FINITE CLINICAL PROGRAMS: AN INSPECTABLE DISCRETE INTERFACE FOR RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study radiology report generation with a complete case-dependent conditioning message reconstructible from discrete IDs. Finite Clinical Subspace Learning (FCSL) combines named clinical states, task and prior-study fields, and quantized residual codes, using 202 bits on MIMIC-RG4 and 160 bits on MIMIC-ABN. A 0.8B language backbone accesses this message through a prefix and layer-local writes. Training uses reference-derived states; inference uses predicted IDs. Three-seed system comparisons improve clinical-label agreement but yield lower language and RadGraph scores. A supervision-matched continuous interface attains similar micro-F1 and higher complementary scores. However, matched RG4 interventions on 3,696 shared eligible requests from 245 patients yield 33.05% joint selective success for FCSL versus 18.60% for continuous conditioning: a paired difference of 14.44 percentage points (95% patient-bootstrap interval [11.87, 16.63]). Joint success requires the requested target change with all thirteen non-target labels unchanged. FCSL also exceeds prefix-only's 29.63%, without a demonstrated reduction in non-target changes. These single-seed, exploratory interventions reveal a report-quality–selective-response trade-off on the common eligible subset, rather than universal accuracy gains or clinical safety. FCSL makes complete conditioning reconstructible and its editing benefits and limits measurable at a compact backbone scale.

1 INTRODUCTION

Radiology report generation must infer clinical findings and express them in language. Flexible generators combine current images with optional views, prior studies, or context through continuous representations (Wang et al., 2025; Yu et al., 2026; Liu et al., 2025). An omitted finding may reflect failed encoding or failed realization; without named intermediate fields, these errors are difficult to distinguish.

We ask what is gained and lost when the decoder's *complete* case-dependent message must be reconstructible from discrete IDs (Figure 1). Named clinical states expose decisions to inspect or modify; quantized residuals retain bounded capacity beyond that checklist. Unlike diagnostic labels alongside an unrecorded feature vector, the tuple specifies all conditioning atoms. Only named fields have prescribed clinical meanings. Whether modifying them produces selective text changes is an empirical question, separate from reconstruction.

Finite Clinical Subspace Learning (FCSL) implements this interface as a finite clinical program. A query encoder predicts fourteen four-state observations and quantized residuals; task and available prior-study fields complete the tuple. On MIMIC-RG4, 202 fixed-width bits reconstruct 52 atoms, including derived changes; MIMIC-ABN uses 160 bits and 24 atoms. The same reconstruction supplies a 0.8B language model through a global prefix and rank-32 writes at eight layers, without a patient-specific evidence bypass.

The backbone is 11.4% of LLM-RG4's nominal 7B decoder size (Wang et al., 2025); the full system has 1.00B parameters, with 79.75M trainable. Continuous controls share this compact backbone. Reference-derived training states and predicted inference states leave a conditioning mismatch.

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.

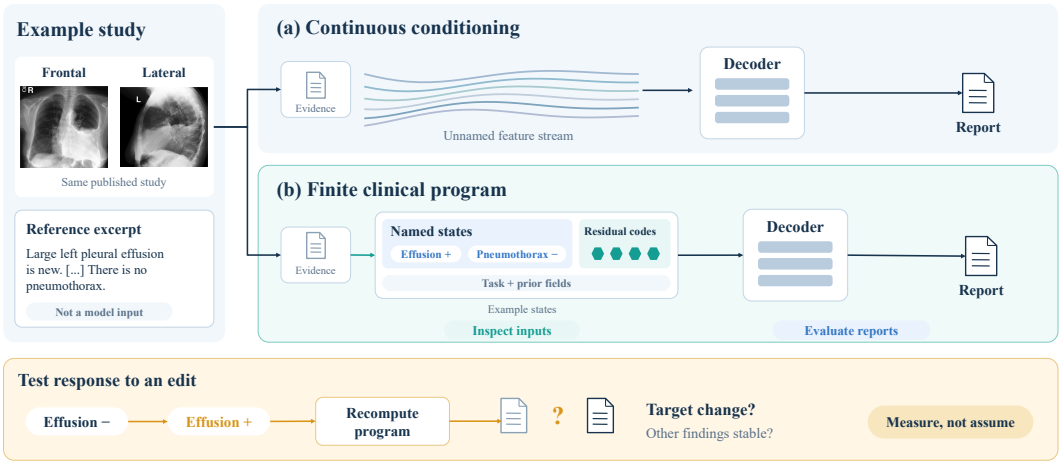

Figure 1: **From opaque conditioning to inspectable report generation.** An omitted finding can arise during evidence encoding or report realization. Named clinical states expose this distinction, while residual IDs complete the finite conditioning message. Paired state edits then test whether the report changes the requested finding while preserving non-target findings.

Matched experiments expose a report-quality–selective-response trade-off. Continuous conditioning attains similar micro-F1 and higher language/RadGraph scores, but Full achieves 33.05% joint selective success versus 18.60% for continuous and 29.63% for prefix-only on 3,696 shared eligible RG4 edit requests. These single-seed results distinguish generation quality from editing response, without isolating quantization's effect or establishing clinical safety.

Our contributions are:

1. A complete finite interface combining clinical states, task and prior-study fields, and residual IDs, with explicit budgets and exact reconstruction of all conditioning atoms.
2. A compact 0.8B realization using a global prefix and layer-local rank-bounded writes, with consistent program edits across both routes and no evidence bypass.
3. Matched evaluations separating report quality from selective response, showing higher joint editing success for Full alongside lower generation scores.

2 Related Work

Radiology report generation. R2Gen uses generation memory (Chen et al., 2020), while cross-modal memory aligns images and text (Chen et al., 2021). MAIRA-2 adds finding localization (Bannur et al., 2024); LLaVA-Rad studies specialized 7B reporting (Zambrano Chaves et al., 2025). LLM-RG4 and AnchorDiff accommodate variable views and prior reports (Wang et al., 2025; Yu et al., 2026); PriorRG integrates longitudinal context (Liu et al., 2025). FCSL focuses on a different design axis: a complete finite conditioning message with measurable selective-editing behavior.

Concept representations and interventions. Concept-first generation (Nooralahzadeh et al., 2021) and CBM-RAG (Alam et al., 2025) already introduce concepts into reporting. Concept bottlenecks expose supervised variables (Koh et al., 2020); CEM expands their capacity (Espinosa Zarlenga et al., 2022), and IntCEM trains intervention response (Espinosa Zarlenga et al., 2023). FCSL instead encodes *all* case-dependent conditioning as named states and quantized residual IDs (van den Oord et al., 2017). Its distinction is this complete categorical interface, not concept supervision or editing alone. Continuous tensors are also storable and editable; residual IDs need not have independent clinical semantics. No superiority over CEM or IntCEM is established.

Parameter and representation adaptation. Low-dimensional adaptation (Aghajanyan et al., 2021), LoRA (Hu et al., 2022), ReFT (Wu et al., 2024), and multimodal subspaces (Tonekaboni et al., 2026) constrain parameters or hidden updates. Prefix-tuning learns task-specific continuous

prompts (Li & Liang, 2021); our input prefix is reconstructed per case. FCSL couples global prefix access and local writes to the same finite message. Alphabet size, write rank, and shared adaptation parameters remain distinct constraints, not a common compression guarantee.

3 Finite Clinical Programs

3.1 From evidence to a finite program

Let $x = (I_f, I_l, I_p, R_p, C, s, \sigma)$ contain the available current, lateral, and prior images, prior-report text, clinical context, scenario s, and requested section σ. FCSL constructs an ID tuple $\iota_\theta(x)$ and reconstructs program atoms $Z = g_\theta(\iota)$ before generating

$$p_\theta(y \mid x) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, p_0, g_\theta(\iota_\theta(x))). \tag{1}$$

The program is a structured conditioning message, not an executable clinical rule set. The fixed instruction p_0 carries no patient-specific evidence; scenario and section are task-provided fields. Both access routes in Figure 2 consume the ID reconstruction, with no image or patient-text bypass. Shared parameters and calibration thresholds belong to the model, not the per-case budget.

3.2 Estimating clinical states from evidence

Chest images share a RAD-DINO encoder (Pérez-García et al., 2024), with its final four blocks fine-tuned; CXR-BERT (Boecking et al., 2022) remains frozen. Source projections and embeddings distinguish current, lateral, prior-image, prior-text, and context memories. Availability masks exclude missing sources.

Four blocks update fourteen clinical and eight residual queries (width 512) through eight-head self-attention, cross-attention to evidence, and feed-forward layers (Appendix B.1). Each clinical concept has an independent four-state head,

$$\ell_c = q_c W_c + b_c, \qquad p_{c,k} = \mathrm{softmax}(\ell_c)_k, \quad k \in \{0,1,2,3\}, \tag{2}$$

corresponding to blank, positive, negative, and uncertain. A validation-fitted threshold t_c first selects an endpoint group, after which the more probable state in that group is retained:

$$\widehat{s}_c = \begin{cases} \arg\max_{k\in\{1,3\}} p_{c,k}, & p_{c,1} + p_{c,3} > t_c, \\ \arg\max_{k\in\{0,2\}} p_{c,k}, & \text{otherwise.} \end{cases} \tag{3}$$

Thresholds maximize endpoint micro-F1 on validation data; patient-level cross-fitting and deployment refitting are specified in Appendix B.2.

3.3 Message fields and exact reconstruction

Residual and task information. Each residual query is projected to width 256 and residual-quantized (Lee et al., 2022) using two shared, 256-entry codebooks:

$$r_{j,1} = P_r q_j^{\mathrm{res}}, \quad k_{j,a} = \arg\min_k \|r_{j,a} - v_{a,k}\|_2^2,$$
$$r_{j,a+1} = r_{j,a} - v_{a,k_{j,a}}, \quad \widehat{r}_j = \sum_{a=1}^{2} v_{a,k_{j,a}}. \tag{4}$$

Eight residual atoms thus use sixteen categorical IDs. Here, residual refers to the remaining quantization error, not a subtraction of clinical-state embeddings. Scenario $s \in [4]$ records evidence availability; section $\sigma \in [3]$ requests findings, impression, or both, where $[n] = \{0, \ldots, n-1\}$.

For RG4, prior-report labels $s_c^p \in [5]$ use the four clinical states plus an unavailable symbol. They are obtained by offline CheXbert labeling of legitimate prior text, independently of the current reference report. SN/MN use unavailable prior symbols. A derived change state $\delta_c = h(\widehat{s}_c, s_c^p)$ indicates new, resolved, stable endpoint, or unavailable/unmentioned status. It is recomputed from the transmitted current and prior IDs. ABN omits the prior and change blocks because its input manifests contain no legitimate prior report to label.

