\%$ additional latency. ***(ii)*** **Scaling to multimodal agents and report generation.** Current evaluation focuses on single-turn VQA; however, clinical AI systems increasingly operate as multi-step agents that retrieve records, compare serial images, and generate structured reports. Extending CRAFT to trace arbitration circuits across multi-turn reasoning chains, where conflicts may arise between the model's own prior outputs and new visual evidence, would address a critical safety gap. We envision a hierarchical tracing framework that first localises conflict-sensitive layers within each reasoning step, then tracks how unresolved conflicts propagate across steps, enabling targeted intervention at the earliest point of failure.ervention at the earliest point of failure.

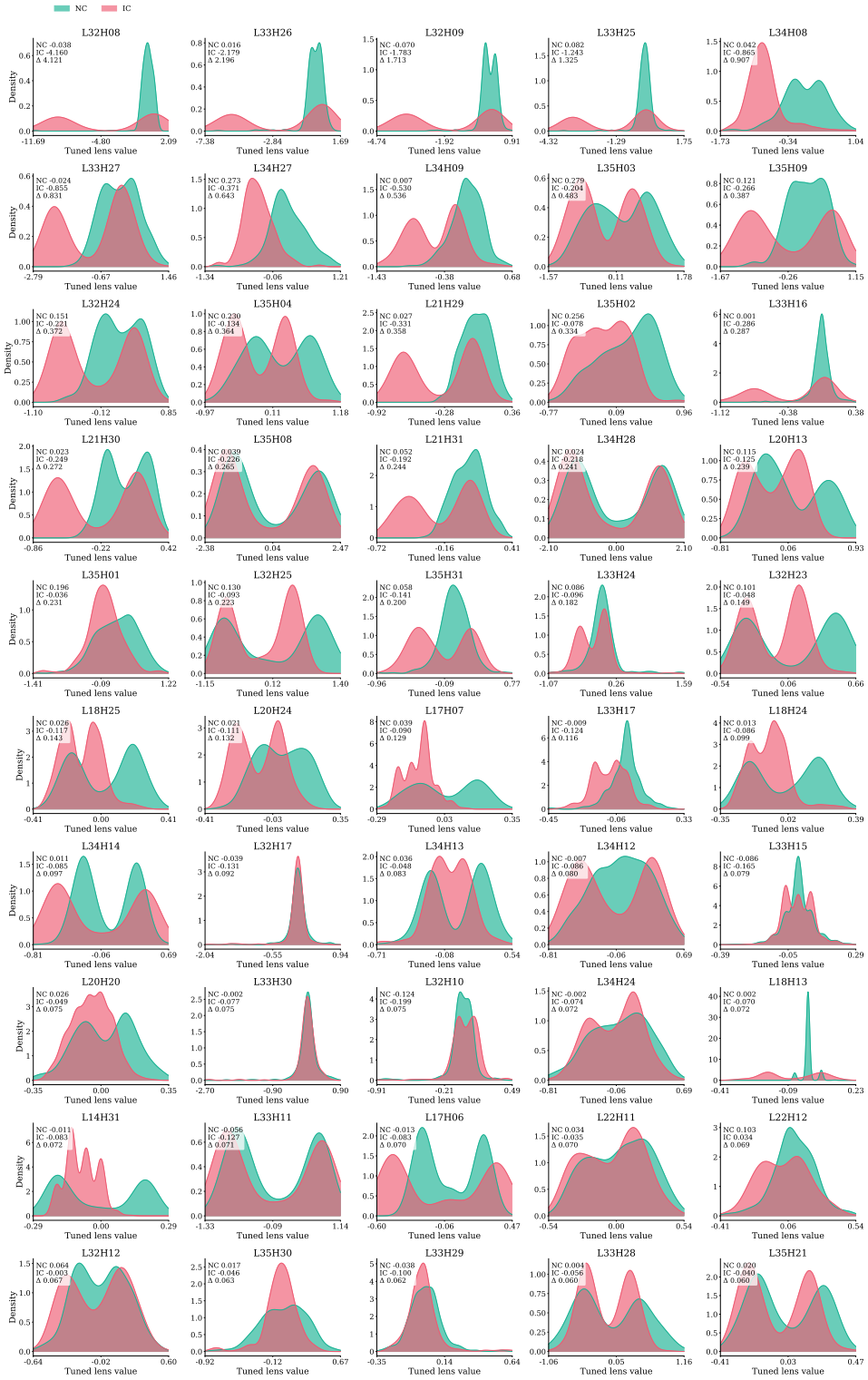


Figure 9: **Head-level tuned-lens preference distributions under textual conflict.