CARAE: MATCHED COUNTERFACTUAL TRANSITION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual supervision can make a model respond to an edited input, but the response may reflect nuisance variation introduced by the edit rather than the intended task change. A targeted counterfactual alone provides no reference for this ambiguity. Matched controls turn absolute response fitting into a relative comparison of target and non-target transitions within the same input. CARAE instantiates this idea with a question-conditioned visual state, a relative transition objective, and question-conditioned token-weight-modulated residual coupling to a vision–language model's generation pathway. Expert localization in chest radiographs permits controlled target and non-target edits; the masks construct interventions offline and are never model inputs. On the canonical CheXlocalize protocol, CARAE raises target-versus-control answer AUROC from 0.5133 to 0.8038 at 3B scale and from 0.7053 to 0.8527 at 7B scale. The 7B paired improvement is 0.1474 (95% patient-bootstrap interval [0.0737, 0.2266]). Frozen checkpoints retain an advantage under held-out control constructions and replacement operators; dataset transfer within chest radiography is moderate on MS-CXR and weak in absolute terms on VinDr-CXR. These results support intervention-specific response under the evaluated construction family, without treating synthetic removal as a clinically adjudicated diagnosis.

1 INTRODUCTION

Counterfactual learning asks a model to change its behavior when a task-relevant part of an input changes. Yet an edit changes more than its intended target: boundary texture, location, intensity, and surrounding context may also shift. A model can satisfy counterfactual labels by responding to these correlated changes. Sensitivity to a counterfactual is therefore an incomplete test of whether the response is specific to the intended intervention.

Let x be a factual input, x_t a targeted edit, and x_c a matched control edit that preserves the target while introducing a comparable perturbation. The transitions $x \to x_t$ and $x \to x_c$ share an origin and an intervention family. The relevant behavioral comparison is the response to x_t relative to x_c, rather than the response to x_t alone. A matched control supplies an operational reference for generic edit sensitivity; its validity depends on how well the interventions are matched.

Output-level supervision can reward the desired answer ordering without requiring a model's internal state to distinguish the two transitions. A generic residual adapter may likewise improve fitting while leaving this distinction implicit. We therefore study whether organizing a question-conditioned representation around the relative target and control transitions improves the resulting response.

CARAE learns a visual state from the image and question, encourages the factual-to-target transition to differ from the factual-to-control transition through Evidence-Specific Transition Learning (EST), and couples the state to native generation. A sample-specific residual direction and scalar gate define the update, while the learned question-conditioned token weights modulate its strength at each visual token. These weights are not mask-supervised and are not interpreted as lesion locations.

Chest radiography supplies a controlled instantiation because expert finding annotations allow localized target edits and approximately matched non-target edits in the same image. Annotation masks serve only to construct and audit interventions offline. They are not given to CARAE, and learned to-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ken weights are not asserted to recover the expert masks. A synthetic target edit is a protocol-defined perturbation, not a clinically adjudicated negative radiograph.

We first test whether conventional response training distinguishes target from control edits, then compare question-conditioned coupling and EST at 3B and 7B scales. We audit recorded intervention shortcuts, evaluate frozen checkpoints under unseen control constructions and replacement operators, and test transfer to MS-CXR and VinDr-CXR. The canonical 7B improvement over response-only training is 0.1474 AUROC with a paired patient-bootstrap interval of [0.0737, 0.2266]; EST adds a smaller 0.0252 [0.0021, 0.0556] over coupling without EST. The evidence supports a relative response within the tested intervention families and leaves broader task generalization open.

Our contributions are:

- We formulate the ambiguity between intended counterfactual response and sensitivity to the intervention realization, using a matched control as the reference.
- We instantiate matched counterfactual transition learning in CARAE through question-conditioned states, EST, and question-conditioned token-modulated coupling to generation.
- We evaluate target-versus-control behavior across model scales, measured nuisance adjustments, held-out constructions and operators, and two external radiograph datasets.

2 Related Work

Counterfactual learning. Counterfactually augmented data and robustness studies use targeted changes to expose shortcut behavior Kaushik et al. (2020); Geirhos et al. (2020); Sagawa et al. (2020). Such work motivates controlled edits, but an edited sample can carry both the intended change and artifacts of its construction. CARAE focuses on the resulting ambiguity in learned response and uses a matched intervention as a within-input reference.

Visual counterfactuals. Counterfactual VQA and CF-VLM-style objectives aim to reduce language priors or promote responses to visual changes Chen et al. (2020); Zhang et al. (2025). Our distinction is the explicit comparison of factual-to-target and factual-to-control transitions in a question-conditioned state, coupled to generation. We do not assert that earlier methods never use controls; the comparison here concerns the particular transition objective and evaluation protocol.

Medical VLMs and grounding. Medical question-answering benchmarks and VLMs establish the application setting Lau et al. (2018); Liu et al. (2021); He et al. (2020); Li et al. (2023); Moor et al. (2023); Wu et al. (2025); Chen et al. (2024). Grounding methods associate findings with regions Deng et al. (2024); Müller et al. (2024); Ouyang et al. (2021), while intervention studies probe whether answers change under image edits Xiong et al. (2026); Zafar et al. (2026). CARAE is neither a lesion-localization method nor a post-hoc attribution method. It measures and optimizes relative behavior under constructed interventions.

Shortcut-aware evaluation. Calibration and selective prediction address confidence or abstention Guo et al. (2017); Wei et al. (2022); Geifman & El-Yaniv (2017). For our intervention study, recorded edit geometry and image differences can themselves predict target versus control. We therefore use grouped cross-fitting and complementary feature scopes to bound what the observed response supports. Measured adjustment does not rule out unrecorded confounding.

3 Matched Counterfactual Setting

3.1 Problem Setup

Consider a factual input x, a target intervention x_t, and a matched control x_c. For a question-conditioned state $h(x, Q)$, define $\Delta_t = h(x_t, Q) - h(x, Q)$ and $\Delta_c = h(x_c, Q) - h(x, Q)$. For an output score $s(x, Q)$, the relative behavioral response is $S_{\text{ans}} = s(x_c, Q) - s(x_t, Q)$. These quantities identify target-versus-control behavior only under the specified intervention and matching protocol.

In our chest-radiograph instantiation, samples are localized-positive tuples $(I_i, Q_i, M_i)_{i=1}^N$, where M_i is an expert annotation used to construct the intervention. All factual answers are positive: $y_i^f = \texttt{Yes}$. Masks are used only to construct and audit interventions and are never provided to the VLM or CARAE.

For an answer sequence a, we define the length-normalized log-likelihood

$$\ell_\Theta(a; I, Q) = \frac{1}{|a|} \sum_{k=1}^{|a|} \log p_\Theta(a_k \mid a_{<k}, I, Q), \tag{1}$$

and the positive-answer margin

$$s_\Theta(I, Q) = \ell_\Theta(\texttt{Yes}; I, Q) - \ell_\Theta(\texttt{No}; I, Q). \tag{2}$$

3.2 Paired Evidence Interventions

For each sample, we construct a factual–targeted–control triplet

$$\mathcal{T}_i = (I_i^f, I_i^t, I_i^c), \qquad I_i^f = I_i, I_i^t = \mathcal{R}(I_i, M_i), I_i^c = \mathcal{R}(I_i, C_i), \tag{3}$$

where $\mathcal{R}$ is the shared intervention operator and C_i is a non-overlapping control mask approximately matched to M_i in area and shape.

The corresponding behavioral targets are

$$y_i^f = \texttt{Yes}, \qquad y_i^t = \texttt{No}, \qquad y_i^c = \texttt{Yes}. \tag{4}$$

These are protocol-defined targets, not independently adjudicated clinical labels for the synthesized images.

3.3 Evidence-Conditional Reliability

Targeted removal alone may capture generic corruption or intervention artifacts. We therefore compare it with the matched non-target control.

Let $s_i^v = s_\Theta(I_i^v, Q_i)$ for $v \in \{f, t, c\}$. Answer-level evidence specificity is

$$S_i^{\text{ans}} = s_i^c - s_i^t, \tag{5}$$

which is positive when targeted removal suppresses the positive answer more strongly than the control.

For an internal visual state $\mathbf{z}_i^v = \Phi(I_i^v, Q_i)$, representation-level specificity is

$$S_i^{\text{repr}} = d(\mathbf{z}_i^f, \mathbf{z}_i^t) - d(\mathbf{z}_i^f, \mathbf{z}_i^c). \tag{6}$$

A positive value indicates a larger internal response to targeted removal.

We assess relative target-versus-control response at both answer and representation levels while tracking factual positive-anchor preservation. We call an insufficient targeted-over-control response *counterfactual rigidity* under this protocol. CARAE instantiates Φ as a projected evidence state and optimizes both criteria jointly.

4 CARAE

4.1 Framework Overview

We propose **CARAE**, short for a method for matched counterfactual transition learning. CARAE learns question-conditioned visual representations by contrasting the effect of removing answer-supporting regions against matched non-target interventions. The framework contains two complementary components: an evidence-specific transition objective that shapes the latent geometry during training, and a question-conditioned token-modulated residual coupling module that injects the learned state into the visual stream for answer generation.

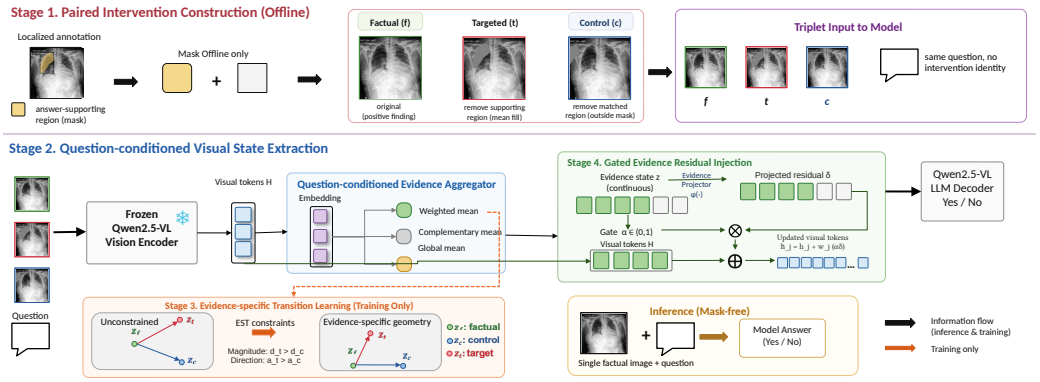


Figure 1: Overview of CARAE. Expert masks construct the image interventions offline and are never model inputs. A question-conditioned state is learned from visual tokens; EST compares factual, targeted, and control states during training only. The gated residual couples this state to generation. Learned token weights are not lesion-localization predictions.

Localization annotations are used only offline to construct the targeted and matched-control images. They are never provided to the VLM, the evidence aggregator, or the answer generation pathway. For each factual–targeted–control triplet,

$$\mathcal{T}_i = (I_i^f, I_i^t, I_i^c), \tag{7}$$

the three conditions are processed independently using shared parameters. For $v \in \{f, t, c\}$, the forward process is

$$(I_i^v, Q_i) \xrightarrow{V_\phi} \mathbf{H}_i^v \xrightarrow{E_\eta} \mathbf{u}_i^v \xrightarrow{P_\theta} \mathbf{z}_i^v \xrightarrow{J_\omega} \widetilde{\mathbf{H}}_i^v \xrightarrow{F_\Theta} p(a|I_i^v, Q_i), \tag{8}$$

where V_ϕ denotes the frozen vision encoder that produces visual tokens $\mathbf{H}$. E_η is the question-conditioned evidence aggregator that extracts a question-conditioned representation $\mathbf{u}$ from visual tokens and the question. P_θ projects this representation into the visual latent space, producing the evidence state $\mathbf{z}$. J_ω performs gated residual injection by updating the visual representation with an evidence-conditioned residual. Finally, F_Θ denotes the native Qwen2.5-VL generation pathway.

The triplet construction and EST objective are used only during training to learn evidence-specific transitions. At inference time, CARAE discards all intervention information and processes only a single factual image–question pair: no localization mask, targeted image, or control image is required.