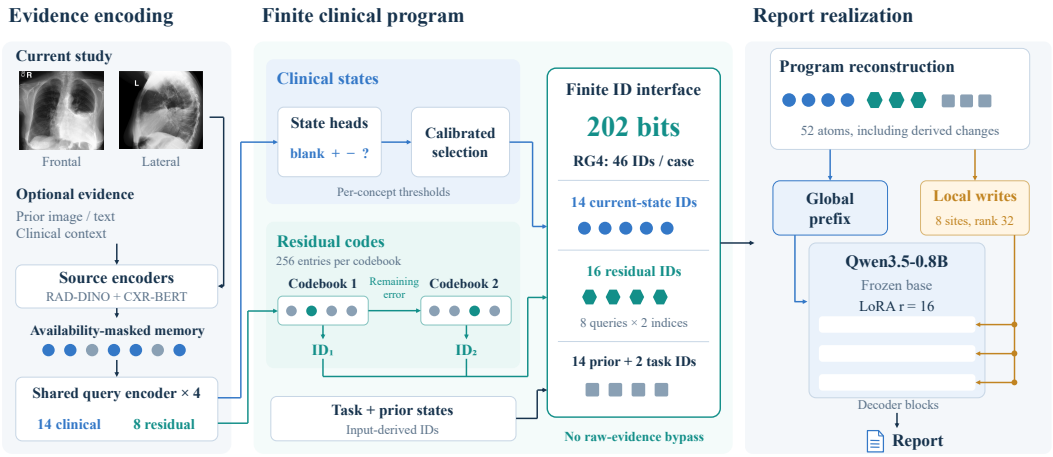

Figure 2: **Finite clinical program architecture.** Shared query blocks produce clinical states and residual codes; task and prior fields complete the IDs. Their reconstruction supplies the prefix and eight rank-32 writes. The diagram shows predicted-state inference; report training supplies reference-derived states (Section 3.5).

Alphabet and atoms. The RG4 tuple is

$$\iota = (s, \sigma, \widehat{s}_{1:14}, s^p_{1:14}, \{k_{j,a}\}), \quad \mathcal{I} = [4] \times [3] \times [4]^{14} \times [5]^{14} \times [256]^{16}. \tag{5}$$

Its 46 IDs reconstruct 52 atoms: one scenario, one section, fourteen current states, fourteen prior states, fourteen derived changes, and eight residual atoms. With learned embeddings and slot identities A,

$$Z = \mathrm{LN}\Big([e_s; e_\sigma; U(\widehat{s}); V(s^p); D(h(\widehat{s}, s^p)); \widehat{R}] + A\Big) \in \mathbb{R}^{52\times 256}. \tag{6}$$

The independent fixed-width budget is

$$b_{\mathrm{RG4}} = 2 + 2 + 28 + 42 + 128 = 202 \text{ bits}. \tag{7}$$

Derived changes add no independent bits. ABN uses 32 IDs, 24 atoms, and $2 + 2 + 28 + 128 = 160$ bits. These are fixed-width ID budgets, not embedding sizes or measured entropy. RG4 retains 202 bits even with unavailable prior IDs; joint log-cardinality bounds appear in Appendix A.2.

Reconstruction and editing. Every conditioning atom is reconstructed from IDs. For fixed shared parameters, instruction, and generated prefix, equal IDs give equal next-token distributions. Thus the tuple is sufficient to reconstruct the conditioning message without retaining patient-specific continuous features. Editing a current-state ID recomputes its change atom and updates both routes. These are representation guarantees, not guarantees of correct states or selective text changes; the latter are evaluated through paired edits. The ID budget excludes shared weights and does not measure runtime memory or privacy.

3.4 Global prefix reads and local writes

For input-level access, a learned affine map projects all program atoms to decoder width:

$$P(Z) = ZW_p + b_p, \qquad H_0 = [P(Z); \mathrm{Embed}(p_0, y_{<t})]. \tag{8}$$

Ordinary causal attention gives the decoder global access to this prefix. For layer-level access, an additional branch at eight evenly spaced layers reads the same atoms and writes through a layer-specific rank-32 basis $U_l = \mathrm{QR}(A_l)$:

$$\begin{aligned} C_l &= \mathrm{MHA}_l(H_l, Z, Z), \\ D_l &= f_l(C_l U_l) \odot \mathrm{sigmoid}(G_l(H_l U_l)), \\ H'_l &= H_l + m_l \odot [\mathrm{sigmoid}(\alpha_l) D_l U_l^\top]. \end{aligned} \tag{9}$$

The single joint read uses all atoms as keys and values; the hidden state gates the program-derived update. The mask m_l suppresses writes at padded positions, and the initial scalar gate is 0.5. Attention projections, f_l, and G_l are bias-free.

Each direct additive write lies in $\mathrm{col}(U_l)$, giving a local rank bound of 32. The global prefix and subsequent nonlinear computation remain unrestricted by that local bound. Rank-16 LoRA adapts all linear modules in decoder blocks, including MLP projections; the base language-model parameters remain frozen.

3.5 Learning states and training report realization

Stage A: clinical pretraining and calibration. The clinical encoder and tuned visual blocks undergo two epochs of four-state focal supervision, with training-frequency class weights (Appendix B.2). The inherited checkpoint is selected by uncalibrated validation header micro-F1 and then recalibrated.

Stage B: gold-state report training. Stage B supplies reference-derived current-state prototypes, using a soft predicted mixture only for missing labels. Gold here is not independent expert adjudication. Clinical supervision continues, residuals use vector quantization, and task/prior fields remain input-derived:

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + 0.3\mathcal{L}_{\mathrm{state}} + 0.1\mathcal{L}_{\mathrm{VQ}} + 0.5\mathcal{L}_{\mathrm{usage}} + 0.01\mathcal{L}_{\mathrm{orth}}. \tag{10}$$

LM loss trains next-token realization; VQ commitment and an entropy-floor penalty support residual use; orthogonality regularizes the raw write bases. EMA codebook updates are committed once per optimizer update.

Validation thresholds are refitted after each epoch and saved with that checkpoint before report generation. Inference uses calibrated hard current states and canonical residual IDs. The main model excludes predicted-state adaptation and additional narrative or adversarial objectives; these ablations are evaluated separately (Appendix C.11).

4 Experimental Protocol

4.1 Datasets and comparison scope

MIMIC-RG4 contains 424,411/3,400/6,215 train/validation/test rows (Wang et al., 2025). One model serves SN (current image), SW (+prior report), MN (+lateral image), and MW (+both), with 2,357/2,026/1,004/828 test cases. Section comes from the official instruction; legitimate prior text supplies offline header labels and encoder-only continuous features.

MIMIC-ABN has 69,641/586/844 rows with current/prior radiographs and context (Liu et al., 2025). Without a legitimate prior report, its program omits prior/change fields. The target is *findings*; nonempty factual serialization is an eligibility filter, consistent with the public PriorRG loader. Unmatched cleaning/scoring limits literature comparisons.

FCSL and a section-conditioned continuous-prefix baseline (D+section) use report-training seeds 17/42/73 on both tasks: twelve validation-selected runs. FCSL seeds share Stage-A weights, with seed-specific calibration folds. Development involved prior test-set feedback, so the study remains exploratory.

4.2 Training and checkpoint selection

FCSL uses Qwen3.5-0.8B (Qwen Team, 2026), rank-16 LoRA, eight rank-32 writes, and global batch 128. Gold-conditioned Stage B runs three RG4 or five ABN epochs; optimization details are in Appendix B.3.

Every epoch generates reports for the complete validation split. The selected checkpoint maximizes

$$S_{\mathrm{val}} = 0.5F_{\mathrm{CheXbert}} + 0.3F_{\mathrm{RadGraph\text{-}ERC}} + 0.2\,\mathrm{BLEU\text{-}4}, \tag{11}$$

with inputs on their native $[0, 1]$ scale and ties resolved by the earliest epoch. Within each run, selection is restricted to Stage B and completed before evaluating that run on the test set.

Deterministic decoding uses three beams, a 256-token prompt cap, and at most 200/80 new tokens for RG4/ABN. Validation-selected penalties are frozen for FCSL/D+section (Appendix B.3). All main results use this frozen policy.

4.3 METRICS AND CONTROLS

CheXbert (Smit et al., 2020) compares generated and reference labels, merging positive/uncertain. It also supplies supervision, so agreement is not independent clinical validation. Header metrics compare encoder states with manifest labels. RadGraph entity–relation-complete F1 (Jain et al., 2021; Yu et al., 2023) complements BLEU-1/2/3/4, METEOR, ROUGE-L, and CIDEr (sources and implementation in Appendix B.3). Corpus scores use complete test splits, not scenario averages.

In D+section, D denotes direct continuous-prefix conditioning (128 queries); +section supplies an evidence-free prompt suffix requesting findings, impression, or both. It shares FCSL's visual initialization, decoder/LoRA, training budget, selection, and decoding, but lacks its clinical encoder, state loss, and local writes (Appendix B.4). Unequal supervision and capacity make this a system-level comparison, distinct from the supervision-matched Continuous control.

Matched interface and route controls. A separate RG4 seed-42 batch retrains Full, supervised continuous, prefix-only, and write-only variants for three Stage-B epochs. Full/continuous share clinical supervision, trainable capacity, and both access routes; continuous uses probability-weighted states and unquantized residuals. Single-route variants remove the other route before training. Quantization-related losses and route capacities differ (Appendix C.1). All use the same selection rule and frozen decoding. A separate 4B FCSL/D+section comparison tests backbone scale; it is not a matched-supervision comparison. Neither batch adds seeds to the original main experiment.

External references. LLM-RG4 (Wang et al., 2025), AnchorDiff (Yu et al., 2026), and PriorRG (Liu et al., 2025) provide literature context (†), not reproduced controls; reference preprocessing and scoring are unmatched.

All reported scores use test data; validation supports calibration and selection, including additional seed-42 ablations (Appendix C.11).

We report seed means and sample SD. Original patient statistics use 2,000 bootstrap draws and cluster permutations with Holm correction; post-hoc PRF/state/edit intervals use separate unadjusted bootstraps. These intervals do not transfer to the new generation controls or 4B runs (Appendix C.6).

Matched edits. The frozen Full, continuous, and prefix-only checkpoints share 3,696 requests from 1,596 cases and 245 patients. Eligibility requires all three diagnostic states to agree on a positive/negative endpoint; a fixed seeded ordering caps each concept at 300, before edit scoring. We flip the target ID or set its continuous probability vector to the requested one-hot state, retain each model's other conditioning, and recompute changes. Directional success requires the report to move from the original to requested endpoint; joint success additionally keeps all thirteen non-target labels unchanged. The primary comparison is Full minus continuous concept-macro joint success; Full minus prefix-only is secondary. Paired patient bootstrap uses 2,000 shared draws without p-values (Appendix C.4). Original-model edits remain separate (Appendix C.10).

5 RESULTS AND ANALYSIS

5.1 CLINICAL-LABEL AGREEMENT AND COMPLEMENTARY REPORT METRICS

FCSL reaches micro-F1 0.637 ± 0.002 on RG4 and 0.543 ± 0.004 on ABN (Table 1). Gains over D+section are 9.48 and 6.78 points, with paired patient-bootstrap 95% intervals [8.46, 10.52] and [4.86, 8.86]. Both pass Holm correction ($p = 0.006$), conditional on the trained models and exploratory history.

RG4 gains primarily accompany recall (0.666 vs. 0.491), with little precision change (0.610 vs. 0.605; difference CI $[-0.48, 1.59]$ points). ABN improves both. Concept-level results are mixed on ABN (Appendix C.7); its macro-F1 and RadGraph differences do not pass Holm correction. Lower language and RadGraph scores preclude overall report-quality superiority.

Table 1: **Clinical-label gains with lower language and RadGraph scores.** Three-seed test means; ±: seed SD. P/R: CheXbert micro-precision/recall; F1: CheXbert. B4: BLEU-4; RL: ROUGE-L; RG: RadGraph-ERC. Bold: better within-task system score, including rounded ties. †: unmatched literature; –: not quoted.