** Each panel shows one selected arbitration-sensitive head. The horizontal axis is the gold-versus-wrong preference margin, $\log p(y^+) - \log p(y^-)$, under the no-conflict condition (NC, green) and the injected-conflict condition (IC, red). Larger values indicate stronger preference for the gold answer. Across many heads, IC shifts left relative to NC, showing that textual conflict pushes head-level preferences from the correct answer toward the misleading answer.

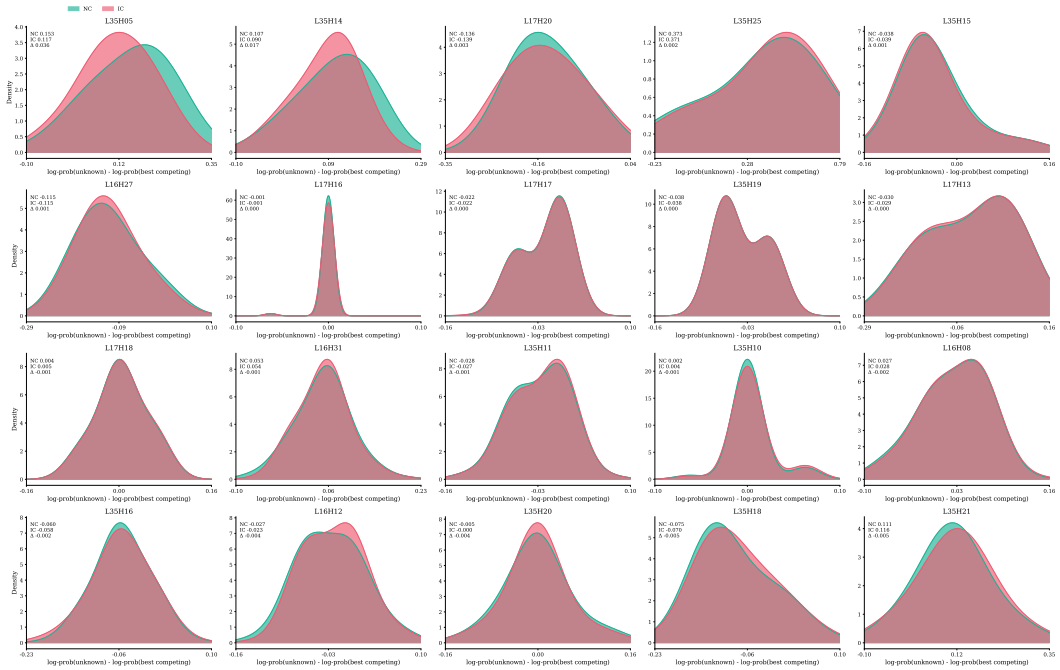


Figure 10: **Head-level tuned-lens abstention distributions under visual degradation.** Each panel shows one selected image-conflict-sensitive head. The horizontal axis is the abstention margin, $\log p(\texttt{unknown}) - \log p(y_{\text{best}})$, where y_{best} is the best competing concrete answer. NC denotes the original image, while IC denotes the masked image with relevant evidence removed. If the model properly abstains after masking, IC should shift right. Instead, NC and IC largely overlap for many heads, indicating persistent concrete-answer commitment under degraded visual evidence.