4.2 Question-Conditioned Visual State

For each image and question, CARAE computes a learned weight $w_{i,j}^v$ for each merged visual token. The same weights define weighted and complementary visual means, which are concatenated with the global mean and projected to the compact state $\mathbf{z}_i^v$. The question representation uses input embeddings without answers, masks, or intervention identity. These weights are reused to modulate the state-conditioned residual at injection; they are learned without lesion-mask supervision and are not interpreted as lesion locations. Full equations and extraction details appear in Appendix A.

4.3 Evidence-Specific Transition Learning

The factual-to-targeted and factual-to-control transitions are

$$\boldsymbol{\Delta}_i^t = \mathbf{z}_i^t - \mathbf{z}_i^f, \qquad \boldsymbol{\Delta}_i^c = \mathbf{z}_i^c - \mathbf{z}_i^f. \tag{9}$$

Let

$$\bar{\mathbf{z}}_i^v = \frac{\mathbf{z}_i^v}{\|\mathbf{z}_i^v\|_2 + \epsilon}, \qquad d_i^t = \|\bar{\mathbf{z}}_i^t - \bar{\mathbf{z}}_i^f\|_2, \qquad d_i^c = \|\bar{\mathbf{z}}_i^c - \bar{\mathbf{z}}_i^f\|_2. \tag{10}$$

The transition-magnitude loss encourages the targeted edit to produce a larger normalized state change than its control:

$$\mathcal{L}_{\text{dist}} = \frac{1}{B}\sum_{i=1}^{B} \tau_d \,\text{softplus}\left(\frac{m_d + d_i^c - d_i^t}{\tau_d}\right). \tag{11}$$

Magnitude alone does not specify how the two transitions differ. To rank their directions in the learned state, define

$$\widehat{\boldsymbol{\Delta}}_i^{\,v} = \frac{\boldsymbol{\Delta}_i^v}{\|\boldsymbol{\Delta}_i^v\|_2 + \epsilon}, \qquad a_i^v = D_\psi(\widehat{\boldsymbol{\Delta}}_i^{\,v}), \quad v \in \{t, c\}, \tag{12}$$

where

$$D_\psi(\mathbf{x}) = \mathbf{w}_\psi^\top \,\text{LN}(\mathbf{x}). \tag{13}$$

The implementation includes a shared scalar bias, which cancels in the pairwise score difference below and receives no gradient from this loss. The directional ranking loss is

$$\mathcal{L}_{\text{dir}} = \frac{1}{B}\sum_{i=1}^{B} \tau_q \,\text{softplus}\left(\frac{m_q + a_i^c - a_i^t}{\tau_q}\right). \tag{14}$$

The evidence-specific transition objective is

$$\mathcal{L}_{\text{EST}} = \lambda_d \mathcal{L}_{\text{dist}} + \lambda_q \mathcal{L}_{\text{dir}}. \tag{15}$$

The objective organizes relative transitions under the training protocol; it does not identify a universal disease-removal direction.

4.4 Evidence-State Coupling to Generation

A discriminative projected state does not by itself ensure that the generation pathway uses it. CARAE predicts a sample-specific residual direction and scalar gate from each condition's state:

$$\begin{aligned} \boldsymbol{\delta}_i^v &= W_\delta \,\text{LN}_\delta(\mathbf{z}_i^v) + \mathbf{b}_\delta, \\ \alpha_i^v &= \sigma\left(\mathbf{w}_\alpha^\top \,\text{LN}_\alpha(\mathbf{z}_i^v) + b_\alpha\right). \end{aligned} \tag{16}$$

Here LN_δ and LN_α denote the separate LayerNorm modules in the residual and gate branches. The token weights $w_{i,j}^v$ are the same question-conditioned weights used to construct $\mathbf{z}_i^v$. The actual update at merged visual token j is

$$\widetilde{\mathbf{h}}_{i,j}^v = \mathbf{h}_{i,j}^v + w_{i,j}^v \left(\alpha_i^v \boldsymbol{\delta}_i^v\right), \qquad j = 1, \ldots, N_i. \tag{17}$$

Thus $\boldsymbol{\delta}_i^v$ and α_i^v are sample-specific, while $w_{i,j}^v$ makes the applied residual token-dependent. For a batch, $\mathbf{z}_i^v \in \mathbb{R}^{D_e}$, $\boldsymbol{\delta}_i^v \in \mathbb{R}^{D_v}$, and $\alpha_i^v \in (0,1)$ is scalar; $w_{i,j}^v \in (0,1)$ is scalar for each merged visual token. The sequence update can equivalently be written $\widetilde{\mathbf{H}}_i^v = [\widetilde{\mathbf{h}}_{i,j}^v]_{j=1}^{N_i}$. In the reported mask-free CARAE implementation the gate is an uncapped sigmoid, and no normalization, clipping, or other transformation is applied to the residual after these branches and before token-wise addition. The weights are not supervised by expert masks and are not interpreted as lesion-localization predictions. The modified visual tokens are assembled into the model's visual input and processed by the original generation pathway:

$$p_\Theta(a \mid I_i^v, Q_i) = F_\Theta(\widetilde{\mathbf{H}}_i^v, Q_i). \tag{18}$$

4.5 Optimization

The training objective combines triplet answer likelihood, a target-versus-control answer-margin loss, EST, and weight coverage and entropy regularizers. Its weights and exact equations are in Appendix A.2. The same canonical objective is used across both model scales and all reported seeds.

5 EXPERIMENTAL SETUP

5.1 DATASETS AND INTERVENTIONS

Canonical evaluation. We use localized-positive CheXlocalize image–finding pairs and the corresponding CheXpert radiographs Irvin et al. (2019); Saporta et al. (2022). The dataset contains 536 image–finding pairs from 157 images and 145 patients across nine findings. Patient-disjoint train/validation/test splits contain $374/81/81$ pairs.

Each pair yields factual, targeted, and matched-control conditions, for $1{,}608$ inputs. All conditions share the same question, and the model receives no intervention identity or mask-derived information. Pairs are the evaluation units and patients the resampling units.

Targeted masks are filled with the per-channel mean of the unmasked image. Control masks are non-overlapping and approximately matched in area and shape, with connected growth as fallback. The targets are `Yes`, `No`, and `Yes` for factual, targeted, and control conditions. The targeted label is a behavioral target, not an independently adjudicated diagnosis. Full construction details are provided in the Detailed Experimental Protocol appendix. Canonical optimization uses the frozen strict-control realization described in the appendix. After the canonical checkpoints are frozen, complementary control constructions and replacement operators are evaluated without parameter updates or checkpoint selection. These experiments test transfer across realizations of the same target-control behavioral interface.

Frozen transfer. Canonical checkpoints are evaluated on MS-CXR and VinDr-CXR without retraining, checkpoint selection, or dataset-specific calibration. We use the same intervention and scoring protocol, with patient-grouped resampling for MS-CXR and image-grouped resampling for VinDr-CXR.

5.2 BACKBONES AND COMPARISONS

We evaluate Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct.

Comparisons include the frozen backbone, standard SFT, LogitNorm, response-only triplet training, and a CF-VLM-style baseline. We further compare global injection, Matched Coupling (the same question-conditioned coupling and matched response objective without EST), and Full CARAE. All methods use identical data, interventions, preprocessing, questions, and scoring.

For Full CARAE and its architectural variants, the vision encoder is frozen and LoRA is applied to `q_proj`, `k_proj`, `v_proj`, and `o_proj` with rank 8, scale 16, and dropout 0.05. The LoRA and Full CARAE module learning rates are 2×10^{-5} and 10^{-4}, respectively. Each configuration is trained for one epoch with seeds 0, 1, and 2 and evaluated at the terminal checkpoint. Further settings are provided in the Detailed Experimental Protocol appendix.

5.3 METRICS

Answer specificity. The primary metric is AUROC for distinguishing control from targeted conditions using the positive-answer margin in Eq. (2). We also report paired accuracy, $\mathbb{I}[s_i^c > s_i^t]$, and the mean gap, $s_i^c - s_i^t$.

Positive-anchor preservation. Because all canonical factual labels are positive, factual accuracy is not informative. We instead report factual NLL and ECE as positive-anchor preservation measures, not balanced diagnostic calibration.

Representation behavior. For methods with a projected evidence state, we report

$$d(\bar{\mathbf{z}}^f, \bar{\mathbf{z}}^t) - d(\bar{\mathbf{z}}^f, \bar{\mathbf{z}}^c), \tag{19}$$

paired visual accuracy, effective rank, and representation–answer Spearman correlation. Partial coupling is computed after residualizing recorded intervention metadata.

5.4 Statistics and Shortcut Audit

Results are averaged over three seeds. We use $2{,}000$ grouped bootstrap replicates, resampling patients for CheXlocalize and MS-CXR and source images for VinDr-CXR. Paired comparisons use aligned predictions and identical draws, with two-sided percentile 95% confidence intervals.

The shortcut audit uses eleven recorded mask and image features. A metadata-only classifier is evaluated with five-fold grouped cross-fitting. We then report residualized AUROC and the incremental AUROC beyond metadata.

The feature scopes separate recorded intervention geometry and local edit signatures from full radiographic context. Residualized and incremental metrics quantify response beyond recorded variables, whereas the full-region audit preserves potentially finding-bearing appearance as a separate sensitivity scope. Together, these analyses make the intervention response auditable across nuisance scopes.

6 Results and Analysis

6.1 RQ1: Does response training distinguish target from control?

Table 1 shows that the frozen backbone, Standard SFT, and LogitNorm preserve the positive factual answer but remain near chance in targeted-versus-control discrimination. Correct factual behavior therefore does not establish discrimination between target and matched-control interventions.

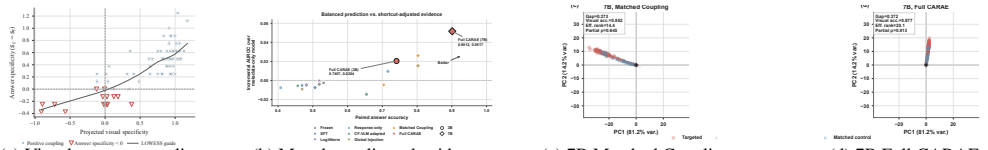
6.2 RQ2–3: What do coupling and EST add?

Table 1: Cross-scale results on Qwen2.5-VL-3B and Qwen2.5-VL-7B. Evidence-specific metrics compare targeted and matched-control interventions; positive-anchor metrics measure factual confidence retention; shortcut-adjusted metrics use grouped cross-fitting to account for recorded intervention metadata.

Model / Method	Evidence Specificity			Positive-Anchor		Shortcut Adjusted	
	Answer AUROC ↑	Spec. Gap ↑	Pair Acc. ↑	NLL ↓	ECE ↓	Residual AUROC ↑	Metadata Increment ↑
Qwen2.5-VL-3B							
Frozen backbone	0.5210	0.0044	0.5309	0.1354	0.1237	0.5150	-0.0026
Standard SFT	0.5248	0.0001	0.4691	0.0021	0.0021	0.5240	-0.0052
LogitNorm	0.5107	-0.0000	0.5185	$\mathbf{4.46\times10^{-5}}$	$\mathbf{4.46\times10^{-5}}$	0.5767	-0.0005
Response-only	0.5133	0.0098	0.4074	0.4153	0.3375	0.5063	-0.0076
CF-VLM-style FT	0.5247	0.0293	0.4568	0.7106	0.1540	0.4528	-0.0056
Generic global injection	0.5786	0.0242	0.4691	0.3999	0.3268	0.5197	-0.0090
Matched Coupling	0.7623	0.1170	0.7037	0.3332	0.2808	0.6220	-0.0046
Full CARAE (Ours)	**0.8038**	**0.2552**	**0.7407**	0.2266	0.2004	**0.6504**	**0.0204**
Qwen2.5-VL-7B							
Frozen backbone	0.4910	-0.0054	0.5185	0.1002	0.0703	0.5799	-0.0040
Standard SFT	0.4836	-0.0002	0.5062	0.0013	0.0012	0.5522	-0.0075
LogitNorm	0.4662	-0.0000	0.4815	$\mathbf{2.22\times10^{-5}}$	$\mathbf{2.22\times10^{-5}}$	0.5054	-0.0047
Response-only	0.7053	0.0673	0.7160	0.2919	0.2512	0.6241	0.0096
CF-VLM-style FT	0.6225	0.5394	0.6543	0.3211	0.1460	0.5501	-0.0145
Generic global injection	0.7458	0.3036	0.8025	0.1353	0.1247	0.6258	0.0154
Matched Coupling	0.8275	0.5770	0.8025	0.1410	0.1286	0.6703	0.0261
Full CARAE (Ours)	**0.8527**	**0.8220**	**0.9012**	0.0877	0.0814	**0.7090**	**0.0517**

Full CARAE gives the strongest answer-level and shortcut-adjusted results at both scales. Response-only training and global injection yield smaller or less consistent gains, showing that triplet supervision or residual adaptation alone is insufficient.