Task	Method	Micro-F1	P	R	Macro-F1	B4	RL	RG
RG4-SN	LLM-RG4†	0.609	–	–	–	0.196	0.384	–
RG4-SN	AnchorDiff†	0.609	–	–	–	0.203	0.379	–
RG4-SN	D+section	0.561 ± 0.0027	**0.623**	0.510	0.392	**0.189**	**0.384**	**0.262**
RG4-SN	FCSL	$\mathbf{0.649 \pm 0.0009}$	0.620	**0.682**	**0.471**	0.172	0.365	0.235
RG4-SW	LLM-RG4†	0.610	–	–	–	0.186	0.382	–
RG4-SW	AnchorDiff†	0.607	–	–	–	0.189	0.358	–
RG4-SW	D+section	0.567 ± 0.0017	0.618	0.523	0.413	**0.180**	**0.375**	**0.255**
RG4-SW	FCSL	$\mathbf{0.653 \pm 0.0031}$	**0.620**	**0.689**	**0.496**	0.168	0.360	0.232
RG4-MN	LLM-RG4†	0.559	–	–	–	0.216	0.405	–
RG4-MN	AnchorDiff†	0.561	–	–	–	0.224	0.402	–
RG4-MN	D+section	0.472 ± 0.0097	0.560	0.408	0.367	**0.208**	**0.403**	**0.278**
RG4-MN	FCSL	$\mathbf{0.596 \pm 0.0029}$	**0.587**	**0.604**	**0.462**	0.196	0.387	0.265
RG4-MW	LLM-RG4†	0.563	–	–	–	0.197	0.401	–
RG4-MW	AnchorDiff†	0.550	–	–	–	0.197	0.401	–
RG4-MW	D+section	0.482 ± 0.0063	0.550	0.428	0.390	**0.190**	**0.393**	**0.267**
RG4-MW	FCSL	$\mathbf{0.595 \pm 0.0074}$	**0.575**	**0.615**	**0.475**	0.183	0.380	0.257
RG4-all	D+section	0.542 ± 0.0008	0.605	0.491	0.406	**0.189**	**0.385**	**0.263**
RG4-all	FCSL	$\mathbf{0.637 \pm 0.0019}$	**0.610**	**0.666**	**0.490**	0.176	0.369	0.242
ABN	PriorRG†	0.471	–	–	–	0.102	0.242	–
ABN	D+section	0.476 ± 0.0029	0.460	0.493	0.327	**0.056**	**0.175**	**0.088**
ABN	FCSL	$\mathbf{0.543 \pm 0.0037}$	**0.534**	**0.553**	**0.366**	0.049	0.165	0.078

Table 2: **Supervision-matched and retrained route controls.** RG4 test ($N = 6{,}215$), seed 42; validation-selected. Full is the same-batch reference, not the original seed-42 run. Bold: highest displayed score, not significance or test-based selection. Train.: trainable parameters (M). Complete validation-selected metrics: Appendix C.2.

Interface	Train.	Micro-F1	Macro-F1	B4	RG
Full	79.75	0.6340	0.4901	0.1772	0.2396
Continuous	79.75	0.6331	**0.5130**	**0.1961**	**0.2596**
Prefix-only	58.47	0.6383	0.4984	0.1776	0.2446
Write-only	79.48	**0.6397**	0.4956	0.1752	0.2444

5.2 Finite conditioning and its empirical trade-offs

The supervised continuous interface has micro-F1 0.6331 versus Full's 0.6340, with higher macro-F1, BLEU-4, and RadGraph at the same trainable count (Table 2). The original F/D+section gains therefore cannot be assigned to discretization. Hard/soft states, residual quantization, and associated regularizers change jointly, not in isolation.

Prefix-only exceeds Full on all seven generation-score point estimates with 58.47M trainable parameters; write-only improves micro-F1 and RadGraph but lowers BLEU-4. These results do not establish a dual-route generation-quality benefit. Matched edits below assess a different capability.

Full's complete 202-bit message specifies all conditioning, whereas the continuous control's diagnostic labels do not specify its soft states and residual vectors. Reconstruction is a design property; comparative editing separately assesses selective response under a specified protocol.

Residual capacity. Removing residuals reduces the budget from 202 to 74 bits and BLEU-4 from 0.177 to 0.131, while micro-F1 changes from 0.639 to 0.636. This does not establish clinically

(a) Report generation: closer to the reference



GT reference	D+section	FCSL
there is enlargement of the cardiac silhouette with **pulmonary vascular congestion** and **bilateral pleural effusions** with **compressive atelectasis**. central catheter remains in place.	[…] pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis at the bases. in the appropriate clinical setting **supervening pneumonia would have to be considered**. […]	there is enlargement of the cardiac silhouette with **pulmonary vascular congestion** and **bilateral pleural effusions** with **compressive atelectasis** at the bases. central catheter remains in place.

FCSL retains the reference findings without the additional pneumonia hypothesis in D+section.

(b) State editing: successful target response



GT (original case)

[…] there is no pneumothorax. […]

Requested state: pneumothorax, negative to positive

Model	Before edit	After edit	Label outcome
FCSL (Full)	[…] no **pneumothorax** is identified. […]	[…] there is a small right apical **pneumothorax**. […]	**Target met** **0 other changes**
Continuous	[…] there is no focal consolidation or **pneumothorax**. […]	[…] there is no **pneumothorax**. […]	**Target not met** **0 other changes**
Prefix-only	[…] no **pneumothorax** is identified. […]	[…] no **pneumothorax** is identified. […]	**Target not met** **0 other changes**

Only FCSL meets the requested state while retaining its edema, effusion and atelectasis descriptions.

Figure 3: **Report generation and targeted state editing.** (a) Reference-supported findings. (b) Only Full realizes the requested pneumothorax reversal; all three models preserve non-target binary labels. These are post hoc examples from separate comparisons. GT describes the original case, not an edited target. Full texts: Appendix E.

Table 3: **Matched state edits.** RG4 seed 42, shared 3,696 requests; fourteen-concept macro rates (%). Align: initial alignment; Direction/Joint include unaligned requests. Any/Mean off: non-target label changes. Bold: best value, not significance. Intervals and coverage: Appendix C.4.

Interface	Align	Direction	Joint	Any off	Mean off
Full	**93.09**	**54.22**	**33.05**	32.72	3.01
Continuous	89.82	43.75	18.60	46.55	5.33
Prefix-only	92.63	47.54	29.63	**31.37**	**2.90**

independent residuals. Style+length and combined-objective extensions do not yield consistent gains (Appendix C.12); the main recipe is unchanged.

5.3 Selective response under matched state edits

Full achieves 33.05% joint selective success versus continuous's 18.60% (Table 3). The primary paired difference is +14.44 points (95% CI [11.87, 16.63]); Full also has fewer requests with any non-target change (32.72% vs. 46.55%). Excluding no finding as an edit target or ignoring its non-target changes preserves the direction of the joint-success difference (Appendix C.5).

Relative to prefix-only, Full gains 3.41 joint-success points (secondary CI [1.21, 5.33]), despite its lower generation scores. It does not show fewer non-target changes: the any-change difference is +1.35 points [−1.35, 4.04]. Thus the dual-route comparison supports a narrower response benefit, not uniformly improved control. Joint success still occurs for only about one third of requests. These outcomes concern specified state edits on the common eligible subset, not image-validated corrections or clinical safety. Original-model RG4/ABN edits and state-correctness diagnostics remain in Appendices C.10 and C.9.

5.4 Matched qualitative comparisons

In Figure 3(a), FCSL retains the reference findings without D+section's additional pneumonia hypothesis. In (b), Full expresses the requested pneumothorax state while retaining its edema,

Table 4: **Compact-backbone parameter accounting.** RG4 Stage B. LM: nominal size; system: registered parameters. Bold: minima among our systems. †: literature reference, not a matched-quality comparison.

Method	LM (B)	LM / 7B	System (M)	Trainable (M)	Trainable (%)
LLM-RG4†	7.0	100%	–	–	–
D+section	**0.8**	**11.4%**	**963.869**	**44.246**	**4.59**
FCSL	**0.8**	**11.4%**	999.506	79.748	7.98

effusion, and atelectasis descriptions; both controls continue to negate pneumothorax. These examples illustrate reference agreement and intervention response, not clinical correctness or population-level superiority.

5.5 Compact backbone and parameter accounting

The 0.8B backbone is approximately one ninth of LLM-RG4's nominal 7B decoder; the full system has 999.51M parameters, with 79.75M trainable. FCSL adds 35.50M trainable parameters over D+section and matches continuous conditioning. This is neither a discretization-induced saving nor an equal-quality 7B comparison.

The 4B FCSL reaches micro-F1 0.6420 versus D+section's 0.5696, with lower BLEU-4/RadGraph. This single-seed extension shows no uniform gain over original 0.8B FCSL, nor establishes equivalence (Appendix C.3). Latency, memory, and training cost remain unmeasured.

5.6 What the interface makes observable

Named states separate observed encoding and realization errors. Among reference-positive RG4 decisions, 4.14% have a correct state but an incorrect report label; 29.28% are missed by both (Appendix C.9). Improving state following alone therefore cannot address every omission. These concept-level fractions localize disagreement, not its cause.

The tuple also specifies an intervention: edit one state, hold other independent IDs fixed, recompute derived changes, and update both routes. This separates the message change from the measured response. Higher generation scores need not imply higher selective success; the interface makes that distinction directly testable.

6 Limitations

Development involved prior test-set feedback, so the study is exploratory. Seeds share Stage-A weights, and both tasks derive from MIMIC-CXR. Matched controls, edits, and the 4B extension cover only RG4 seed 42. Coupled representation/regularization changes and unequal route capacities limit attribution. Residuals may encode clinical information without prescribed semantics; gold-to-predicted conditioning, shared supervision/scoring, and absent laterality fields further limit interpretation. Most matched edits fail joint success, and unchanged labels ensure neither fine-grained text stability nor clinical correctness. The requested edits are not adjudicated corrections; external-site and expert validation remain necessary.

7 Conclusion

FCSL combines complete finite conditioning with a 0.8B backbone. Matched controls reveal a trade-off: continuous conditioning yields higher complementary report scores, while Full achieves higher joint selective response under shared-subset edits. Reconstructibility, generation quality, and clinical safety remain distinct properties.

AI USE STATEMENT

AI tools assisted with manuscript refinement. The authors reviewed all content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the interface, training recipe, and validation-based selection. Appendices A and B detail reconstruction, calibration, decoding, and scoring. Appendices C.1–C.3 describe the interface, access-route, and backbone-scale controls; Appendix C.4 specifies the shared-request intervention protocol and patient-bootstrap analysis. Parameter accounting and qualitative selection are documented in Appendices D and E. The appendices provide experimental settings and aggregate results. An accompanying partial code supplement demonstrates dataset loading and input organization with synthetic examples; it does not include model training, generation, checkpoints, or full result reproduction. Patient-level artifacts remain subject to data-access restrictions. Scorer metadata were reconstructed retrospectively; complete original environment locks are unavailable.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 7319–7328, 2021. doi: 10.18653/v1/2021.acl-long.568. URL https://aclanthology.org/2021.acl-long.568/.

Hasan Md Tusfiqur Alam, Devansh Srivastav, Abdulrahman Mohamed Selim, Md Abdul Kadir, Md Moktadirul Hoque Shuvo, and Daniel Sonntag. CBM-RAG: Demonstrating enhanced interpretability in radiology report generation with multi-agent RAG and concept bottleneck models. In *Companion Proceedings of the 17th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, pp. 59–61. Association for Computing Machinery, 2025. doi: 10.1145/3731406.3731970. URL https://www.dfki.de/en/web/research/projects-and-publications/publication/15937.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005. URL https://aclanthology.org/W05-0909/.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024. URL https://arxiv.org/abs/2406.04449.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *arXiv preprint arXiv:1506.03099*, 2015. URL https://arxiv.org/abs/1506.03099.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision – ECCV 2022*, pp. 1–21. Springer, 2022. doi: 10.1007/978-3-031-20059-5_1. URL https://arxiv.org/abs/2204.09817.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449, 2020. doi: 10.18653/v1/2020.emnlp-main.112. URL https://aclanthology.org/2020.emnlp-main.112/.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5904–5914, 2021. doi: 10.18653/v1/2021.acl-long.459. URL https://aclanthology.org/2021.acl-long.459/.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/867c06823281e506e8059f5c13a57f75-Abstract-Conference.html.

Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/770cabd044c4eacb6dc5924d9a686dce-Abstract-Conference.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021. URL https://arxiv.org/abs/2106.14463.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11523–11532, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Lee_Autoregressive_Image_Generation_Using_Residual_Quantization_CVPR_2022_paper.html.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004. URL https://aclanthology.org/W04-1013/.

Kang Liu, Zhuoqi Ma, Zikang Fang, Yunan Li, Kun Xie, and Qiguang Miao. PriorRG: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation. *arXiv preprint arXiv:2508.05353*, 2025. URL https://arxiv.org/abs/2508.05353.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2824–2832, 2021. doi: 10.18653/v

1/2021.findings-emnlp.241. URL https://aclanthology.org/2021.findings-emnlp.241/.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390, 2024. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. RAD-DINO: Exploring scalable medical image encoders beyond text supervision. *arXiv preprint arXiv:2401.10815*, 2024. URL https://arxiv.org/abs/2401.10815.

Qwen Team. Qwen3.5: Towards native multimodal agents. Model card and technical blog, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1500–1519, 2020. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/.

Sana Tonekaboni, Viktoria Schuster, and Caroline Uhler. MultiLoReFT: Decoupling shared and modality-specific subspaces in multimodal learning via low-rank representation fine-tuning. *arXiv preprint arXiv:2607.16789*, 2026. URL https://arxiv.org/abs/2607.16789.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.html.

Zhuhao Wang, Yihua Sun, Zihan Li, Xuan Yang, Fang Chen, and Hongen Liao. LLM-RG4: Flexible and factual radiology report generation across diverse input contexts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8250–8258, 2025. doi: 10.1609/aaai.v39i8.32890. URL https://ojs.aaai.org/index.php/AAAI/article/view/32890.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63908–63962, 2024. doi: 10.52202/079017-2041. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/75008a0fba53bf13b0bb3b7bff986e0e-Abstract-Conference.html.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4(9):100802, 2023. doi: 10.1016/j.patter.2023.100802. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10499844/.

Shiying Yu, Jielei Wang, and Guoming Lu. AnchorDiff: Topology-aware masked diffusion with confidence-based rewriting for radiology report generation. *arXiv preprint arXiv:2605.17071*, 2026. URL https://arxiv.org/abs/2605.17071.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. *Nature Communications*, 16:3108, 2025. doi: 10.1038/s41467-025-58344-x. URL https://www.nature.com/articles/s41467-025-58344-x.

A Program Representation and Reconstruction

This appendix gives the reconstruction and capacity details supporting Section 3. Appendix B documents training and data access; Appendix C presents supplementary results.

A.1 Reconstruction and local-rank properties

Fix the program layout, shared embeddings, codebooks, slot identities, normalization, decoder parameters, and calibration policy. Each legal ID tuple selects a unique scenario/section embedding and concept-state prototype; each residual atom is the sum of its selected codewords. Derived changes are deterministic functions of current and prior states. Concatenation and normalization therefore define a unique $g_\theta(\iota)$. Both decoder routes reconstruct from this same tuple.

For a fixed instruction and generated prefix, equal IDs give the same mathematical next-token distribution. Stochastic sampling can still produce different reports, and hardware, batching, or floating-point computation can affect exact numerical reproducibility. We consequently state distributional reconstruction, not an unconditional claim that every report is a deterministic function of IDs.

For token t at write site l, using column-vector notation, $\Delta h_{l,t} = \beta_{l,t} U_l d_{l,t}$, where $\beta_{l,t}$ contains the scalar gate and mask. Thus $\Delta h_{l,t} \in \mathrm{col}(U_l)$ and the direct-write dimension is at most 32. The basis differs across layers; later nonlinear propagation and the prefix route are not restricted to any one such subspace. Alphabet size, write rank, and shared LoRA parameters describe distinct properties.

Editing a current-state ID reconstructs the dependent change atom and both access routes. A generated report can nevertheless change other findings or sentence structure through attention and autoregressive decoding. Per-concept target and off-target label changes must be measured directly before claiming localized control.

The finite message specifies the full case-dependent conditioning representation, not merely a set of diagnostic labels alongside an unrecorded feature vector. Its exact reconstruction assumes fixed shared weights and codebooks; IDs need not retain their meaning across model versions. Continuous tensors can also be serialized and replayed, and finite-precision floats have finite storage. The distinction here is the explicit compact categorical layout, not an assertion that practical continuous representations have infinite capacity. Quantized residuals may still convey clinical information absent from, or redundant with, named states. The message budget is not a privacy guarantee or a measured reduction in decoder memory or compute.

A.2 Alphabet accounting

Table 5: RG4 uses a 202-bit fixed-width layout in every scenario. Derived changes increase the atom count, not the independent ID budget.

Field	IDs	Choices per ID	Fixed bits	Atoms
Scenario	1	4	2	1
Section	1	3	2	1
Current observations	14	4	28	14
Prior observations	14	5	42	14
Derived changes	0	–	0	14
Residual codes	16	256	128	8
RG4 total	46	–	202	52
ABN (without prior/change)	32	–	160	24

For categorical fields of cardinality n_i, independent fixed-width coding uses $\sum_i \lceil \log_2 n_i \rceil$ bits, whereas the joint log-cardinality upper bound is $\sum_i \log_2 n_i$. These are different for the three-way section and five-way prior fields. RG4 has joint bound $2 + \log_2 3 + 28 + 14 \log_2 5 + 128 = 192.092$ bits, versus the declared 202-bit layout; ABN has 159.585 versus 160 bits. Scenario-dependent constraints can further reduce attainable combinations.

In SN/MN, fourteen prior IDs are fixed to symbol 4 (unavailable), and the corresponding change atoms are unavailable. The implemented layout remains 52 atoms and its declared serialized budget remains 202 bits. Omitting known constants could define a conditional coding scheme, but is not the fixed-layout budget reported here. ABN disables these blocks entirely. Section is a task instruction, not a target-derived label, and is included in accounting even when fixed within a task.

The four change categories are unavailable/unmentioned, new positive endpoint, resolved positive endpoint, and stable endpoint. Positive and uncertain states share the endpoint. Blank current/prior states or unavailable prior states map to unavailable/unmentioned change. These derived categories are not an independently annotated temporal-disease ontology.

B Training and Evaluation Details

B.1 Query encoder

Four query blocks operate on fourteen clinical and eight residual queries of width 512, using eight-head self-attention across queries, cross-attention to evidence memory M, and a feed-forward transformation:

$$\begin{aligned} Q^a &= Q + \mathrm{SA}(\mathrm{LN}(Q)), \\ Q^b &= Q^a + \mathrm{CA}(\mathrm{LN}(Q^a), M, M), \\ Q^+ &= Q^b + \mathrm{FFN}(\mathrm{LN}(Q^b)). \end{aligned} \tag{12}$$

B.2 Clinical supervision and calibration

The two-epoch Stage-A pretraining excludes the language model and uses four-state focal supervision:

$$\mathcal{L}_{\mathrm{state}} = -\frac{1}{|\Omega|} \sum_{(i,c)\in\Omega} w_{c,s_{ic}} (1 - p_{ic,s_{ic}})^2 \log p_{ic,s_{ic}}, \tag{13}$$

where Ω contains observed training labels. State supervision uses training-only inverse-square-root class-frequency weights, capped at eight. The four explicitly reweighted rare concepts are enlarged cardiomediastinum, lung lesion, other pleural abnormality, and fracture. The inherited Stage-A checkpoint was trained with rare positive/uncertain weight eight; Stage B uses four. Missing state labels are excluded from supervision. Observed gold labels condition Stage B; only missing labels fall back to the predicted soft mixture at temperature 0.5.

Threshold candidates are $\{.05, .06, \ldots, .95\}$. Each concept first maximizes its own endpoint F1. Three sequential coordinate-ascent passes then optimize pooled micro-F1 by holding other concepts' confusion counts fixed. Ties prefer the threshold closest to .5 and then the larger threshold; concepts without validation positives retain .5. Within an endpoint, ties favor positive over uncertain and blank over negative.

Validation patients are deterministically assigned to two folds. Thresholds trained on one fold predict the other, yielding out-of-fold header diagnostics. A separate full-validation fit is stored with the checkpoint for inference. Seed 17/73 calibration reuses the inherited Stage-A probability cache and weights; changing the fold seed is not independent encoder pretraining. During Stage B, recalibration repeats each epoch before epoch/resume checkpoints and generated validation reports. The teacher-forced LM diagnostic precedes this refit.

Cross-fitting removes overlap for the current threshold-fitting diagnostic, but prior validation selection and repeated threshold feedback remain. It does not establish calibrated probabilities or an unbiased performance estimate for an independently selected model.

The gold-to-predicted state mismatch differs from the previous-token exposure bias addressed by scheduled sampling (Bengio et al., 2015). Here, reference-derived states condition report training, whereas predicted states condition inference. An explicit interface makes this mismatch observable but does not resolve it.

B.3 Optimization and decoding settings

The 0.8B Stage B runs three RG4 or five ABN epochs at global batch 128. Learning rates are 2×10^{-4} for interface/LoRA, $5 \times 10^{-5}/10^{-5}$ for the RG4/ABN clinical encoder, and 5×10^{-6} for visual parameters. AdamW uses cosine decay, 3% warmup, weight decay .01 with parameter-type exclusions, and gradient clipping at 1.0. The matched RG4 interface comparison uses these learning rates and three-epoch budget.

The clinical encoder has width 512, four query blocks, and eight attention heads; reconstructed atoms have width 256. Eight residual roles each use two 256-way codebooks. Codebook EMA updates occur once per optimizer update, with data-based initialization and dead-code restarts. The usage penalty is the normalized positive gap below an entropy floor of $0.6 \times 128 = 76.8$ bits, weighted by 0.5. It is not a penalty requiring full use of the maximum alphabet. Learned entropy priors remain frozen.

LoRA uses rank 16, alpha 32, dropout .05, and all decoder-block linear modules. Trainable master parameters are FP32 with BF16 autocast on CUDA. The write bases have 32 active columns, an initial gate of .5, and orthogonality regularization on their raw matrices. Sequence length during chest report training is capped at 384 tokens, in addition to the program prefix. Training shuffles each row once per epoch, with minimal distributed-sampler padding.

The final decoding grid uses three beams and repetition penalties $\{1, 1.2, 1.5, 2\}$ crossed with length penalties $\{1, 2\}$, zero minimum new tokens, plus the $(2, 2, 50)$ tuple. Each tuple is evaluated with repetition penalties applied either to the whole sequence or only generated tokens, giving 18 candidates. Selection maximizes validation composite subject to micro-F1 no more than .005 below the baseline and zero runaway rate, with grid-order tie breaking. The chosen dataset-specific policy is frozen for FCSL and D+section before test replay.

Table 6: Decoding settings for all twelve main runs. All runs use deterministic decoding, generation batch 8, and prompt cap 256 tokens.