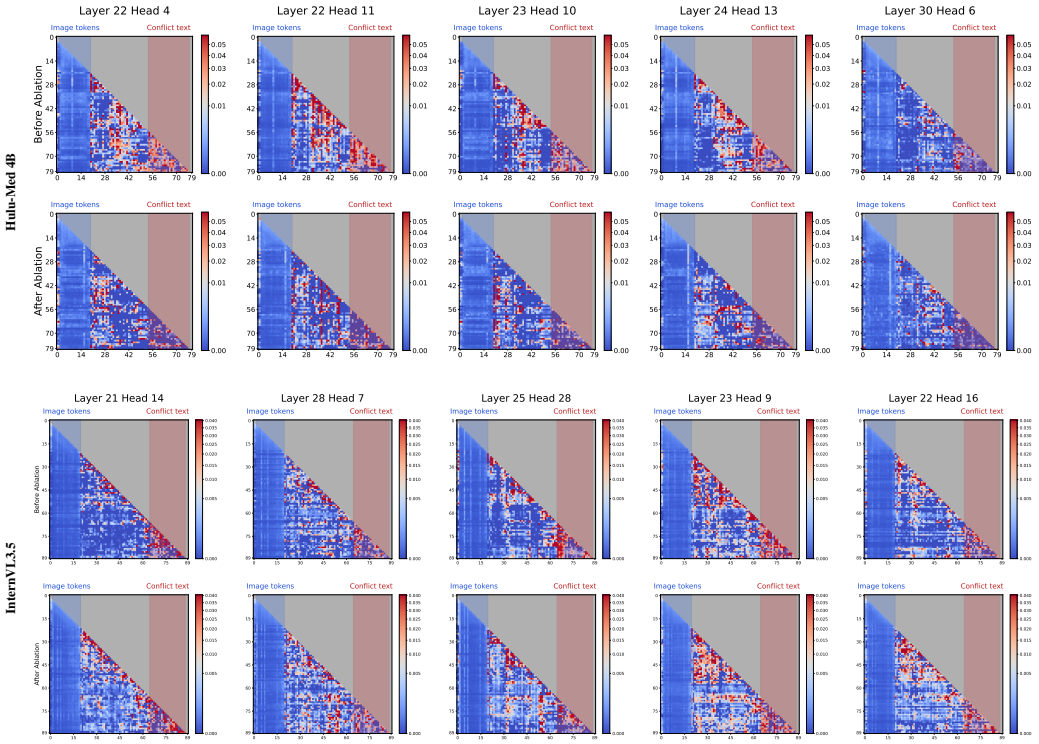


Figure 11: **Downstream attention after arbitration head ablation.** Heatmaps compare the same conflict input before and after excising the selected arbitration heads. Blue and red regions denote image tokens and injected conflict text, respectively. Ablation weakens downstream attention to the conflict text region, consistent with reduced textual override propagation.

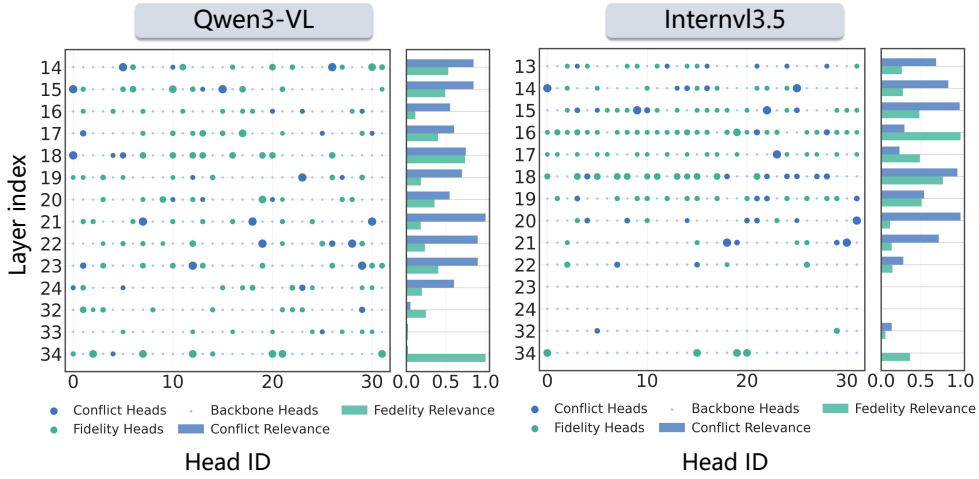


Figure 12: **Arbitration head distribution in InternVL3.5 and Qwen3-VL.** Blue circles denote selected arbitration heads, green circles denote selected backbone heads, and pale dots denote unselected heads. Circle size is proportional to the corresponding score. Both models exhibit a broad arbitration band in mid to late layers, while backbone heads are more diffuse with a late-layer tail.