Question-conditioned aggregation and coupling provide the main measured improvement over response-only training. EST provides an additional numerical gain over Matched Coupling and is associated with more separated projected transitions. On 7B, Full CARAE exceeds Matched



(a) Visual–answer coupling (b) Metadata-adjusted evidence ordering (c) 7B Matched Coupling (d) 7B Full CARAE

Figure 2: **Representation and response diagnostics.** (a) Each point is a canonical 3B Full CARAE test pair. (b) Each point is a method–backbone configuration. (c,d) Held-out 7B target and control transitions share a PCA basis; colors denote target and matched-control transitions. Projected geometry is illustrative; quantitative gaps and answer associations use the original state space. The complete 3B/7B plots appear in Appendix Figure 4.

Coupling by $+0.0252$ AUROC with a paired 95% interval of $[0.0021, 0.0556]$; on 3B the difference is positive but less precisely estimated. We therefore view EST as a mechanism-level refinement of the complete model rather than the sole source of its performance.

Table 2. Paired comparisons on canonical 7B. Patient-clustered; 2,000 aligned bootstrap replicates.

Comparison	Δ AUROC [95% CI]	Δ Pair
Full / Resp.	**+0.1474 [0.0737, 0.2266]**	+0.1852
Full / Global	**+0.1069 [0.0512, 0.1673]**	+0.0988
Full / w/o EST	**+0.0252 [0.0021, 0.0556]**	+0.0988
w/o EST / Resp.	**+0.1222 [0.0601, 0.1907]**	+0.0864

Table 3. Representation diagnostics. Partial coupling is residualized against metadata.

Scale / Method	Gap	Vis. Acc.	Rank	ρ_p
3B Global	0.067	0.778	36.80	0.351
3B Matched	0.330	0.790	19.01	0.729
3B Full	**0.538**	**0.815**	29.31	**0.866**
7B Global	0.172	0.864	20.08	0.656
7B Matched	0.273	0.852	14.40	0.645
7B Full	**0.372**	**0.877**	20.08	**0.813**

Representation diagnostics.

Question-conditioned variants produce positive targeted-over-control transition gaps (Table 3). Full CARAE has larger projected ordering gaps and stronger metadata-adjusted visual–answer association. Figure 2(a) illustrates this association.

These associations are consistent with the intended representation structure. They do not establish that the learned state functionally mediates the answer.

6.3 RQ4: Are recorded intervention shortcuts sufficient?

Recorded intervention metadata predicts target versus control to some degree. We therefore use five-fold grouped cross-fitting to report metadata-only prediction, residualized AUROC, and incremental AUROC (Table 1). Full CARAE retains positive increments of 0.0204 (3B) and 0.0517 (7B). An image-difference audit finds weak direct edit signatures under its validation gate (AUROC 0.5688–0.6385), while full-region context remains informative (HGB/RF AUROC 0.8673/0.8599). Full-context residualization is a conservative sensitivity analysis. These checks address measured intervention-correlated variation; they do not remove unmeasured confounding. Full protocols appear in the appendix.

6.4 RQ5: Does the response transfer across intervention realizations?

This is the main cross-construction generalization test: the canonical 7B checkpoint is frozen, then evaluated with unseen random, anatomy-matched, and appearance-matched controls. Full CARAE remains ahead of response-only training and Matched Coupling across the three evaluated families. The detailed paired intervals and overlap analysis are described below.

Table 4. Held-out control AUROC.

Method	Random	Anatomy	Appearance
Metadata-only	0.653	0.675	0.662
Response-only	0.654	0.661	0.663
Matched Coupling	0.792	0.774	0.795
Full CARAE	**0.849**	**0.824**	**0.837**

Held-out controls. The canonical model defines the target-control behavioral interface using the frozen strict-control realization. Its learned response geometry is then evaluated, without retraining,

with three unseen control families—random, anatomy-matched, and appearance-matched—that are not used in the canonical triplets (Table 4). Full CARAE consistently outperforms response-only training and Matched Coupling across all three constructions. All paired patient-bootstrap confidence intervals exclude zero, and the advantage persists on the subset with the greatest cross-fitted metadata overlap. If the model relied only on the canonical control construction, this advantage would be expected to collapse. Its persistence directly tests transfer across the evaluated control realizations; it does not establish invariance to arbitrary edits.

Held-out replacement operators. We additionally evaluate the frozen canonical checkpoints under three alternative localized replacement operators: constant-gray filling, local-ring-mean filling, and locally matched texture noise. Across both model scales, Full CARAE outperforms response-only training under every evaluated replacement, with AUROC gains ranging from 0.0688 to 0.2449 for 3B and from 0.1486 to 0.2007 for 7B. If the model relied only on the canonical global-mean fill, its advantage would be expected to collapse under these replacements. The frozen advantage persists across the three tested operators; full results are in the alternative intervention-operator appendix.

6.5 RQ6: Does the response transfer across datasets?

On MS-CXR, Full CARAE improves answer AUROC over response-only training at both scales (0.5756 vs. 0.5090 for 3B; 0.6647 vs. 0.5879 for 7B). Pair accuracy and metadata-adjusted AUROC provide complementary checks: the 7B Full model reaches 0.7109 and 0.6829, respectively. Matched Coupling is close on 3B AUROC (0.5729), so the added value of Full CARAE is modest there. On VinDr-CXR, Full CARAE reaches AUROC 0.5444 and pair accuracy 0.4871; absolute transfer remains weak. Complete metrics, including anchor NLL and representation measures, appear in Appendix Table 6. Representative cases are shown in Appendix Figure 3.

Table 5. Frozen external-transfer metrics.
No retraining or dataset-specific calibration. Ans. AUC is target–control answer AUROC; Resid. AUC adjusts for recorded metadata.

Dataset	Scale	Method	AUC	Pair	Resid. AUC
MS-CXR	3B	Response	0.5090	0.4093	0.5139
		Global inj.	0.5122	0.4197	0.5104
		Matched	0.5729	0.5337	0.5610
		Full CARAE	**0.5756**	**0.5275**	**0.5742**
	7B	Response	0.5879	0.5927	0.6300
		Global inj.	0.6260	0.6487	0.6425
		Matched	0.6515	0.6642	0.6758
		Full CARAE	**0.6647**	**0.7109**	**0.6829**
VinDr-CXR	3B	Frozen backbone	0.5016	0.4411	0.5122
		Response	0.5085	0.3592	0.5136
		Global inj.	0.5191	0.3807	0.5238
		Matched	0.5110	0.3865	0.5140
		Full CARAE	**0.5444**	**0.4871**	**0.5509**

7 Discussion and Limitations

CARAE estimates and optimizes differential responsiveness between targeted and matched-control interventions under a specified intervention family. This quantity is stronger than sensitivity to a targeted edit alone, but it does not establish that synthetic removal eliminates all clinically relevant evidence or yields a clinically adjudicated counterfactual diagnosis. Recorded metadata adjustment cannot eliminate unmeasured confounding. Chest radiography is the current empirical setting; other modalities and tasks remain unevaluated. The state-conditioned residual direction and gate are sample-specific, and the learned question-conditioned token weights modulate the residual at each merged visual token. This computational token modulation is not evidence of lesion localization. A direct state intervention would further test the functional role of the learned state; the current representation and coupling analyses are associative.

8 Conclusion

Matched controls help distinguish intended counterfactual response from sensitivity to an edit. CARAE uses a question-conditioned state and token-modulated gated residual to improve target-versus-control response in chest radiographs at two model scales and across several intervention realizations. These findings are limited to the tested constructions and datasets.

AI USE STATEMENT

Generative AI tools were used for minor language editing to improve clarity and for presentation checks, including Figure 1 layout refinement and reference consistency checks. They were not used to generate research data or determine experimental results. The authors take full responsibility for the methods, analyses, citations, and final manuscript.

ETHICS STATEMENT

This study uses existing chest-radiograph research datasets and synthetic image interventions for offline evaluation. It involves no new patient recruitment or clinical deployment. A targeted edit is a protocol-defined perturbation, not a clinically adjudicated disease-negative image; the results should not be interpreted as diagnostic validation.

REPRODUCIBILITY STATEMENT

The main paper specifies the intervention setting, model comparisons, scoring rules, and grouped statistical analysis in Sections 3–6. The appendix documents the state and optimization equations (Appendix A), data composition and intervention construction (Appendix B), and metric and bootstrap definitions (Appendix B.9 and Appendix B.10). We report the training seeds and frozen-checkpoint evaluation protocol, including the resampling unit for each dataset.

Anonymous code release. An anonymized release of the mask-free CARAE implementation, formal metric code, and canonical experiment runner is available at `https://anonymous.4open.science/r/carae-6046/`. Exact reruns additionally require the frozen intervention manifests and edited-image assets, which are not included in the release; the repository README describes these requirements and the supported reproduction scope.

REFERENCES

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10800–10809, 2020.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. CheXagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium Series on Clinical Foundation Models*, 2024. URL `https://openreview.net/forum?id=P3LOmrZWGR`.

Qiao Deng, Zhongzhen Huang, Yunqi Wang, Zhichuan Wang, Zhao Wang, Xiaofan Zhang, Qi Dou, Yeung Yu Hui, and Edward S. Hui. Grounded knowledge-enhanced medical vision-language pre-training for chest x-ray. *arXiv preprint arXiv:2404.14750*, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, International Convention Centre, Sydney, Australia, 2017. PMLR. URL `https://proceedings.mlr.press/v70/guo17a.html`.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=GSuP99u2kR.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021. doi: 10.1109/ISBI48211.2021.9434010.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: A multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 2023. URL https://proceedings.mlr.press/v225/moor23a.html.

Philip Müller, Georgios Kaissis, and Daniel Rueckert. ChEX: Interactive localization and region description in chest x-rays. *arXiv preprint arXiv:2404.15770*, 2024.

Xi Ouyang, Srikrishna Karanam, Ziyan Wu, Terrence Chen, Jiayu Huo, Xiang Sean Zhou, Qian Wang, and Jie-Zhi Cheng. Learning hierarchical attention for weakly-supervised chest x-ray abnormality localization and diagnosis. *IEEE Transactions on Medical Imaging*, 40(10):2698–2710, 2021. doi: 10.1109/TMI.2020.3042773.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 2020.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven Q. H. Truong, Chanh D. T. Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022. doi: 10.1038/s42256-022-00536-x.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23631–23644. PMLR, 17–23 Jul 2022.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, Weidi Xie, et al. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16:7866, 2025. doi: 10.1038/s41467-025-62385-7.

Guangzhi Xiong, Qiao Jin, Sanchit Sinha, Zhiyong Lu, and Aidong Zhang. Rethinking visual attribution for chest x-ray reasoning in large vision language models. *arXiv preprint arXiv:2605.20158*, 2026.

Anas Zafar, Leema Krishna Murali, Siddhant Bharadwaj, Ashish Vashist, and Jia Wu. Do medical vision language models actually see? a counterfactual grounding framework and hard-negative contrastive training for visually-reliant medical vlms. *arXiv preprint arXiv:2607.03647*, 2026.

Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. CF-VLM: Counterfactual vision-language fine-tuning. *arXiv preprint arXiv:2506.17267*, 2025.