Task	Beams	Max new	Min new	Repetition penalty	Length penalty
RG4	3	200	0	1.5	2.0
ABN	3	80	0	1.0	1.0

`penalize_generated_only=False`: the default repetition penalty uses the supplied token history, including prompt IDs, not only generated tokens. Continuous prefix embeddings are not vocabulary IDs. ABN's repetition penalty of 1.0 is inactive. Generation stops at the first generated `im_end` or `endoftext`. Only the continuation is decoded, special tokens are skipped, and surrounding whitespace is stripped; no clinical sentence-cleaning regex is applied.

Scorers and preprocessing. Language metrics are BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). Scoring uses the official LLM-RG4 evalcap implementation, METEOR, CheXbert, and RadGraph. Recorded package versions are torch 2.7.1+cu118, transformers 5.7.0, tokenizers 0.22.2, numpy 1.26.4, radgraph 0.1.18, and nltk 3.10.0. RadGraph uses `radgraph-xl`, `reward_level=all`, with the complete entity–relation score reported as ERC.

The NLG adapter passes reference and generated strings directly, without an extra PTBTokenizer pass. BLEU/CIDEr use whitespace tokenization; ROUGE-L uses literal-space tokenization with $\beta = 1.2$; METEOR uses the bundled English `-norm` implementation. CheXbert uses BERT-base-uncased, truncation at its configured positional limit, and the source's literal string replacements; backslash-s patterns are not regex whitespace normalization. Scorer metadata were reconstructed retrospectively; complete original environment locks and per-process Java versions are unavailable.

Clinical evaluation scope. RadGraph-based evaluation and RadCliQ examine alignment with radiologist error judgments (Yu et al., 2023), while GREEN identifies and explains clinically significant report errors (Ostmeier et al., 2024). They motivate distinguishing lexical similarity, label agreement, and clinical correctness. We report RadGraph-ERC, not RadCliQ or GREEN; no uncomputed metric or expert assessment is implied by citing these studies.

B.4 SECTION-CONDITIONED CONTINUOUS-PREFIX BASELINE

D+section and FCSL use the same available evidence, selected Stage-A visual weights, language backbone, LoRA configuration, report-training update budget, checkpoint-selection rule, and final decoding policy. D+section therefore benefits from the same visual pretraining; it does not inherit FCSL's clinical query encoder. The remaining differences are summarized in Table 7.

Table 7: Conditioning-design differences in the main comparison. Shared training and evaluation settings do not match supervision, representation capacity, or decoder access.

Component	D+section	FCSL
Evidence fusion	128 continuous queries	14 clinical + 8 residual queries
Stage-B state supervision	None	Four-state focal loss
Current-state conditioning in training	No explicit state interface	Reference-derived; soft fallback for missing labels
Requested section	Evidence-free prompt suffix	Discrete program field
Decoder input prefix	128 continuous tokens	52 RG4 / 24 ABN reconstructed atoms
Layer-local writes	None	Eight rank-32 writes

This comparison evaluates complete systems and does not identify quantization separately from clinical supervision, learned fusion, or local writes. The supervision-matched continuous and retrained single-route controls are specified in Appendix C.1 and evaluated in Table 2. They narrow these confounds but retain joint representation/regularization or capacity differences. Parameter counts are given in Appendix D.

B.5 DATA SOURCES AND ACCESS

MIMIC-CXR is a de-identified, credentialed-access dataset (Johnson et al., 2019). The motivation and framework illustrations adapt cropped radiographs and reference excerpts from Fig. 1 of the data article (CC BY 4.0); they illustrate the interface, not example-specific FCSL predictions, and the reference is not a model input. The matched cases in Figure 3 use author-authorized MIMIC-CXR-JPG images, RG4 references, and saved model outputs. Case aliases replace dataset identifiers, and the credentialed dataset is not redistributed.

C SUPPLEMENTARY RESULTS

C.1 MATCHED INTERFACES AND RETRAINED ACCESS ROUTES

The four-arm comparison uses RG4 seed 42, the same fixed Stage-A initialization, and three complete Stage-B epochs without predicted-state adaptation. Each arm receives 9,948 planned optimizer updates with global batch 128 (eight GPUs, microbatch eight, accumulation two), generation batch eight, and full validation after each epoch. Checkpoints use Equation (11), earliest-epoch tie breaking, and the frozen RG4 decoding policy in Table 6. The same-batch Full is a reference retraining, not a replacement for original seed 42 in Table 1. All arms first instantiate the canonical Full model and then remove inapplicable modules, preserving identical initialization for the shared trainable tensors.

Supervision-matched continuous interface. The continuous arm retains the clinical query encoder, fourteen four-state embedding sets, residual projection, task/prior inputs, prefix and write routes, rank-16 LoRA, and the state loss with weight .3. Like Full, training uses reference-derived states and inference does not. At inference, current-state embeddings are probability mixtures at temperature .5, change atoms are expectations under the same current/prior lookup, and residuals use pre-RVQ projected vectors. Known scenario, section, and legitimate prior fields retain the shared categorical definitions. Diagnostic calibrated hard labels do not reconstruct this continuous message; the decoder instead consumes unthresholded temperature-scaled probabilities and residual features.

Removing RVQ sets VQ and usage penalties to zero, versus .1 and .5 in Full; the state loss and write orthogonality penalty (.01) remain. Full and continuous each have 79,747,648 trainable parameters,

but their total registered counts differ because of removed frozen parameters. This comparison jointly changes hard/soft clinical states, residual quantization, and its regularizers. It is not an independent estimate of any one factor's effect.

Single-route retraining. Prefix-only removes write modules and hooks before optimization, with no write orthogonality penalty. Write-only removes the prefix projection before optimization. Both retain the finite program, clinical supervision, and relevant shared initialization. Their respective trainable counts are 58,473,016 and 79,484,480. Thus these are retrained access designs, not inference-time toggles or equal-capacity ablations; removing the prefix also changes decoder context length.

Scope of the interface guarantee. All finite variants use the same 202-bit RG4 layout. The generation path reconstructs the code from IDs and checks equality with the encoder-produced program. This implements the representation property in Appendix A.1; it does not prove selective editing. The continuous control also exposes clinical predictions and supports probability interventions. Appendix C.4 evaluates a separately implemented target-state intervention for the same frozen Full, continuous, and prefix-only models. Their original training code and checkpoint selection remain unchanged.

C.2 Complete metrics for validation-selected interface controls

Table 8: Complete metrics for the four 0.8B interface controls on RG4 test ($N = 6{,}215$), seed 42. E: validation-selected epoch; M: METEOR; C: CIDEr; other abbreviations follow Table 1. Bold marks the highest point estimate in each metric, not statistical significance.

Model	E	Micro-F1	Macro-F1	B4	M	RL	C	RG
Full	3	0.6340	0.4901	0.1772	0.1947	0.3691	0.3250	0.2396
Continuous	2	0.6331	**0.5130**	**0.1961**	**0.2091**	**0.3807**	**0.3511**	**0.2596**
Prefix-only	2	0.6383	0.4984	0.1776	0.1953	0.3708	0.3338	0.2446
Write-only	2	**0.6397**	0.4956	0.1752	0.1938	0.3694	0.3381	0.2444

The continuous arm exceeds same-batch Full in macro-F1, BLEU-4, METEOR, ROUGE-L, CIDEr, and RadGraph, with a micro-F1 point estimate lower by 0.0009. Prefix-only exceeds Full on all seven displayed scores. A single report-training seed supplies no estimate of seed variability or test of equivalence; the original main-result patient intervals do not apply to these models.

C.3 Backbone-scale extension

Table 9: Validation-selected 4B system comparison on RG4 test ($N = 6{,}215$), seed 42. Train.: trainable parameters (M). Bold marks the higher displayed score within this pair, not statistical significance. D+section lacks FCSL's clinical encoder/state supervision; it is not the matched continuous control.

Model	Train.	Micro-F1	Macro-F1	B4	M	RL	C	RG
FCSL (4B)	196.55	**0.6420**	**0.4991**	0.1776	0.1957	0.3682	0.3304	0.2406
D+section (4B)	66.68	0.5696	0.4357	**0.1916**	**0.2019**	**0.3787**	**0.3770**	**0.2581**

The 4B batch uses the same RG4 splits, inherited Stage-A seed-42 weights, three Stage-B epochs, global batch 128, and validation/decoding rules. Microbatch four and accumulation four preserve the global batch on eight GPUs; generation batch is four. FCSL uses eight rank-32 writes at zero-indexed layers 3, 7, 11, 15, 19, 23, 27, and 31, with rank-16 LoRA. D+section restores the shared visual initialization and uses its continuous 128-query prefix. No narrative extension, residual adversary, calibration variant, or state-weighted token objective is added.

The FCSL–D+section micro-F1 difference is 0.0723, accompanied by lower BLEU-4 (0.1776 versus 0.1916) and RadGraph (0.2406 versus 0.2581). Original 0.8B FCSL seed 42 has micro-F1 0.6390, BLEU-4 0.1766, and RadGraph 0.2421; the respective 4B values are 0.6420, 0.1776, and 0.2406. This is no uniform scaling gain. Different batches and generation batch sizes, one seed, and unmatched

F/D supervision prevent an equivalence claim, scaling law, or isolated discretization effect. The 4B and matched-interface comparisons use separate training runs from the three-seed main experiment.

C.4 SHARED-REQUEST EDITING PROTOCOL

This evaluation reuses the RG4 seed-42 validation-selected Full, continuous, and prefix-only models from Table 2: epochs 3, 2, and 2, respectively. No weights, calibration thresholds, decoding settings, or selected checkpoints are changed. Write-only, ABN, 4B, and additional seeds are outside this evaluation. The original main-model edit results in Appendix C.10 use different checkpoints and request sets and are not pooled here.

Eligibility and coverage. Each model first extracts its native conditioning for all 6,215 test cases. A case–concept pair is eligible only when all three calibrated diagnostic states agree and are explicitly positive or negative. These diagnostic IDs determine eligibility but do not replace the continuous model's probability-weighted conditioning. Within each concept, a fixed seeded ordering of sample identities selects at most 300 cases, with dataset order breaking ties. The ordering is fixed before edit scoring and does not depend on generated reports. The target reverses the common endpoint. Neither baseline report alignment nor edit success determines eligibility. The frozen list contains 3,696 requests, 1,596 unique cases, and 245 of the test set's 287 patients. There are 2,149 positive-to-negative and 1,547 negative-to-positive requests, without direction balancing.

Table 10: Eligibility on 6,215 RG4 test cases per concept. F/C/P: individually positive/negative diagnostic states in Full/continuous/prefix-only; Any: at least one arm eligible; Binary: all three arms binary; Same: all three agree on the same binary endpoint. Counts are case–concept pairs, not distinct patients.

Concept	F	C	P	Any	Binary	Same
Enl. cardiomediastinum	1718	1751	1734	1892	1581	1573
Cardiomegaly	4109	4173	4069	4354	3884	3828
Lung opacity	3208	3325	3457	3602	3046	3028
Lung lesion	139	292	323	347	126	126
Edema	2431	2464	2532	2754	2192	2154
Consolidation	1832	1624	1595	1967	1448	1448
Pneumonia	1641	1524	1514	1765	1369	1361
Atelectasis	2412	2258	2537	2690	2107	2103
Pneumothorax	3500	3701	3740	3849	3429	3401
Pleural effusion	4746	4825	4796	4980	4572	4488
Pleural other	98	138	100	140	85	85
Fracture	225	208	201	246	185	185
Support devices	2681	2299	2418	2687	2265	2265
No finding	315	409	359	418	306	306

Eligibility varies markedly by concept; the common subset is not representative of all test case–concept pairs. Pleural other, support devices, and no finding have no selected negative-to-positive requests. Atelectasis has only one such request. Empty concept–direction cells remain undefined in the aggregate data, not zero, and sparse cells do not establish reliable directional control.