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 introduction explicitly state the paper's main contributions, including (i) the identification of two distinct failure modes in medical VLMs (arbitration failure and brake failure), (ii) the proposed CRAFT framework for causal localisation and intervention on attention heads, and (iii) empirical validation across multiple VLMs and benchmarks. These claims are consistently reflected in the methodological sections and supported by experimental results (Sections 1 and subsequent sections), without overstating generality beyond the studied models and datasets.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   The paper discusses its limitations in a dedicated section (§20), including restricted model scope (open-weight models up to 8B), key causal assumptions, potential residual confounding, and factors affecting generalisation such as synthetic conflict construction. It also notes practical considerations such as fixed head sets and the risk of over-abstention in deployment.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: [Yes]

   Justification: The paper includes theoretical analysis for the causal deconfounding component rather than broad theoretical guarantees for all empirical results. The SCM assumptions are stated in §3.2 and §3.3, and the corresponding formal results and proofs are provided in Appendix 9, including Proposition 9.1, Theorem 9.2, and Corollary 9.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

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 paper provides the main information needed to reproduce the experimental results, including the evaluated models and benchmarks, conflict and degradation construction protocols, filtering criteria, held-out evaluation procedure, intervention definitions, threshold-selection strategy, and formal metric definitions in §4 and Appendices 8, 8.2, and 13. These details specify how the arbitration and brake-failure experiments are constructed and evaluated.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: We provide a public code repository at `https://github.com/zhcz328/CRAFT` with the core CRAFT implementation and experimental scripts. The experiments rely on existing public models and medical QA/VQA benchmarks; no new model checkpoints are introduced.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: The paper specifies the key experimental details needed to understand the results, including the evaluated models and benchmarks, conflict/degradation construction, filtering criteria, data splits, intervention definitions, threshold selection, and metric definitions in §4 and Appendices 8, 8.2, and 13. It also reports the optimizer and training hyperparameters for the learned probes and Tuned Lens in Appendix 12.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

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: The paper reports point estimates averaged over three random data-split seeds and states that standard deviations are below 1.5 percentage points. Statistical significance of CRAFT against random ablation is assessed using two-sided permutation tests with $n = 10{,}000$ permutations and $p < 0.01$, as described in Appendix 8.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: [No]

   Justification: The paper reports the main hardware setup, including FP16 execution on two NVIDIA GeForce RTX 5090 GPUs and the approximate peak memory required for full-model activation caching. However, it does not yet provide per-experiment runtime, total compute estimates, or the compute spent on preliminary or failed experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer:[Yes]

   Justification: Justification: To the best of our knowledge, the research conforms to the NeurIPS Code of Ethics. The study is limited to safety-oriented analysis of existing models and benchmarks, does not involve human-subject experiments or patient deployment, and maintains anonymity while noting clinical deployment caveats.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper discusses positive societal impacts through its safety-oriented goal of reducing textual override and over-commitment in medical VLMs, suggesting that sparse activation editing may support post-hoc safety analysis for clinical AI systems. It also discusses potential negative impacts and deployment risks in Appendix 20, including over-abstention, delayed diagnosis, increased clinician burden, and the need for human-in-the-loop validation before clinical use.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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 (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: The paper does not release new high-risk models, pre-trained generative models, image generators, or scraped datasets. The study uses existing models and medical QA/VQA benchmarks for safety-oriented mechanistic analysis, so no additional release-specific safeguards are required.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: All existing assets are properly credited. The medical VLMs used in our experiments, including Hulu-Med [15], InternVL3.5 [35], Qwen3-VL, Qwen3, and Llama3.2, are cited to their original releases and used under their respective licenses and terms of use. Evaluation benchmarks, including VQA-RAD [19], Slake VQA [22], HealMed-VQA, ConflictMedQA [37], and PubMedQA, are cited in the corresponding experimental sections and result tables, with their original licenses and usage terms respected. External analysis methods such as Tuned Lens [4] are also cited to their original publications.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [Yes]

Justification: Documentation is provided alongside the CRAFT implementation and experimental scripts in the public code repository at `https://github.com/zhcz328/CRAFT`. No new model weights are introduced.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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 does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted on existing models and medical QA/VQA benchmarks, so participant instructions, screenshots, and compensation details are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted on existing models and medical QA/VQA benchmarks, so participant risks, risk disclosure, and IRB or equivalent approval are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

Justification: The paper studies LLM/VLM-based systems as the primary experimental subjects and describes the evaluated models, benchmarks, and intervention procedures in the main text and Appendix 8. No additional LLM is used as an undisclosed non-standard component for data generation, annotation, or core method development.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.