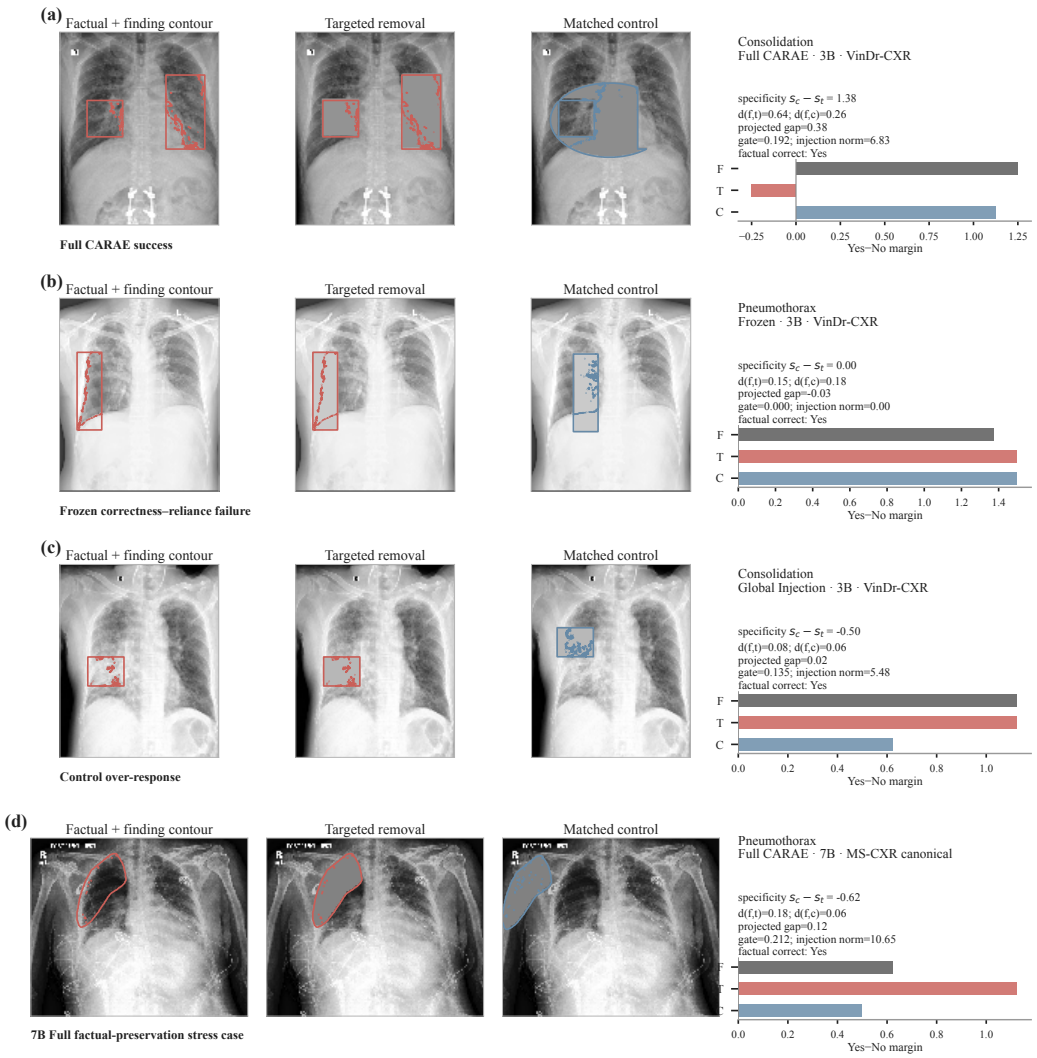


Figure 3: **Representative successes and failure modes under paired visual interventions.** Each row shows the factual radiograph with the annotated finding contour, the targeted-removal image, the matched-control image, and the corresponding positive-answer margins. (a) Full CARAE suppresses the answer after targeted removal while preserving it under the matched control. (b) The frozen model remains correct on the factual image but responds similarly to both interventions, illustrating counterfactual rigidity. (c) Global injection responds more strongly to the matched control than to targeted removal, producing negative answer specificity. (d) Under MS-CXR transfer, Full CARAE exhibits a positive projected transition gap but negative answer specificity, illustrating representation–answer mismatch. Red and blue contours denote targeted and matched-control regions, respectively.

Table 6: Frozen cross-dataset transfer without additional training, checkpoint selection, or dataset-specific calibration. MS-CXR uses patient-grouped resampling and VinDr-CXR uses image-grouped resampling.

Dataset	Scale	Method	Answer AUROC ↑	Pair Acc. ↑	Residual AUROC ↑	Projected Gap ↑	Partial Coupling ↑	Anchor NLL ↓
MS-CXR	3B	Response-only	0.5090	0.4093	0.5139	-0.0396	0.2224	0.4441
		Generic global injection	0.5122	0.4197	0.5104	-0.0145	0.3412	0.4412
		Matched Coupling	0.5729	**0.5337**	0.5610	**0.1835**	0.6017	0.3832
		Full CARAE (Ours)	**0.5756**	0.5275	**0.5742**	0.0604	**0.6463**	**0.2647**
	7B	Response-only	0.5879	0.5927	0.6300	-0.0184	0.1634	0.3456
		Generic global injection	0.6260	0.6487	0.6425	0.0301	0.4238	0.3368
		Matched Coupling	0.6515	0.6642	0.6758	0.1014	0.5538	**0.3187**
		Full CARAE (Ours)	**0.6647**	**0.7109**	**0.6829**	**0.1249**	**0.6414**	0.3400
VinDr-CXR	3B	Frozen backbone	0.5016	0.4411	0.5122	0.0001	0.0100	**0.1912**
		Response-only	0.5085	0.3592	0.5136	0.0035	0.1076	0.5182
		Generic global injection	0.5191	0.3807	0.5238	0.0094	0.1591	0.5161
		Matched Coupling	0.5110	0.3865	0.5140	0.0049	0.4766	0.4851
		Full CARAE (Ours)	**0.5444**	**0.4871**	**0.5509**	**0.0308**	**0.4944**	0.3384

A State and Optimization Details

A.1 Question-Conditioned Visual Aggregation

For condition v, the merged visual tokens are

$$\mathbf{H}_i^v = V_\phi(I_i^v) = [\mathbf{h}_{i,1}^v, \ldots, \mathbf{h}_{i,N_i}^v] \in \mathbb{R}^{N_i \times D_v}. \tag{20}$$

To obtain the question representation, we tokenize the raw question without special tokens and locate its last matching token subsequence in the multimodal prompt before the assistant-response boundary. Let $\mathcal{S}_i$ denote the resulting question-token positions and $\mathbf{E}_{\text{in}}$ the Qwen input embedding layer. We compute

$$\mathbf{q}_i^Q = \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \mathbf{E}_{\text{in}}[x_{i,j}]. \tag{21}$$

This representation uses non-contextualized input embeddings only. It does not contain answer tokens, contextual LLM hidden states, intervention identifiers, image-conditioned states, or mask-derived information. Because the same question is used across the triplet, $\mathbf{q}_i^Q$ is shared by the three conditions.

CARAE assigns each visual token a question-conditioned evidence weight:

$$\begin{aligned} e_{i,j}^v &= \mathbf{w}_e^\top \,\text{GELU}\left(W_h \mathbf{h}_{i,j}^v + W_Q \mathbf{q}_i^Q + \mathbf{b}_e\right), \\ w_{i,j}^v &= \sigma(e_{i,j}^v). \end{aligned} \tag{22}$$

Using these learned weights, we summarize weighted visual, complementary visual, and global visual features:

$$\begin{aligned} \mathbf{r}_i^v &= \frac{\sum_j w_{i,j}^v \mathbf{h}_{i,j}^v}{\max\left(\sum_j w_{i,j}^v, \epsilon\right)}, \\ \mathbf{c}_i^v &= \frac{\sum_j (1 - w_{i,j}^v) \mathbf{h}_{i,j}^v}{\max\left(\sum_j (1 - w_{i,j}^v), \epsilon\right)}, \\ \mathbf{g}_i^v &= \frac{1}{N_i} \sum_j \mathbf{h}_{i,j}^v, \qquad \mathbf{u}_i^v = [\mathbf{r}_i^v; \mathbf{c}_i^v; \mathbf{g}_i^v]. \end{aligned} \tag{23}$$

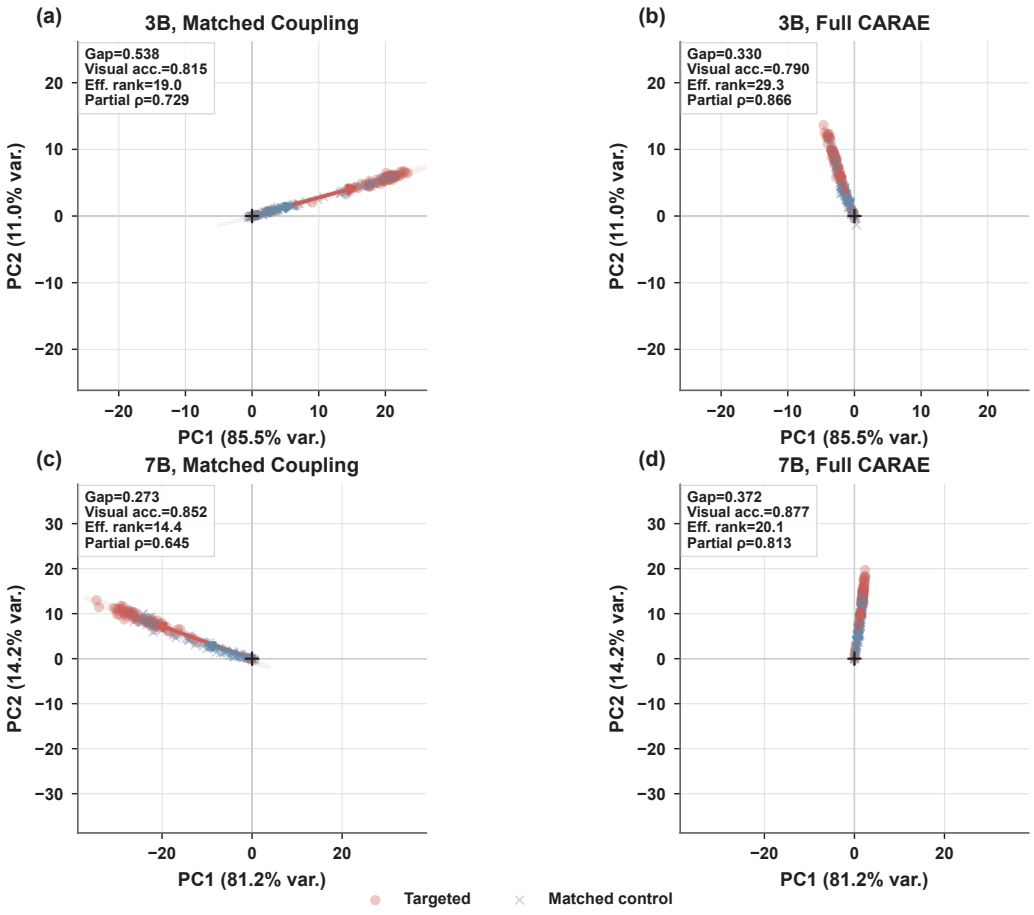


Figure 4: **Geometry of factual-to-targeted and factual-to-control transitions in the projected evidence space.** Transition vectors are translated to a common factual origin and visualized using a shared PCA basis within each backbone. Panels compare Matched Coupling and Full CARAE for the 3B and 7B backbones. Both variants exhibit distinct targeted and matched-control transition patterns, while Full CARAE changes their organization in the projected space. PCA is used only for visualization; projected-gap and coupling metrics are computed in the original representation space and reported in Table 3.

The learned weighting produces token-level weights that are reused in the generation coupling below. It is neither supervised by the expert mask nor interpreted as a lesion prediction.