Interventions. For Full and prefix-only, the target current-state ID is replaced and dependent change atoms are recomputed; all other IDs, including each model's own residuals, remain fixed. For continuous conditioning, only the target concept's temperature-.5 probability vector is set to the requested one-hot state. Other concept probabilities and that model's unquantized residual projections remain unchanged; expected-change atoms and normalization are recomputed. The continuous code is not globally hardened, and residuals are not exchanged between models. The experiment tests these specified operators, not all possible continuous interventions or an isolated quantization effect.

Paired generation. Intact and edited reports are generated in matching request batches with the same frozen policy (three beams, generation batch eight per GPU, 200-token limit). Implementation checks verify conditioning reconstruction, no-op generation, and unchanged non-target conditioning

Table 11: Frozen request selection and direction coverage. Same/test: percentage of all 6,215 cases eligible before the per-concept cap; N: selected requests. Directions denote requested state reversals, not verified diagnostic corrections.

Concept	Same/test (%)	N	Positive → negative	Negative → positive
Enl. cardiomediastinum	25.31	300	16	284
Cardiomegaly	61.59	300	220	80
Lung opacity	48.72	300	286	14
Lung lesion	2.03	126	104	22
Edema	34.66	300	168	132
Consolidation	23.30	300	27	273
Pneumonia	21.90	300	112	188
Atelectasis	33.84	300	299	1
Pneumothorax	54.72	300	16	284
Pleural effusion	72.21	300	132	168
Pleural other	1.37	85	85	0
Fracture	2.98	185	84	101
Support devices	36.44	300	300	0
No finding	4.92	300	300	0

slots. Each arm yields 3,696 intact/edited pairs with no empty outputs. These checks do not establish clinical validity or byte-exact generation across environments.

Estimands and uncertainty. CheXbert positive/uncertain labels form one report endpoint, with blank/negative forming the other. For target concept c, original/requested endpoints u_c, v_c, and intact/edited report labels b, e, directional success is $\mathbb{1}[b_c = u_c, e_c = v_c]$. Joint success additionally requires $e_k = b_k$ for all $k \neq c$. Both use all selected requests, including initially unaligned outputs. Target attainment $\mathbb{1}[e_c = v_c]$ is reported separately. The primary estimand is Full minus continuous joint success, averaged equally over the fixed fourteen concepts; Full minus prefix-only is secondary. Request-pooled and initially aligned-only rates have different weights or denominators.

Two thousand patient-cluster bootstrap draws use identical patient weights for all models and requests. Percentile intervals for differences are computed from paired draw-wise differences, not by subtracting interval endpoints. An interval requires at least 95% valid draws; all primary and secondary joint-difference draws are valid. No p-values are reported; secondary and sensitivity intervals are descriptive and not multiplicity-adjusted. Intervals condition on these models and the fixed request-selection rule. They do not quantify training-seed variability, remove historical test feedback, or generalize beyond the common eligible subset.

C.5 Matched editing results and sensitivity analyses

Tables 12 and 13 distinguish equal-concept and request-pooled rates. In both, Full has higher joint success than continuous conditioning, but most requests still fail the criterion. Full's initial alignment is also higher. Restricting to initially aligned requests retains the ordering (Table 14), but the resulting model-specific subsets are not the primary paired estimand. Bold in comparative tables marks the best point estimate (lower for non-target changes, higher otherwise), not significance. These labels do not measure unchanged laterality, severity, temporal detail, or clinical safety.

Table 12: Fixed fourteen-concept macro edit rates (%) and 95% patient-bootstrap intervals on the common request set. All-request denominators apply unless stated otherwise. NF: no finding.

Metric	Full	Continuous	Prefix-only
Initial alignment	**93.09** [91.65, 94.33]	89.82 [87.90, 91.15]	92.63 [90.83, 93.89]
Directional success	**54.22** [51.64, 56.34]	43.75 [41.32, 45.83]	47.54 [45.28, 49.98]
Joint selective success	**33.05** [30.64, 35.33]	18.60 [17.25, 20.32]	29.63 [27.37, 31.87]
Any non-target change	32.72 [30.36, 35.04]	46.55 [44.06, 49.00]	**31.37** [29.11, 33.61]
Mean non-target change	3.01 [2.77, 3.25]	5.33 [4.93, 5.84]	**2.90** [2.66, 3.14]
Joint, ignoring non-target NF	**35.78** [33.52, 37.95]	20.32 [18.90, 22.08]	32.71 [30.64, 34.97]
Target endpoint attained	**60.51** [58.02, 62.75]	52.21 [49.74, 54.77]	54.32 [52.04, 57.06]

Table 13: Request-pooled edit rates (%) with 95% patient-bootstrap intervals. Each of the 3,696 requests has equal weight, unlike Table 12.

Metric	Full	Continuous	Prefix-only
Initial alignment	94.29 [93.22, 95.34]	90.80 [89.55, 91.97]	**94.56** [93.39, 95.64]
Directional success	**52.79** [49.64, 55.82]	42.45 [39.41, 45.41]	47.67 [44.15, 51.02]
Joint selective success	**33.06** [30.39, 35.77]	17.99 [16.43, 19.70]	30.25 [27.45, 32.84]
Any non-target change	30.41 [27.80, 33.24]	45.54 [41.94, 48.89]	**29.71** [27.19, 32.31]
Mean non-target change	2.79 [2.52, 3.08]	5.20 [4.69, 5.73]	**2.76** [2.48, 3.03]
Joint, ignoring non-target NF	**35.69** [33.21, 38.27]	19.72 [18.23, 21.45]	32.90 [30.24, 35.45]
Target endpoint attained	**57.90** [54.64, 61.03]	50.11 [46.87, 53.31]	52.49 [48.69, 56.21]

Table 14: Conditioning on initially aligned reports changes the denominator separately for each model. Rates (%) and 95% patient-bootstrap intervals are descriptive, not a replacement for all-request comparisons.

Model	Aligned requests	Conditional directional	Conditional joint
Full	3485	55.98 [52.57, 59.24]	35.06 [32.35, 37.79]
Continuous	3356	46.75 [43.39, 50.08]	19.82 [18.11, 21.67]
Prefix-only	3495	50.41 [46.58, 54.08]	31.99 [29.06, 34.73]

Table 15: Paired differences in percentage points [95% patient-bootstrap interval]. The first row is the primary Full–continuous comparison and secondary Full–prefix-only comparison. Remaining rows are descriptive or sensitivity analyses; no formal p-values or multiplicity-adjusted claims are made.

Estimand	Full − Continuous	Full − Prefix-only
Macro14 joint (primary / secondary)	+14.44 [11.87, 16.63]	+3.41 [1.21, 5.33]
Pooled joint	+15.07 [12.67, 17.36]	+2.81 [1.08, 4.56]
Macro14 directional	+10.47 [8.26, 12.78]	+6.68 [4.41, 8.50]
Macro14 any non-target change	-13.83 [-16.79, -10.94]	+1.35 [-1.35, 4.04]
Macro14 mean non-target change	-2.33 [-2.82, -1.91]	+0.11 [-0.16, 0.37]
Macro13 joint, excluding NF target	+14.96 [12.25, 17.27]	+3.09 [0.76, 5.24]
Macro14 joint, ignoring non-target NF	+15.46 [12.92, 17.63]	+3.07 [0.77, 4.85]

The primary joint-success difference is +14.44 points [11.87, 16.63]. It remains positive when no finding is excluded as a target (+14.96 [12.25, 17.27]) or its non-target changes are ignored (+15.46 [12.92, 17.63]). These sensitivity checks address an endpoint dependency, not medically adjudicated consistency. Against prefix-only, Full's smaller joint gain (+3.41 [1.21, 5.33]) accompanies an any-non-target-change difference of +1.35 [-1.35, 4.04]; it does not establish a reduction in such changes. Higher joint success can therefore coexist with no demonstrated improvement in that non-target metric.

Table 16: Joint selective success by concept (%) with patient-bootstrap intervals. Request and direction counts are in Table 11; small or direction-limited groups require caution.

Concept	Full	Continuous	Prefix-only
Enl. cardiomediastinum	25.67 [20.06, 31.69]	16.33 [12.35, 20.91]	**28.00** [21.68, 35.02]
Cardiomegaly	57.00 [51.57, 62.23]	33.67 [28.01, 39.34]	**57.33** [51.64, 62.93]
Lung opacity	**31.67** [26.26, 37.23]	18.33 [14.13, 23.29]	29.67 [24.16, 35.47]
Lung lesion	**29.37** [20.18, 39.26]	21.43 [13.46, 30.72]	27.78 [20.18, 35.30]
Edema	**60.00** [52.98, 67.24]	35.67 [29.71, 42.14]	56.67 [50.85, 62.15]
Consolidation	**17.00** [12.68, 21.37]	9.67 [6.78, 12.66]	16.33 [11.38, 21.32]
Pneumonia	**30.33** [23.87, 37.07]	10.67 [7.53, 14.29]	24.67 [18.94, 31.12]
Atelectasis	**28.00** [22.86, 32.98]	7.00 [3.95, 9.93]	23.67 [18.56, 28.13]
Pneumothorax	**34.33** [28.24, 41.13]	15.00 [11.05, 19.18]	23.00 [17.96, 28.71]
Pleural effusion	**51.00** [44.49, 57.09]	30.00 [24.50, 35.74]	**51.00** [45.49, 57.14]
Pleural other	**29.41** [11.32, 43.04]	18.82 [12.77, 30.00]	15.29 [5.26, 33.33]
Fracture	**44.86** [32.64, 57.01]	33.51 [24.19, 42.74]	39.46 [26.31, 52.25]
Support devices	9.67 [5.38, 14.61]	3.67 [1.48, 6.43]	**15.33** [10.49, 20.34]
No finding	**14.33** [8.92, 18.45]	6.67 [3.26, 10.00]	6.67 [3.86, 10.69]

Table 17: Direction- and scenario-stratified pooled joint success (%) with patient-bootstrap intervals. N: requests; Patients: unique patients within the stratum. Patients can occur in multiple strata; their counts must not be summed.

Stratum	N	Patients	Full	Continuous	Prefix-only
Positive → negative	2149	229	**33.92** [30.74, 37.10]	18.38 [16.43, 20.76]	31.69 [28.03, 35.25]
Negative → positive	1547	207	**31.87** [28.46, 35.50]	17.45 [15.27, 19.85]	28.25 [25.25, 31.59]
SN	1298	207	**36.83** [33.21, 40.46]	20.03 [17.36, 22.81]	32.67 [28.93, 36.37]
SW	1178	198	**29.46** [26.16, 33.03]	15.96 [13.46, 18.52]	29.12 [25.60, 32.79]
MN	693	146	**32.32** [28.62, 35.97]	19.19 [16.55, 22.04]	31.75 [27.96, 35.79]
MW	527	114	**32.83** [27.22, 38.45]	15.94 [12.83, 19.61]	24.86 [20.08, 30.17]

Interpretation. Full exceeds continuous conditioning in each concept's joint-success point estimate, but this is not fourteen independent significance claims. Prefix-only is competitive or better for several concepts. The data support a response advantage for the specified frozen models and operators on their common eligible subset, not reliable control for every finding or superiority of all finite interfaces. This batch adds no BLEU, RadGraph, or ordinary diagnostic-F1 scores, no runtime-efficiency benchmark, and no expert assessment. Per-request texts and labels are subject to the dataset's access restrictions; the supplementary tables report aggregate results.