A shared projector produces the compact visual evidence state

$$\mathbf{z}_i^v = P_\theta(\mathbf{u}_i^v) = W_2\,\mathrm{Dropout}\left[\mathrm{GELU}\left(W_1\,\mathrm{LN}(\mathbf{u}_i^v) + \mathbf{b}_1\right)\right] + \mathbf{b}_2 \in \mathbb{R}^{D_e}. \tag{24}$$

A.2 Optimization objective

A.3 Optimization

For condition v, let y_i^v denote the intervention-conditioned answer target. The triplet answer loss is

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{3B}\sum_{i=1}^{B}\sum_{v\in\{f,t,c\}} \ell_\Theta\left(y_i^v; \widetilde{\mathbf{H}}_i^v, Q_i\right). \tag{25}$$

Let

$$s_i^v = s_{\Theta,\eta,\theta,\omega}(I_i^v, Q_i) \tag{26}$$

denote the positive-answer margin produced by the complete CARAE forward path. The response-specificity objective is

$$\mathcal{L}_{\text{resp}} = \frac{1}{B}\sum_{i=1}^{B}[m_r + s_i^t - s_i^c]_+. \tag{27}$$

To prevent degenerate evidence weights, let

$$\bar{w}_i^v = \frac{1}{N_i}\sum_{j=1}^{N_i} w_{i,j}^v. \tag{28}$$

The coverage loss constrains the mean weight to the interval $[\rho_{\min}, \rho_{\max}]$:

$$\mathcal{L}_{\text{cov}} = \frac{1}{3B}\sum_{i=1}^{B}\sum_{v\in\{f,t,c\}}\left([\rho_{\min} - \bar{w}_i^v]_+^2 + [\bar{w}_i^v - \rho_{\max}]_+^2\right). \tag{29}$$

The entropy regularizer is the negative Bernoulli entropy:

$$\begin{aligned}\mathcal{L}_{\text{ent}} = \frac{1}{3B}\sum_{i=1}^{B}\sum_{v\in\{f,t,c\}}\frac{1}{N_i}\sum_{j=1}^{N_i}\Big[&w_{i,j}^v \log(w_{i,j}^v + \epsilon) \\ &+ (1 - w_{i,j}^v)\log(1 - w_{i,j}^v + \epsilon)\Big].\end{aligned} \tag{30}$$

Because this quantity is non-positive, minimizing it encourages higher-entropy evidence weights rather than sparse binary assignments.

The complete objective is

$$\mathcal{L}_{\text{CARAE}} = \mathcal{L}_{\text{CE}} + 0.2\mathcal{L}_{\text{resp}} + \mathcal{L}_{\text{EST}} + 0.01\mathcal{L}_{\text{cov}} + 0.001\mathcal{L}_{\text{ent}}, \tag{31}$$

where

$$\lambda_d = \lambda_q = 1, \qquad \rho_{\min} = 0.05, \qquad \rho_{\max} = 0.70. \tag{32}$$

The same canonical objective is used for the 3B and 7B backbones and across all reported random seeds. The remaining optimization and parameter-efficient adaptation settings are provided in the Appendix.

B Detailed Experimental Protocol

B.1 Canonical Dataset Composition and Patient-Level Splits

We construct the canonical training and evaluation dataset from the CheXlocalize validation annotations and the corresponding CheXpert chest radiographs Irvin et al. (2019); Saporta et al. (2022). A sample is defined as one image–finding pair with a non-empty positive localization mask. The resulting dataset contains 536 samples from 157 images and 145 patients.

The split is constructed at the patient level with a target ratio of $70/15/15$. Images and patients are mutually exclusive across splits, and all conditions derived from the same image–finding pair remain in the same split. Table 7 reports the final counts.

Each sample produces a factual, targeted, and matched-control condition:

$$I_i^f, \qquad I_i^t, \qquad I_i^c.$$

The complete canonical dataset therefore contains 1,608 model inputs. These inputs are not statistically independent. The image–finding pair is the paired sample unit, while the patient is the split and resampling unit.

The restricted dataset size is determined by annotation and image availability rather than artificial subsampling. Among 187 annotated images in the available CheXlocalize validation set, the nine findings define 1,683 potential image–finding positions. Of these, 1,147 have empty masks and 536 have non-empty positive masks. All retained positive masks satisfy the area and strict-control feasibility criteria. Additional CheXlocalize test annotations are not used because the corresponding CheXpert test images were unavailable in the experimental environment.

Table 7: Patient-grouped dataset split. A sample denotes one positive image–finding pair before intervention expansion.

Split	Samples	Images	Patients
Train	374	109	102
Validation	81	25	23
Test	81	23	20
Total	536	157	145

Table 8: Finding distribution across patient-grouped splits. Some rare findings are absent from the validation or test split; consequently, the experiment should not be interpreted as independent held-out validation for every finding category.

Finding	Train	Validation	Test	Total
Cardiomed.	75	17	17	109
Cardiomegaly	44	11	13	68
Lung Lesion	1	0	0	1
Airspace Opacity	88	19	18	125
Edema	32	7	6	45
Consolidation	24	7	2	33
Atelectasis	58	11	11	80
Pneumothorax	6	0	2	8
Pleural Effusion	46	9	12	67
Total	374	81	81	536

B.2 Frozen External-Transfer Evaluation

We evaluate the trained canonical checkpoints on MS-CXR and VinDr-CXR without additional training, parameter updates, checkpoint selection, or dataset-specific calibration. The external evaluations use the same binary question format, intervention operator, answer-scoring rule, and targeted-versus-control criterion as the canonical experiment.

For each external localized finding, the original radiograph forms the factual condition. The available localization annotation is converted to the targeted intervention region, and a matched non-target control is generated outside the annotated region. The same per-channel mean-filling operator is applied to both targeted and control masks. No external sample is used to update the LoRA parameters, evidence projector, direction scorer, or coupling module.

MS-CXR provides patient identifiers, and its uncertainty estimates therefore use patient-grouped resampling. For the retained VinDr-CXR evaluation set, patient identifiers are unavailable; samples are consequently grouped by source image during cross-fitting and bootstrap resampling. All image–finding pairs derived from the same image are kept together.

These evaluations measure frozen transfer of target-versus-control responsiveness across radiograph sources and localization formats. They do not establish cross-architecture, cross-modality, or clinical deployment generalization.

B.3 Question and Intervention-Conditioned Answer Protocol

For each image–finding pair, we use a controlled binary question asking whether the associated finding is present. The factual, targeted, and control conditions share the same question template and finding name. No mask location, intervention identity, control-generation type, or intervention metadata is included in the prompt.

The protocol targets are

$$y_i^f = \texttt{Yes}, \qquad y_i^t = \texttt{No}, \qquad y_i^c = \texttt{Yes}. \tag{33}$$

Table 9: Final strict-control construction types.

Control type	Samples
Translation	204
Connected-region growth	193
Horizontal reflection	94
Vertical reflection	45
Total	536

The targeted target is intervention-conditioned: removing the annotated answer-supporting region is expected to reduce support for the factual positive answer. It is not an independently adjudicated clinical diagnosis for the synthesized image. The experiments therefore evaluate controlled behavioral responsiveness to localized evidence removal rather than diagnostic validity of the modified radiographs.

Binary answer scores are computed from length-normalized sequence log-likelihoods:

$$s_i^v = \ell(\texttt{Yes}; I_i^v, Q_i) - \ell(\texttt{No}; I_i^v, Q_i), \qquad v \in \{f, t, c\}. \tag{34}$$

The same tokenization, prompt, and scoring procedure is used for all methods within a backbone.

B.4 Mask Preprocessing

The original localization annotation is mapped to the model input resolution before intervention construction. Binary masks are resized using nearest-neighbor interpolation. No morphological dilation, erosion, smoothing, or boundary expansion is applied in the canonical protocol.

A canonical mask is retained only when

$$\frac{|M_i|}{H_i W_i} > 5 \times 10^{-4}. \tag{35}$$

The loader additionally requires the mask fraction to be below 0.75. The strict equal-area control construction imposes an additional feasibility requirement because a non-overlapping control region of the same area must remain available.

External localization annotations are converted to binary masks at the corresponding input resolution before applying the same intervention pipeline. The original external annotation is used only to define the answer-supporting region; it is not provided to the model.

B.5 Image Intervention Operator

For an image I and intervention mask A, the fill vector is computed from all pixels outside the current mask:

$$\boldsymbol{\mu}(I, A) = \frac{\sum_{p:\, A_p = 0} I_p}{|\{p : A_p = 0\}|} \in \mathbb{R}^3. \tag{36}$$

The intervention operator is

$$\mathcal{R}(I, A)_p = \begin{cases} \boldsymbol{\mu}(I, A), & A_p = 1, \\ I_p, & A_p = 0. \end{cases} \tag{37}$$

The mean is calculated independently for each RGB channel over the full unmasked image rather than a local neighborhood. Because targeted and control masks differ, their estimated fill vectors may differ slightly. No fixed grayscale value, learned inpainting model, or condition-specific texture generator is used.

B.6 Strict Matched-Control Construction

The formal canonical experiments use a frozen strict-control realization and do not use the earlier pilot construction based on a limited number of random translations.

For each target mask M_i, the procedure first enumerates rigid control candidates at a stride of four pixels. The candidate transformations include:

1. spatial translation;
2. horizontal reflection;
3. vertical reflection.

Only candidates that remain inside the valid image region and have zero overlap with the target mask are retained. At most 64 rigid candidates are kept per sample.

When no valid rigid candidate is available, the procedure generates up to 32 equal-area candidates using connected-region growth outside the target region. A candidate is accepted only if it satisfies the frozen metadata-distance criterion. Samples without an acceptable candidate would be excluded. All 536 canonical samples have a valid strict-control realization.

The selected control types are reported in Table 9.

The construction exactly controls intervention area and enforces zero target–control overlap, while approximately controlling shape and lower-order mask statistics. It does not guarantee matching of anatomical location, surrounding texture, radiographic density, or higher-order image structure. The intervention identity consequently remains partly predictable from recorded metadata, motivating the explicit audit in the Intervention-Shortcut Audit appendix.

B.7 Backbones, Adaptation, and Training Configuration

We use Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct to evaluate scale consistency within one VLM family. The vision encoder is frozen. For Full CARAE and its architectural variants, LoRA is applied to

$$\{\texttt{q_proj}, \texttt{k_proj}, \texttt{v_proj}, \texttt{o_proj}\}$$

with rank 8, scaling factor 16, and dropout 0.05.

The learning rates are

$$\eta_{\mathrm{LoRA}} = 2 \times 10^{-5}, \qquad \eta_{\mathrm{agg}} = \eta_{\mathrm{proj}} = \eta_{\mathrm{dir}} = \eta_{\mathrm{inj}} = 10^{-4}, \tag{38}$$

where the latter rate is used for the evidence aggregator, projector, direction scorer, residual projection, and gate.

Each method–backbone configuration is trained with seeds 0, 1, and 2. Every run uses one factual–targeted–control triplet per optimization step, gradient accumulation of one, and one prespecified epoch, yielding 374 updates. The terminal checkpoint is evaluated without validation-based checkpoint selection. Validation summaries are recorded only for auditing.

Main point estimates are computed independently for each seed and then averaged across the three runs. Seed-level variability is reported where space permits. Grouped bootstrap intervals quantify uncertainty over held-out patient or image groups and should not be interpreted as confidence intervals over the population of training seeds.

For the CF-VLM-style baseline, LoRA is applied to

$$\{\texttt{q_proj}, \texttt{k_proj}, \texttt{v_proj}\}$$

with rank 16, scaling factor 32, and dropout 0.1, following its adapted configuration. Because this baseline is not parameter-matched to CARAE, comparisons should be interpreted as method-level rather than strict capacity-controlled comparisons. Trainable parameter counts are reported with the released configurations.

B.8 Comparison Methods

The *frozen backbone* is evaluated without adaptation. *Standard SFT* applies conventional answer supervision without paired answer-specificity or evidence-transition objectives. *LogitNorm* tests whether output-scale regularization alone improves intervention discrimination.

Response-only uses factual, targeted, and control answer supervision together with $\mathcal{L}_{\text{resp}}$, but does not construct or inject a CARAE evidence state. The *CF-VLM-style* baseline applies counterfactual fine-tuning to the same intervention triplets without CARAE's paired transition ranking or evidence-state coupling.

Generic global injection uses a global-mean visual representation and constant token weights of 0.5, applying the same gated residual to every merged visual token. It controls for residual modulation and additional trainable capacity without learned question-conditioned token weighting or weighted visual aggregation.

Matched Coupling (without EST) retains the same mask-free question-conditioned aggregation, projected evidence state, gated coupling, triplet answer loss, response-specificity loss, coverage regularization, and entropy regularization as Full CARAE, but sets

$$\lambda_{\text{EST}} = 0.$$

It therefore isolates the contribution of explicit distance- and direction-sensitive transition learning.

Full CARAE uses the complete objective

$$\mathcal{L}_{\text{CE}} + 0.2\mathcal{L}_{\text{resp}} + \mathcal{L}_{\text{EST}} + 0.01\mathcal{L}_{\text{cov}} + 0.001\mathcal{L}_{\text{ent}}.$$

All applicable methods use the same patient split, questions, intervention triplets, image preprocessing, answer scoring, training schedule, and random seeds. They do not necessarily have identical trainable parameter counts.

B.9 Answer- and Representation-Level Metrics

The primary answer-level metric is AUROC for distinguishing control from targeted conditions using the positive-answer margin s_i^v. Control rows are treated as evidence-preserving and targeted rows as evidence-removed. We additionally report

$$S_i^{\text{ans}} = s_i^c - s_i^t, \qquad \text{Acc}_{\text{pair}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[s_i^c > s_i^t]. \tag{39}$$

AUROC measures global ranking, paired accuracy measures within-pair ordering, and the mean specificity gap measures response magnitude.

Factual accuracy, negative log-likelihood, and expected calibration error are computed on the unmodified images. Because all retained factual labels are positive, these metrics measure positive-answer preservation and confidence rather than balanced diagnostic performance.

For methods with a defined projected evidence state, we compute

$$d_i^t = \|\bar{\mathbf{z}}_i^t - \bar{\mathbf{z}}_i^f\|_2, \qquad d_i^c = \|\bar{\mathbf{z}}_i^c - \bar{\mathbf{z}}_i^f\|_2, \tag{40}$$

and report the mean projected gap $d_i^t - d_i^c$, paired visual accuracy $\mathbb{I}[d_i^t > d_i^c]$, and effective rank. Effective rank is treated only as a diagnostic of dimensional concentration.

Visual–answer coupling is the Spearman association between $S_i^{\text{repr}} = d_i^t - d_i^c$ and S_i^{ans}. Partial coupling rank-transforms both variables and the metadata covariates, residualizes both specificity variables against the metadata, and then computes their association. It measures adjusted association rather than causal mediation.

B.10 Grouped Bootstrap and Multi-Seed Aggregation

All canonical configurations are trained with seeds 0, 1, and 2. For each seed, predictions are aligned by dataset, group, image–finding pair, and intervention condition.

Uncertainty is estimated using $2{,}000$ grouped bootstrap replicates with bootstrap seed 0. Patients are resampled for the canonical and MS-CXR evaluations. VinDr-CXR uses source-image groups because retained patient identifiers are unavailable. All findings and intervention conditions belonging to a sampled group are retained together.

For paired method comparisons, the same grouped draw is applied to both methods within each training seed. The metric difference is computed for each seed and then averaged across the three seeds for that replicate. Two-sided percentile 95% confidence intervals are obtained from the resulting distribution of mean seed-level differences.

The canonical test split contains only 20 patients. The intervals therefore reflect uncertainty under the observed patient clusters but remain limited by the small number of independent groups and the availability of only three training runs.

B.11 Intervention-Shortcut Audit

The audit uses eleven continuous features: mask area, area fraction, normalized centroid coordinates, width, height, aspect ratio, perimeter, compactness, mean intervened-image intensity, and edge density. Each targeted or control row uses the corresponding intervention mask and image.

For the canonical and MS-CXR evaluations, models are evaluated with five-fold patient-grouped cross-fitting. VinDr-CXR uses source-image grouping. The metadata-only classifier and metadata-plus-score classifier use

$$\texttt{StandardScaler} + \texttt{LogisticRegression}$$

with `max_iter`$= 2000$ and `random_state`$= 0$.

For residualization, an out-of-fold

$$\texttt{StandardScaler} + \texttt{Ridge}(\alpha = 1)$$

model predicts each method score $u_{i,v}$ from the metadata. The residual score is

$$r_{i,v} = u_{i,v} - \widehat{u}_{i,v}, \tag{41}$$

and residualized AUROC is computed from the concatenated out-of-fold residuals.

Let A_{meta} denote the out-of-fold AUROC of the metadata-only classifier and $A_{\text{meta+score}}$ the out-of-fold AUROC obtained after adding the method score. The metadata increment is

$$\Delta A_{\text{meta}} = A_{\text{meta+score}} - A_{\text{meta}}. \tag{42}$$

Both terms are computed using the same folds and aggregation protocol. Differences between this increment and arithmetic combinations of rounded table entries may arise from using unrounded out-of-fold predictions.

The eleven-feature audit is intentionally focused on recorded intervention geometry and local image-difference statistics. This keeps the adjustment auditable and avoids treating all regional texture as nuisance: full-region appearance may encode the finding itself. Residualized and incremental results therefore quantify signal beyond the recorded covariates, while full-context adjustment is reported as a conservative over-control sensitivity analysis.

B.12 Debugged Appearance-Control Audit

We performed a post-hoc debug of the appearance-matched control using frozen cached predictions and intervention artifacts only. No model inference, training, or intervention generation was performed for this audit. The validation gate was fixed before evaluating the test split. Each split contains 81 paired samples; the exploratory test split contains 20 patients, and each model scale contributes 243 frozen predictions across the factual, targeted, and control conditions. Every pair preserves patient, image, and finding identity, uses an equal-area non-overlapping control, and was checked for zero overlap. Rigid residual transport was feasible for $74/81$ validation pairs and $77/81$ exploratory-test pairs; the remaining pairs use the documented rank-matched fallback. RGB clipping affected 18.1% of validation control pixels and 16.9% of exploratory-test control pixels.

Strict artifact-only shortcut check. The primary debug classifier uses only the image-difference field and its derived statistics; it does not use source/edited-region texture. The prespecified validation gate requires logistic AUROC at most 0.60 and independent HGB/RF AUROC at most 0.65. Table 10 shows that the gate is satisfied on validation. The exploratory test values are reported only as a post-freeze diagnostic because the canonical test split had already been inspected.

Table 10: Debugged appearance-control shortcut classifiers. The primary scope is image-difference only. Test values are exploratory.

Feature scope	Classifier	Validation	Test[†]
Image difference only	Logistic	0.5688	0.5383
Image difference only	HGB	0.6385	0.5948
Image difference only	RF	0.6236	0.5715
Full-region context	HGB	0.8673	0.8465
Full-region context	RF	0.8599	0.7971

[†]Exploratory because the canonical test split was previously viewed. The full-context logistic classifier is directionally inverted (AUROC 0.0000; separability AUROC 1.0000) and is therefore not interpreted as chance.

Table 11: Exploratory cached CARAE audit on the canonical test split. Raw and residualized values are AUROC with patient-bootstrap 95% intervals. The full-context column is an over-control sensitivity analysis because its features can contain pathology appearance.

Scale	Score	Raw AUROC	Strict residual AUROC	Full-context over-control
3B	Answer	0.8580 [0.8011, 0.9166]	0.7508 [0.6680, 0.8331]	0.5040 [0.4088, 0.6077]
3B	Representation	0.8669 [0.8013, 0.9351]	0.8593 [0.7871, 0.9250]	0.5018 [0.3708, 0.6381]
7B	Answer	0.8332 [0.7643, 0.9118]	0.7734 [0.6850, 0.8663]	0.5207 [0.4171, 0.6318]
7B	Representation	0.8023 [0.7452, 0.8676]	0.7904 [0.7191, 0.8541]	0.4954 [0.4154, 0.5785]

The strict image-difference classifiers support robustness to the measured edit-difference signature. The independent full-region classifiers remain strongly separable, providing a complementary construct-validity result: target and control regions preserve different radiographic content, and the target content can include the pathology being tested. Removing this content to chance would be an over-control. We therefore use the strict scope for primary nuisance adjustment and retain full-context residualization as a deliberately conservative sensitivity bound.

Frozen CARAE residualization audit. Table 11 reports the cached exploratory-test CARAE scores. Residualization is patient-grouped and cross-fitted. "Strict residual" adjusts for finding identity, anatomy, area, pixel-change statistics, boundary changes, and higher-order difference-map texture. "Full-context over-control" also adjusts for source/edited-region texture and is intentionally interpreted conservatively.

The strict residualized answer AUROC remains above chance for both scales, while the full-context over-control is near chance. We interpret this pattern as evidence that CARAE retains target-specific responsiveness after removing measured edit artifacts, with the important qualification that remaining source/edited-region context is not fully ruled out.

Ordinary factual QA diagnostic. The finding-balanced factual QA diagnostic is reported separately from intervention responsiveness. Validation-calibrated balanced accuracy is 0.4681 for 3B and 0.5000 for 7B; at the fixed 0.5 threshold both models answer "Yes" for every test item. The exploratory test AUROC is 0.4717 for 3B and 0.5183 for 7B. This diagnostic distinguishes the controlled evidence-responsiveness claim from general-purpose factual diagnosis.

Because the canonical test split was previously inspected before this debugged run, all test quantities in this subsection are explicitly exploratory rather than fresh confirmatory evidence. A confirmatory version should freeze the protocol on an unseen patient split or an external dataset before test access.

B.13 External Mixed-Polarity Factual Validation

We complement the positive localized benchmark with a small external mixed-polarity validation subset from the SLAKE test split Liu et al. (2021). The frozen source contains $1{,}061$ image–question groups, of which 355 have natural binary answers (175 Yes and 180 No). We selected 80 Yes and 80 No groups by a deterministic hash of the group identifier, without reading model

Table 12: External mixed-polarity factual QA on a frozen 80/80 SLAKE subset.

Method	Bal. Acc.	Sens.	Spec.	Yes rate
Text-only	0.6750	0.5875	0.7625	0.4062
Full CARAE	0.6750	0.5875	0.7625	0.4062

Table 13: Validation-frozen natural factual answer-path calibration.

Answer path	Acc.	Bal. Acc.	Sens.	Spec.
Uncalibrated answer path	0.6188	0.6188	0.4500	0.7875
Prompt-calibrated answer path	0.6750	0.6750	0.5875	0.7625
Calibrated minus uncalibrated	0.0563	0.0563	0.1375	-0.0250

predictions. The natural factual block uses the no-support condition; the same image–question–answer groups are then evaluated under supporting, misleading, and irrelevant evidence contexts. This makes the label balance external and natural while keeping the evidence stress protocol aligned with the Full CARAE evidence-path evaluation. Here, Full CARAE denotes the same complete model used throughout the paper, not a separate router or R1 variant. Text-only and the Support-only baseline are evaluation-path baselines.

The experiment reuses completed Qwen2.5-VL-3B answer caches and holds the validation-selected answer path fixed. On the identical frozen 80/80 subset, calibration increases accuracy and balanced accuracy from 0.6188 to 0.6750 ($+5.63$ percentage points). With this path fixed, Text-only and Full CARAE both obtain 0.6750 on the same no-support factual input (Table 12). The evidence-specific comparison is therefore the stress evaluation below. Table 14 adds the evidence-stress view. On supporting evidence, Full CARAE reaches 0.7375 accuracy, compared with 0.6750 for Text-only. On misleading evidence, the Support-only baseline falls to 0.5250 and selects the evidence-conditioned path on every row, whereas Full CARAE retains 0.6750 accuracy and selects that path on none of the rows in this subset. The corresponding Text-only selection rate is also 0.0000. These selection rates are routing diagnostics, not claims of exact textual answer copying. The balanced factual block broadens label coverage, while the evidence-stress block tests whether the complete model benefits from supporting context without blindly following misleading context. We use this small cached 3B experiment as complementary evidence for evidence-path behavior.

B.14 Scope and Interpretation

The experiments evaluate measured-nuisance-adjusted target-versus-control intervention responsiveness under a localized chest-radiograph protocol. The canonical benchmark supplies dense localized intervention pairs; the mixed-polarity supplement broadens factual-label coverage; and the multi-scope audit separates direct edit signatures from regional radiographic content. Results are patient-clustered and averaged over three seeds, with the canonical test set containing 20 independent patients. The target-control behavioral interface is defined by the canonical objective, while held-out control families and replacement operators test transfer with checkpoints fixed.