C.6 Selected runs and patient-level statistics

Patient-cluster bootstrap resamples subjects, retaining all their reports and paired method predictions. Each of 2,000 draws is shared across the three fixed seeds; corpus metrics are recomputed within each seed and then averaged. Percentile intervals quantify patient-sampling uncertainty conditional on these models, not training-seed uncertainty. RG4 has 287 patients and 6,215 reports; ABN has 134 patients and 844 reports.

The two-sided randomization test swaps FCSL/D+section assignments at patient-cluster level, sharing each swap across seeds. We use 2,000 permutations, a plus-one correction, and Holm adjustment over twelve registered task–metric endpoints. Four METEOR/CIDEr endpoints reserve family positions with $p = 1$ for adjustment but are marked untested and have no reported p-values. Other endpoints use exact confusion-count, RadGraph-sum, or BLEU sufficient-statistic recomputation. This procedure does not remove earlier test-informed development.

Table 18: All twelve validation-selected test runs. Bold marks the better FCSL/D+section score within each task and seed, including displayed ties. Epoch identifies the selected Stage-B checkpoint. All runs use the fixed dataset-specific decoding policy.

Task	Method	Seed	Epoch	Micro-F1	Macro-F1	B4	RL	RG	CIDEr
RG4	FCSL	17	3	**0.636**	**0.499**	0.177	0.368	0.241	0.332
RG4	FCSL	42	2	**0.639**	**0.496**	0.177	0.369	0.242	0.336
RG4	FCSL	73	2	**0.635**	**0.475**	0.174	0.369	0.242	0.321
RG4	D+section	17	3	0.543	0.407	**0.190**	**0.386**	**0.264**	**0.400**
RG4	D+section	42	3	0.542	0.407	**0.190**	**0.386**	**0.264**	**0.390**
RG4	D+section	73	3	0.542	0.404	**0.187**	**0.384**	**0.261**	**0.382**
ABN	FCSL	17	5	**0.539**	**0.363**	0.049	0.164	0.080	0.116
ABN	FCSL	42	5	**0.546**	**0.370**	0.050	0.167	0.078	**0.126**
ABN	FCSL	73	5	**0.545**	**0.366**	0.047	0.162	0.074	0.106
ABN	D+section	17	3	0.473	0.326	**0.053**	**0.175**	**0.089**	**0.131**
ABN	D+section	42	5	0.474	0.332	**0.057**	**0.175**	**0.087**	**0.126**
ABN	D+section	73	4	0.479	0.321	**0.057**	**0.175**	**0.089**	**0.109**

Table 19: FCSL minus D+section differences in percentage points. Intervals are paired patient-bootstrap 95% percentile intervals, not seed SD or multiplicity-adjusted intervals; p is Holm-adjusted over twelve endpoints.

Task	Metric	Difference	95% CI	Adjusted p
RG4	Micro-F1	9.48	[8.46, 10.52]	0.0060
RG4	Macro-F1	8.37	[6.70, 9.89]	0.0060
RG4	BLEU-4	-1.29	[-1.59, -0.97]	0.0060
RG4	RadGraph	-2.12	[-2.54, -1.70]	0.0060
ABN	Micro-F1	6.78	[4.86, 8.86]	0.0060
ABN	Macro-F1	3.98	[1.42, 6.24]	0.0570
ABN	BLEU-4	-0.70	[-1.19, -0.27]	0.0175
ABN	RadGraph	-1.08	[-1.82, -0.32]	0.0725

C.7 Post-hoc clinical-label decomposition

This analysis uses the twelve validation-selected test outputs from the main comparison. Each corpus metric is first computed per trained model and then averaged equally across the three Stage-B seeds; pooling all predictions before computing F1 would be a different estimand. Positive and uncertain CheXbert labels map to the positive endpoint; blank and negative map to nonpositive. These results measure agreement with reference-report labels, not independently adjudicated diagnosis, four-state accuracy, or full report correctness.

Intervals use 2,000 paired patient-cluster bootstrap draws, shared across the fixed training seeds. They are pointwise, post-hoc intervals without multiplicity adjustment. These draws are separate from those used for Table 19, so interval endpoints may differ slightly even when point estimates agree.

Table 20: Post-hoc precision/recall decomposition. D/F are three-seed means; differences and intervals are percentage points. These intervals are not the original Holm-adjusted tests.

Task	Metric	D+section	FCSL	F−D	95% CI
RG4	Micro-precision	0.6050	0.6104	0.54	[-0.48, 1.59]
RG4	Micro-recall	0.4911	0.6658	17.47	[16.32, 18.65]
ABN	Micro-precision	0.4596	0.5343	7.47	[5.18, 9.97]
ABN	Micro-recall	0.4929	0.5533	6.04	[3.44, 8.72]

RG4's micro-F1 gain is primarily accompanied by higher recall; its precision-difference interval spans zero. ABN improves both aggregate precision and recall, but enlarged cardiomediastinum, other pleural abnormality, and fracture have lower F1. ABN has no reference-positive pneumothorax cases: its zero PRF scores follow the scorer's zero-denominator convention and do not measure positive-case detection. All fourteen RG4 concept F1 point estimates increase, without establishing

fourteen multiplicity-adjusted discoveries. Rare-label support is displayed in Tables 21 and 22; bold identifies point estimates only. Paired FN/FP transitions reconcile these differences as descriptive error accounting, not causal module effects.

Table 21: RG4 per-concept CheXbert PRF: three-seed means on 6,215 test reports from 287 patients. Support counts positive reference labels per seed. Bold: better matched estimate, including rounded ties. F1 difference/CI in percentage points; intervals are post-hoc and unadjusted.

		D+section			FCSL			F1 difference	
Concept	Support	P	R	F1	P	R	F1	F−D	95% CI
Enl. cardiomediastinum	1030	0.388	**0.334**	0.359	**0.450**	0.305	**0.360**	0.15	[-4.07, 4.14]
cardiomegaly	2630	**0.717**	0.585	0.644	0.675	**0.820**	**0.740**	9.56	[7.36, 12.13]
lung opacity	2264	**0.615**	0.334	0.433	0.569	**0.598**	**0.582**	14.88	[12.17, 17.70]
lung lesion	364	**0.476**	0.060	0.107	0.450	**0.197**	**0.261**	15.42	[9.12, 21.74]
edema	1781	0.641	0.483	0.551	**0.649**	**0.745**	**0.691**	14.00	[11.23, 17.36]
consolidation	587	0.308	0.133	0.186	**0.361**	**0.182**	**0.241**	5.46	[-0.85, 10.62]
pneumonia	1240	**0.536**	0.309	0.392	0.503	**0.555**	**0.527**	13.53	[10.11, 17.29]
atelectasis	2036	**0.554**	0.476	0.512	0.521	**0.773**	**0.622**	11.04	[8.55, 13.49]
pneumothorax	154	0.399	0.383	0.389	**0.442**	**0.491**	**0.465**	7.55	[-1.94, 15.85]
pleural effusion	2299	**0.753**	0.646	0.696	0.714	**0.826**	**0.766**	7.00	[5.15, 9.19]
Other pleural abnormality	256	0.218	0.057	0.090	**0.354**	**0.098**	**0.153**	6.25	[-1.83, 13.08]
fracture	324	**0.575**	0.110	0.184	0.468	**0.135**	**0.204**	1.93	[-4.24, 8.70]
support devices	2125	**0.788**	0.826	0.806	0.772	**0.851**	**0.810**	0.33	[-1.10, 1.75]
no finding	428	0.219	**0.706**	0.335	**0.379**	0.512	**0.435**	10.05	[5.08, 14.81]

Table 22: ABN per-concept CheXbert PRF: three-seed means on 844 test reports from 134 patients. Definitions and bolding follow Table 21. Sparse labels must be interpreted with their support; zero scores are not missing observations.

		D+section			FCSL			F1 difference	
Concept	Support	P	R	F1	P	R	F1	F−D	95% CI
Enl. cardiomediastinum	105	0.111	**0.105**	**0.107**	**0.225**	0.016	0.029	-7.78	[-13.88, -1.28]
cardiomegaly	252	0.474	**0.798**	0.594	**0.511**	0.783	**0.617**	2.29	[-1.35, 6.40]
lung opacity	398	0.587	0.490	0.533	**0.635**	**0.627**	**0.631**	9.74	[4.18, 15.02]
lung lesion	103	0.245	0.032	0.057	**0.501**	**0.136**	**0.214**	15.70	[4.27, 27.51]
edema	139	0.413	0.597	0.488	**0.480**	**0.767**	**0.591**	10.26	[3.60, 17.24]
consolidation	65	0.221	0.169	0.191	**0.406**	**0.359**	**0.380**	18.92	[6.46, 30.68]
pneumonia	124	0.316	0.207	0.250	**0.345**	**0.339**	**0.341**	9.13	[-3.03, 20.54]
atelectasis	208	0.358	0.606	0.450	**0.416**	**0.663**	**0.511**	6.13	[1.09, 11.23]
pneumothorax	0	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.00	[0.00, 0.00]
pleural effusion	188	0.573	0.711	0.634	**0.677**	**0.752**	**0.713**	7.84	[3.68, 12.24]
Other pleural abnormality	62	**0.221**	**0.151**	**0.175**	0.167	0.005	0.010	-16.44	[-28.39, -2.31]
fracture	78	**0.278**	**0.013**	**0.024**	0.000	0.000	0.000	-2.45	[-5.65, 0.00]
support devices	207	**0.706**	0.758	0.731	0.684	**0.800**	**0.738**	0.71	[-2.58, 4.23]
no finding	58	0.299	**0.385**	0.337	**0.376**	0.333	**0.354**	1.67	[-18.92, 16.06]

C.8 Complete language and relation metrics

Table 23 expands the main comparison to all recorded language and relation metrics, using the same validation-selected checkpoints. Additional columns are descriptive; they do not expand the set of tested hypotheses.

Table 23: Full test language/relation metrics for the same twelve selected runs. B1–B4: BLEU-1–4; M: METEOR; RL: ROUGE-L; RG: RadGraph-ERC. Bold: better matched task/seed score, including rounded ties. CIDEr retains its native scale. No new significance test is inferred for these added columns.

Task	Method	Seed	B1	B2	B3	B4	M	RL	CIDEr	RG
RG4	D+section	17	0.453	**0.325**	**0.244**	**0.190**	**0.198**	**0.386**	**0.400**	**0.264**
RG4	FCSL	17	**0.460**	0.320	0.233	0.177	0.196	0.368	0.332	0.241
RG4	D+section	42	0.453	**0.325**	**0.244**	**0.190**	**0.198**	**0.386**	**0.390**	**0.264**
RG4	FCSL	42	**0.454**	0.316	0.231	0.177	0.194	0.369	0.336	0.242
RG4	D+section	73	**0.451**	**0.323**	**0.241**	**0.187**	**0.197**	**0.384**	**0.382**	**0.261**
RG4	FCSL	73	0.450	0.314	0.229	0.174	0.193	0.369	0.321	0.242
ABN	D+section	17	0.216	**0.123**	**0.077**	**0.053**	**0.118**	**0.175**	**0.131**	**0.089**
ABN	FCSL	17	**0.217**	0.121	0.074	0.049	0.114	0.164	0.116	0.080
ABN	D+section	42	**0.233**	**0.132**	**0.083**	**0.057**	**0.123**	**0.175**	**0.126**	**0.087**
ABN	FCSL	42	0.219	0.122	0.074	0.050	0.116	0.167	**0.126**	0.078
ABN	D+section	73	**0.237**	**0.135**	**0.084**	**0.057**	**0.124**	**0.175**	**0.109**	**0.089**
ABN	FCSL	73	0.212	0.117	0.071	0.047	0.113	0.162	0.106	0.074

C.9 State correctness versus report correctness

Table 24: Reference-stratified state–report correctness (%). Correct/incorrect is relative to binary reference-report CheXbert labels. Each row partitions its label decisions and averages the three models; decisions are not independent patients. S: predicted state; R: generated report label.