The results support claims about:

1. the distinction between factual preservation and targeted-versus-control responsiveness;
2. the contribution of mask-free question-conditioned evidence aggregation and coupling;
3. the mechanism-level refinement provided by evidence-specific transition learning;
4. representation-level transition separation and its association with answer behavior;
5. frozen transfer to the evaluated external chest-radiograph datasets; and
6. measured-nuisance-adjusted target-versus-control intervention responsiveness, supported by a multi-scope audit that separates direct edit signatures from regional appearance.

These results concern the evaluated intervention interface and do not establish unrestricted clinical accuracy, cross-architecture robustness, or causal mediation by the projected state. The MS-CXR 7B evaluation further shows that improved evidence discrimination can coexist with reduced factual preservation under distribution shift.

Table 14: Mixed-polarity evidence stress on the same external subset.

Condition	Text-only	Support-only baseline	Full CARAE
No Support	0.6750	0.6750	0.6750
Supporting	0.6750	0.6875	0.7375
Misleading	0.6750	0.5250	0.6750
Irrelevant	0.6750	0.6125	0.6375

Table 15: Method-specific adaptation and optimization settings.

Method	LoRA Targets	Rank	Scale	Dropout	Evidence Projector	EST
Standard SFT	q,k,v,o	8	16	0.05	No	No
Response-only	q,k,v,o	8	16	0.05	No	No
CF-VLM-style	q,k,v	16	32	0.10	No	No
Global injection	q,k,v,o	8	16	0.05	Global	No
Matched Coupling	q,k,v,o	8	16	0.05	Question-conditioned	No
Full CARAE	q,k,v,o	8	16	0.05	Question-conditioned	Yes

B.15 Mask-Free Implementation and Shortcut Boundary

The formal mask-free implementation constructs question-conditioned weights from visual tokens and the question-token embedding. The same weights form the pooled state and modulate the state-conditioned residual at each visual token. Localization masks are used offline to construct and evaluate interventions, but are not inputs to the mask-free evidence encoder or answer path. A source-level audit and checkpoint-reload check cover the completed 3B/7B formal artifacts, each with 243 prediction rows.

On the frozen test split, the multi-scope audit gives metadata-only AUROC 0.7688 at both scales, while grouped cross-fitted metadata-residualized CARAE reaches 0.6342 and 0.7171 for 3B and 7B. Under the stronger appearance-matched adjustment, the corresponding answer AUROC is 0.6313 and 0.6949. These values quantify response beyond the recorded nuisance scopes. The legacy mask-conditioned runner is kept separate; the reported Full CARAE results use the mask-free branch.

C Additional Mechanism Analyses

This appendix provides additional diagnostics for the representation–answer coupling learned by CARAE. All method names and checkpoint assignments follow the mapping used in the main-paper tables. These analyses complement, rather than replace, the primary answer-level and shortcut-adjusted evaluations.

C.1 Visual–Answer Coupling

Table 16 expands the coupling analysis reported in the main paper. For both backbones, the mask-free evidence-aware variants exhibit strong positive associations between projected visual specificity and answer specificity. These associations remain substantial after rank residualization against the recorded intervention metadata, indicating that the representation-level and answer-level responses are not explained solely by the measured intervention statistics.

On the 3B backbone, Matched Coupling attains a partial correlation of 0.729, whereas Full CARAE increases this value to 0.866. The corresponding confidence intervals, $[0.512, 0.837]$ and $[0.761, 0.921]$, remain clearly above zero. This result indicates that EST improves the sample-wise alignment between the projected transition and the answer response, in addition to increasing the targeted-over-control projected gap.

The same pattern is observed on the 7B backbone. Matched Coupling already exhibits a positive metadata-adjusted association of 0.645, while Full CARAE further increases the partial correlation to 0.813. The corresponding confidence intervals are $[0.416, 0.802]$ and $[0.595, 0.889]$, respectively.

Table 16: Detailed visual–answer coupling on the canonical test set. Raw ρ is the Spearman association between projected visual specificity and answer specificity. Partial ρ is computed after rank residualization against the recorded intervention metadata. Confidence intervals correspond to the partial association. Reported statistics follow the same three-seed aggregation protocol as the main evaluation, and N denotes the number of image–finding pairs.

Model	Method	Raw ρ	Partial ρ	95% CI	N
3B	w/o EST	0.794	0.729	[0.512, 0.837]	81
3B	Full CARAE	0.909	0.866	[0.761, 0.921]	81
7B	w/o EST	0.867	0.645	[0.416, 0.802]	81
7B	Full CARAE	0.928	0.813	[0.595, 0.889]	81

Table 17: Frozen-transfer AUROC gains of Full CARAE over response-only training. Confidence intervals use the available grouped bootstrap protocol for each dataset.

Setting	Δ AUROC	95% CI
MS-CXR, 3B	+0.0666	[0.0522, 0.0808]
MS-CXR, 7B	+0.0769	[0.0644, 0.0892]
VinDr-CXR, 3B	+0.0359	[0.0235, 0.0485]

Thus, EST strengthens visual–answer coupling on both backbones, with the effect extending beyond a single model scale.

Together with the projected-gap and effective-rank results in the main paper, this analysis separates three related but distinct properties: transition separability, transition-space organization, and sample-level alignment with answer behavior. Full CARAE improves both transition separation and behavioral alignment across the evaluated backbones, while effective rank remains a descriptive measure of how the transition signal is distributed across dimensions rather than a metric with a universally preferred direction.

C.2 Trainable Parameter Analysis

To examine whether the improvements of CARAE are attributable to additional optimization capacity, we report trainable parameter statistics for all adaptation methods in Table 19. All methods use parameter-efficient adaptation, with the frozen vision-language backbone excluded from optimization.

Standard SFT and response-only training optimize the same LoRA parameters, while injection-based variants additionally introduce lightweight evidence coupling modules. Full CARAE introduces only a negligible number of additional parameters compared with Matched Coupling: the two configurations differ by only 769 trainable parameters for both Qwen2.5-VL-3B and Qwen2.5-VL-7B. The 769-parameter difference is too small to explain the observed comparison by capacity alone. This controlled comparison supports an additional benefit from the transition-learning objective.

Overall, Full CARAE trains 5.87M and 8.81M parameters on the 3B and 7B backbones, corresponding to only 0.156% and 0.106% of the total model parameters, respectively. This is a lightweight adaptation under the evaluated target-versus-control protocol.

C.3 Projected Transition Geometry

Figure 4 qualitatively visualizes the factual-to-targeted and factual-to-control transitions in the projected evidence space. Within each backbone, Matched Coupling and Full CARAE are projected using a shared PCA basis, enabling direct comparison of their relative transition geometry.

The visualization complements the quantitative results in Table 3. Both variants exhibit distinct targeted and matched-control transition patterns, while Full CARAE achieves larger projected gaps

Table 18: Gate and residual-injection diagnostics for the Qwen2.5-VL-3B terminal checkpoints. Gate–answer and gate–area values are Spearman correlations. All statistics follow the same three-seed aggregation protocol as the main evaluation; mean and standard deviation summarize the resulting sample-level diagnostics. These quantities characterize learned coupling behavior and are not interpreted as causal mediation effects.

Evaluation	Method	Gate Mean ± Std	Injection Norm Mean ± Std	Gate–Answer ρ	Gate–Mask Area ρ
Canonical	Global injection	0.143 ± 0.003	5.97 ± 0.18	-0.204	0.029
Canonical	Matched Coupling	0.156 ± 0.016	6.29 ± 0.89	-0.185	-0.096
Canonical	Full CARAE	0.181 ± 0.053	6.47 ± 2.28	0.388	0.352
VinDr-CXR	Global injection	0.140 ± 0.002	5.81 ± 0.16	0.020	0.043
VinDr-CXR	Matched Coupling	0.162 ± 0.013	6.61 ± 0.73	-0.026	-0.477
VinDr-CXR	Full CARAE	0.181 ± 0.031	6.32 ± 1.23	0.023	0.333

Table 19: Trainable parameter statistics of adaptation methods. The frozen backbone parameters are excluded from optimization.

Method	Backbone	Params	Ratio
SFT	Qwen2.5-VL-3B	3.69M	0.098%
Response-only	Qwen2.5-VL-3B	3.69M	0.098%
Global Injection	Qwen2.5-VL-3B	5.87M	0.156%
Matched Coupling	Qwen2.5-VL-3B	5.87M	0.156%
Full CARAE	Qwen2.5-VL-3B	5.87M	0.156%
SFT	Qwen2.5-VL-7B	5.05M	0.061%
Response-only	Qwen2.5-VL-7B	5.05M	0.061%
Global Injection	Qwen2.5-VL-7B	8.81M	0.106%
Matched Coupling	Qwen2.5-VL-7B	8.81M	0.106%
Full CARAE	Qwen2.5-VL-7B	8.81M	0.106%

and stronger metadata-adjusted visual–answer coupling in the original representation space. PCA is used only for visualization; all reported mechanism metrics are computed before dimensionality reduction.

C.4 Gate and Token-Weighted Residual-Injection Behavior

Table 18 reports descriptive statistics of the sample-specific gate, residual direction, and token-weighted update. On the canonical test set, Full CARAE exhibits a larger mean gate and injection norm than the two ablated variants, together with a positive gate–answer association. This pattern indicates that EST changes not only the projected transition space, but also how the learned evidence state modulates the generation pathway.

The behavior is less uniform under VinDr-CXR transfer. Full CARAE retains the largest mean gate, whereas Matched Coupling produces a slightly larger average injection norm. Gate–answer correlations are close to zero for all three methods, suggesting that raw gate magnitude alone does not explain the transferred answer behavior.

Gate–mask-area associations vary substantially across methods and datasets, including a negative association for Matched Coupling on VinDr-CXR. Because localization masks are not supplied to the model forward pass, these correlations reflect post-hoc associations between the learned gate and intervention geometry rather than direct mask conditioning. They further motivate interpreting the gate only as a descriptive coupling diagnostic rather than as evidence of causal mediation or spatial grounding.

As an auxiliary diagnostic, we additionally combine factual `Yes` and targeted `No` conditions into the prespecified intervention-conditioned binary task. Under this diagnostic, LogitNorm obtains NLL/ECE values of $5.0876/0.4999$ for 3B and $5.8912/0.5000$ for 7B. Thus, its near-zero positive-

Table 20: Small-scale visual-space sanity check. The experiment verifies activation of the evidence-transition objective and excludes immediate representation collapse. It is not used as a replacement for the formal canonical evaluation.

Setting	N	**Normalized Gap**	**Visual Accuracy**	**Effective Rank**
EST check	64	0.5645	0.9330	6.58

Table 21: Auxiliary intervention-conditioned calibration diagnostic for LogitNorm. The task combines factual `Yes` and targeted `No` conditions. It is reported only to clarify the interpretation of positive-anchor NLL/ECE.

Backbone	Paired-task NLL ↓	Paired-task ECE ↓
Qwen2.5-VL-3B	5.0876	0.4999
Qwen2.5-VL-7B	5.8912	0.5000

anchor NLL/ECE arises from highly confident retention of the factual `Yes` answer rather than successful adaptation to targeted evidence removal. Because the targeted labels are behavioral intervention targets rather than independently adjudicated clinical diagnoses, this paired-task diagnostic is not treated as a primary calibration metric.

D Label-Free Target–Control Evidence Responsiveness

The canonical evaluation protocol assigns the targeted intervention a protocol-defined response target. A potential concern is that the edited targeted image is not a clinically adjudicated negative radiograph. To verify that the conclusions do not depend on this protocol label, we perform an additional label-free analysis that never uses the edited answer label. For each image–finding pair, let m^f, m^t, and m^c denote the factual, targeted, and matched-control Yes/No margins. We define the targeted and control response magnitudes as

$$g^t = m^f - m^t, \qquad g^c = m^f - m^c,$$

and the evidence gap

$$\Delta = g^t - g^c.$$

Positive values indicate that removing the localized evidence induces a larger answer-margin change than an equally strong matched-control intervention. The protocol-defined edited answer is never used during this analysis.

The evaluation reuses the cached predictions from the main experiments without additional model training or inference, and patient-group bootstrap follows the same protocol as the main paper. Table 22 shows that the conclusions remain unchanged after removing the protocol-defined edited answer labels from the evaluation. Response-only training exhibits only limited separation between targeted and matched-control interventions. Adding evidence injection substantially enlarges the evidence gap, while the complete CARAE model consistently achieves the strongest target–control discrimination and the largest average evidence gap on both Qwen2.5-VL-3B and Qwen2.5-VL-7B. These results indicate that the improvements are not explained solely by the protocol-defined edited answers, but reflect stronger relative responses to localized evidence removal. Figure 5 visualizes the distribution of sample-level evidence gaps. Compared with response-only training, Matched Coupling shifts the distribution toward positive values, whereas Full CARAE further increases both the median and the overall distribution shift across both backbones. Rather than relying on a small number of extreme examples, the improvement is observed for the majority of evaluation pairs. Figure 6 further reports the empirical cumulative distributions of evidence gaps. The CARAE curves are consistently shifted to the right of all baselines, indicating that the larger target-versus-control response occurs across many evaluated samples rather than isolated outliers.

Overall, this analysis does not interpret the edited targeted image as a clinically validated negative study. Instead, it evaluates whether removing the localized evidence induces a larger model response than an equally strong matched-control intervention.

Table 22: Label-free target–control responsiveness. The analysis directly compares targeted and matched-control interventions without using protocol-defined edited answer labels.

Method	Paired Acc. ↑	Target–Control AUROC ↑	Mean Evidence Gap ↑
Qwen2.5-VL-3B			
Response-only	0.531	0.514	0.032
Matched Coupling	0.815	0.769	0.324
Full CARAE	**0.938**	**0.799**	**0.719**
Qwen2.5-VL-7B			
Response-only	0.704	0.707	0.199
Matched Coupling	0.901	0.816	1.321
Full CARAE	**0.951**	**0.828**	**1.938**

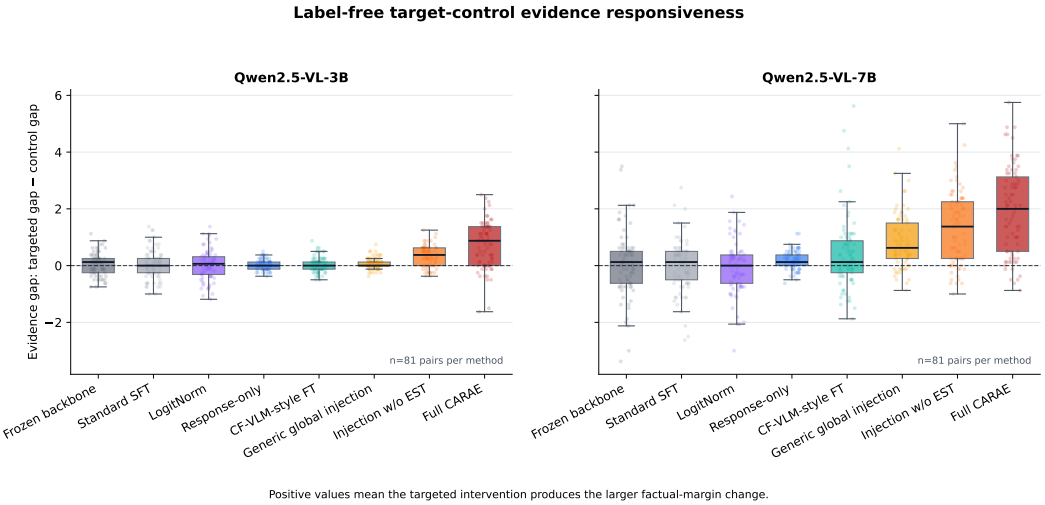


Figure 5: Distribution of label-free evidence gaps ($g^t - g^c$). Positive values indicate that removing the localized evidence induces a larger answer-margin change than an equally strong matched-control intervention. CARAE consistently shifts the distribution toward larger positive values on both Qwen2.5-VL-3B and Qwen2.5-VL-7B.

The label-free results reinforce Full CARAE's measured-nuisance-adjusted target-versus-control responsiveness without relying on the protocol-defined answer labels. Together with held-out controls and replacement operators, they support transfer across intervention realizations while preserving the distinction between direct edit signatures and regional radiographic content.

D.1 Qualitative Success and Failure Analysis

Figure 3 presents representative sample-level behaviors under the paired-intervention protocol. Each row contains the factual image with the annotated finding contour, the targeted intervention, the matched control, and the corresponding positive-answer margins.

The successful case shows the intended behavior: targeted evidence removal suppresses the positive answer, whereas the matched control preserves it. The remaining examples illustrate three limitations. A frozen model can retain the correct factual answer while responding almost identically to targeted and control interventions, corresponding to counterfactual rigidity. A matched control can also induce a larger answer change than targeted removal, producing negative answer specificity. Finally, under external transfer, a positive projected transition gap may coexist with negative answer specificity, showing that representation-level separation does not guarantee aligned generation behavior.

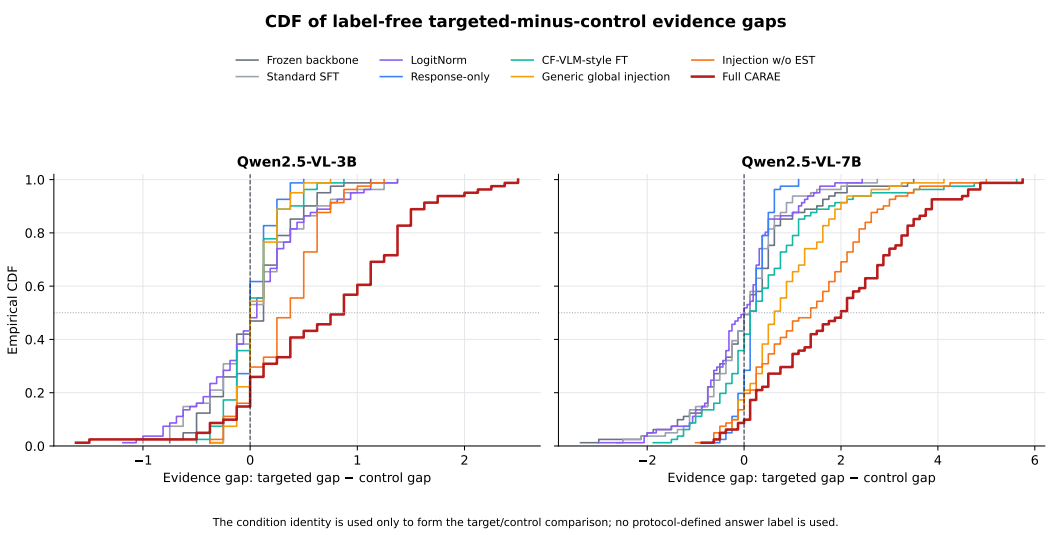


Figure 6: Empirical cumulative distributions of the label-free evidence gap ($g^t - g^c$). The comparison directly evaluates targeted versus matched-control interventions without using protocol-defined edited answer labels. CARAE exhibits a consistent rightward shift relative to the baselines, indicating a larger target-versus-control gap across evaluation samples under this protocol.

These cases motivate evaluating factual preservation, representation specificity, and answer specificity separately. They also show that residual anatomical and intervention differences may still affect individual predictions despite the matched-control construction and shortcut-adjusted evaluation.

E Robustness to Alternative Evidence-Removal Operators

The canonical experiments use per-channel global-mean filling to remove the annotated target region and its matched control. A potential concern is that CARAE may specialize to the appearance of this particular intervention rather than learning a response that transfers across different realizations of localized evidence removal. We therefore perform a frozen cross-operator evaluation using the terminal canonical checkpoints. No additional training, checkpoint selection, threshold adjustment, or dataset-specific calibration is performed. These operators are held out from canonical optimization: the model is trained only with the global-mean realization, and all alternative operator results are obtained after checkpoint freezing.

E.1 Alternative Intervention Operators

We retain the same factual images, target masks, matched-control masks, questions, and answer-scoring protocol as in the canonical evaluation. Only the image-replacement operator is changed. We consider three alternative operators:

- **Constant gray:** pixels inside the intervention region are replaced by a fixed neutral gray value;
- **Local-ring mean:** the fill value is estimated from an eight-pixel ring outside the intervention boundary, with deterministic fallback to the global unmasked mean when insufficient ring pixels are available;
- **Local texture noise:** the intervention region is replaced by fixed-seed noise matched to the mean and variance of the surrounding local ring.

For every operator, the same construction is applied to the targeted and matched-control regions. The operator identity is never provided to the model. The factual branch always receives the original

Table 23: Frozen robustness to alternative localized evidence-removal operators. Full CARAE values are mean ± standard deviation over three seeds. Δ AUROC compares the corresponding cached Full CARAE and response-only seed-0 evaluations; 95% CIs use 2,000 paired patient-grouped bootstrap replicates.

Scale	Operator	AUROC ↑	Spec. Gap ↑	Proj. Gap ↑	Δ AUROC [95% CI]
3B	Gray	0.7703 ± 0.0275	0.1772 ± 0.0564	0.3225 ± 0.0407	$+0.2449\,[0.1672, 0.3282]$
	Local mean	0.7044 ± 0.0165	0.1068 ± 0.0266	0.2075 ± 0.0212	$+0.2149\,[0.1244, 0.3190]$
	Local texture	0.5770 ± 0.0251	0.0320 ± 0.0099	0.1037 ± 0.0494	$+0.0688\,[0.0303, 0.1071]$
7B	Gray	0.7961 ± 0.0446	0.5980 ± 0.1830	0.2848 ± 0.0680	$+0.1680\,[0.0912, 0.2463]$
	Local mean	0.7593 ± 0.0376	0.4732 ± 0.1203	0.2291 ± 0.0339	$+0.1486\,[0.0811, 0.2257]$
	Local texture	0.6665 ± 0.0296	0.1806 ± 0.0527	0.1723 ± 0.0020	$+0.2007\,[0.1204, 0.2866]$

unmodified radiograph. These experiments therefore evaluate frozen transfer across replacement realizations rather than robustness to arbitrary image corruption.

E.2 Cross-Operator Results

Table 23 shows that the learned target-versus-control response is not restricted to the canonical global-mean intervention. Without additional training, Full CARAE retains positive targeted-over-control discrimination under all three replacement operators and at both model scales.

For Qwen2.5-VL-3B, Full CARAE obtains answer AUROCs of 0.7703, 0.7044, and 0.5770 under constant-gray, local-ring-mean, and local-texture replacement, respectively. The corresponding improvements over response-only training are $+0.2449$, $+0.2149$, and $+0.0688$. All three patient-grouped confidence intervals exclude zero.

The same pattern holds for Qwen2.5-VL-7B. Full CARAE reaches AUROCs of 0.7961, 0.7593, and 0.6665, improving over response-only by $+0.1680$, $+0.1486$, and $+0.2007$, respectively. The gains remain positive when the replacement value is estimated locally and when the intervention contains locally matched stochastic texture.

The projected transition gap also remains positive for every reported operator. This indicates that the frozen projected evidence state continues to respond more strongly to targeted replacement than to the matched non-target intervention, even when the pixel-level realization differs from the canonical training operator. Factual positive-answer accuracy remains 1.0000 for the 3B checkpoints and 0.9959 for the 7B checkpoints across the evaluated replacements, since the factual branch is kept unmodified.

These results do not imply robustness to arbitrary image corruptions. Rather, they show that CARAE's target-versus-control response transfers across multiple localized replacement operators and is not explained solely by the exact per-channel global-mean fill used in canonical training.

E.3 Visual-Space Sanity Check

Before the full canonical experiment, we conducted a small-scale visual-space check to verify that the evidence-specific transition objective was active and did not immediately collapse the projected representation. As shown in Table 20, targeted removal produced a positive normalized transition gap and was correctly ordered ahead of the matched control for most evaluated pairs.

The transition representation also retained a non-trivial effective rank, indicating that the learned difference was not reduced to an identical constant response across samples. This experiment serves only as an implementation and mechanism sanity check. Because it uses a restricted sample and training setting, it is not used to claim formal training-seed stability or to replace the patient-grouped canonical results.