Task	Reference	Decisions	S right/R right	S right/R wrong	S wrong/R right	Both wrong
RG4	Positive	17,518	65.91	4.14	0.67	29.28
RG4	Nonpositive	69,492	88.09	0.67	1.20	10.04
ABN	Positive	1,987	54.79	4.98	0.54	39.69
ABN	Nonpositive	9,829	89.26	0.24	0.98	9.53

For reference-positive labels, both state and report are correct in 65.91% of RG4 and 54.79% of ABN decisions, while both are incorrect in 29.28% and 39.69%. Correct states with incorrect report labels account for a further 4.14%/4.98%. The nonpositive stratum is substantially easier and more frequent; mixing it with positives would obscure missed findings. The decomposition separates observed error locations but does not prove which module caused an error. In particular, binary correctness does not distinguish blank from explicit negation or positive from uncertain, and does not evaluate location, degree, or relations.

C.10 Original-model state realization and interventions

These measurements use the original main models, not the retrained arms or shared request set in Appendix C.4. They remain separate evidence, not extra observations in the new paired comparison.

Table 25: **Original-model edits on two tasks.** Original seed-42 checkpoints; concept-macro rates (%). Direction and Joint use all eligible pairs. Any off: at least one non-target label changes; Mean off: label-averaged change. These cohorts differ from Table 3.

Task	Pairs	Target flip	Direction	Joint	Any off	Mean off
RG4	3,729	54.35	53.89	34.63	29.20	2.69
ABN	2,098	47.11	44.62	24.21	43.67	4.82

Positive-state realization is 92.25%/91.89% on RG4/ABN, not state accuracy; negative-state overreach is 0.98%/7.26% (only 234 ABN negative-header conditions). Both state and report miss 29.28%/39.69% of reference-positive decisions (Appendix C.9). Joint success is 34.63%/24.21%, versus directional success of 53.89%/44.62%; any non-target change occurs at 29.20%/43.67%. These binary-label criteria do not adjudicate clinical harm or fine-grained text stability.

For realization, the denominator contains explicitly positive header states (ID 1); a positive or uncertain report label counts as realized. For overreach, the denominator contains explicitly negative header states (ID 2), and the same report endpoint counts as overreach. Pooled counts across concepts and three seeds are 45,932/49,790 and 372/38,111 on RG4, and 5,436/5,916 and 17/234 on ABN. These are repeated concept–report–seed conditions, not independent patient samples. Empty denominators remain undefined.

The intervention uses the fixed seed-42 validation-selected model. Its starting program is anchored to saved model-predicted headers, not reference labels; residual IDs are recomputed once and held fixed within each intact/edited pair. Intact and edited reports are regenerated in the same batch with identical decoding. Exact-match rates against the saved main-evaluation reports are 78.71% on RG4 and 74.88% on ABN, so regenerated intact reports are not treated as exact replays of those outputs.

The binary change matrix groups positive/uncertain report labels as one endpoint. A diagonal entry counts any change of this endpoint on the edited concept; desired-polarity compliance instead asks whether the edited report reaches the requested endpoint. Compliance can be high when the intact report already has that endpoint, so it is not interchangeable with flip rate. Target flip averages fourteen diagonal entries, and mean off-target flip averages 182 off-diagonal entries, without weighting by concept sample count. All rates are computed from the same paired outputs.

Directional and joint success. For edited concept c, let b_c and e_c be binary labels of the intact and edited reports, and let u_c and v_c be the original and requested binary program endpoints. Directional success is $\mathbb{1}[b_c = u_c,\ e_c = v_c]$; joint selective success additionally requires $e_k = b_k$ for all $k \neq c$. The main denominator includes all header-eligible pairs, not just those with an initially aligned report. Conditional rates instead restrict to $b_c = u_c$: 3,492/3,729 RG4 and 1,934/2,098 ABN pairs. Already-desired outputs are not directional successes. Any off-target change is $\mathbb{1}[\exists k \neq c : e_k \neq b_k]$, distinct from the average over the thirteen non-target labels.

Patient bootstraps retain all cached pairs belonging to a sampled patient: 272 RG4 and 128 ABN patients have eligible edits. The concept-macro estimand weights all fourteen concepts equally; pair-pooled estimates weight by eligible pair count. RG4's macro joint interval is [32.06%, 36.77%]. Only 773 of 2,000 ABN draws contain all fourteen concepts for the all-pair rates (1,272 for the conditional rates), so their fixed-concept macro intervals are suppressed rather than averaging over a changing concept set. Pair-pooled intervals remain available. All intervals condition on the fixed model and original capped/eligible sample selection.

Table 26: Post-hoc seed-42 edit success and patient-cluster intervals (%). All: all header-eligible pairs; aligned: the intact report matches the original endpoint. Concept-macro and pair-pooled are different estimands. ABN macro intervals are suppressed for sparse missing concepts, not replaced by changing-set averages. NF: no finding, excluded only from non-target stability in the sensitivity row.

	RG4		ABN	
Metric / denominator	Rate	95% CI	Rate	95% CI
Concept-macro				
Direction / all	53.89	[51.09, 55.98]	44.62	–
Direction / aligned	58.48	[56.01, 60.60]	54.60	–
Joint / all	34.63	[32.06, 36.77]	24.21	–
Joint / aligned	37.07	[34.45, 39.38]	29.13	–
Any non-target change	29.20	[26.63, 31.38]	43.67	–
Joint / all, excluding NF	38.05	[35.82, 39.90]	26.44	–
Pair-pooled				
Direction / all	52.64	[49.47, 55.79]	46.04	[42.58, 49.72]
Direction / aligned	56.21	[52.79, 59.57]	49.95	[46.25, 53.71]
Joint / all	34.89	[31.95, 37.90]	28.36	[24.23, 32.71]
Joint / aligned	37.26	[34.12, 40.32]	30.77	[26.40, 35.30]
Any non-target change	27.27	[24.98, 29.59]	36.18	[31.26, 41.14]
Joint / all, excluding NF	38.24	[35.64, 40.83]	30.89	[26.41, 35.61]

Correlated concepts, particularly *no finding* versus disease states, make endpoint stability different from clinical consistency. Excluding no finding only from the non-target invariance check raises macro joint success to 38.05%/26.44%; this is a sensitivity analysis, not clinical adjudication. Fixed residuals test response to an edited message, not a medically validated counterfactual image. Blank/uncertain-state edits and fine-grained text changes are outside this analysis. Tables 26 to 28 retain aggregation and direction-specific denominators.

Table 27: RG4 direction-specific edit rates (%). N: eligible edits; N_a: initially aligned report labels; Dir./Joint: successes divided by N, not N_a. Directions refer to edited program endpoints, not clinically validated counterfactuals. – in a cell denotes zero denominator.

	Positive → negative				Negative → positive			
Concept	N	N_a	Dir.	Joint	N	N_a	Dir.	Joint
Enl. cardiomediastinum	22	17	77.27	54.55	278	278	49.28	31.29
cardiomegaly	220	216	91.36	68.18	80	79	80.00	57.50
lung opacity	266	221	52.26	32.71	34	31	47.06	26.47
lung lesion	115	76	63.48	22.61	24	24	66.67	33.33
edema	162	161	91.36	65.43	138	138	80.43	56.52
consolidation	35	23	65.71	31.43	265	264	24.53	14.34
pneumonia	116	111	80.17	47.41	184	177	45.65	20.11
atelectasis	295	290	37.29	30.17	5	2	20.00	20.00
pneumothorax	15	14	93.33	60.00	285	284	38.60	29.12
pleural effusion	148	144	87.84	60.81	152	152	85.53	57.89
Other pleural abnormality	83	70	60.24	32.53	0	0	–	–
fracture	94	88	82.98	61.70	113	104	61.95	34.51
support devices	300	263	13.00	9.67	0	0	–	–
no finding	300	265	14.67	12.67	0	0	–	–

Table 28: ABN direction-specific edits, with the definitions of Table 27. Very small or absent cells cannot establish reliable concept-level control; a 100% rate with one pair is not a precise estimate.

	Positive → negative				Negative → positive			
Concept	N	N_a	Dir.	Joint	N	N_a	Dir.	Joint
Enl. cardiomediastinum	2	1	50.00	0.00	9	8	66.67	22.22
cardiomegaly	414	407	69.57	45.41	7	5	42.86	42.86
lung opacity	465	406	34.62	23.01	13	12	76.92	61.54
lung lesion	39	25	46.15	20.51	0	0	–	–
edema	159	157	59.12	41.51	10	9	90.00	70.00
consolidation	57	50	73.68	26.32	9	9	22.22	11.11
pneumonia	72	65	30.56	15.28	11	9	27.27	0.00
atelectasis	294	278	19.05	11.22	0	0	–	–
pneumothorax	6	5	83.33	33.33	1	1	0.00	0.00
pleural effusion	208	195	55.77	25.00	3	3	100.00	33.33
Other pleural abnormality	4	0	0.00	0.00	0	0	–	–
fracture	0	0	–	–	2	2	0.00	0.00
support devices	252	231	33.73	21.03	10	9	60.00	60.00
no finding	51	47	70.59	62.75	0	0	–	–

Direction-specific point estimates must be interpreted with their denominators; sparse or degenerate cells do not establish precise control. The main concept-macro interval requires the complete fixed concept set, as described above.

Table 29: Per-concept paired measurements. N_+/N_- pool positive/negative header conditions across three seeds; realization (Real.) and overreach (Over.) use those denominators. Edit N, target flip, and desired-polarity compliance (Comp.) use seed 42. Rates are percentages; – means an empty denominator.

Task	Concept	N_+	Real.	N_-	Over.	Edit N	Flip	Comp.
RG4	enlarged cardiomediastinum	553	77.9	4683	0.1	300	51.3	53.0
RG4	cardiomegaly	8719	98.1	3585	0.9	300	88.7	89.7
RG4	lung opacity	8748	79.7	1011	3.6	300	53.3	66.0
RG4	lung lesion	735	69.9	74	2.7	139	64.0	92.1
RG4	edema	4783	98.4	2892	1.3	300	86.3	86.7
RG4	consolidation	634	77.6	4525	0.2	300	29.3	33.7
RG4	pneumonia	1709	95.8	3040	4.2	300	60.0	62.0
RG4	atelectasis	7097	97.9	57	29.8	300	37.7	39.0
RG4	pneumothorax	470	95.7	10346	0.4	300	41.3	42.0
RG4	pleural effusion	6990	98.7	7481	0.4	300	86.7	88.0
RG4	pleural other	237	76.4	0	–	83	60.2	75.9
RG4	fracture	224	91.1	413	7.3	207	72.9	77.3
RG4	support devices	7663	90.2	4	100.0	300	14.3	24.0
RG4	no finding	1228	84.5	0	–	300	14.7	26.3
ABN	enlarged cardiomediastinum	5	60.0	27	3.7	11	72.7	72.7
ABN	cardiomegaly	1147	99.1	23	17.4	421	69.6	70.8
ABN	lung opacity	1348	85.5	46	8.7	478	38.3	45.8
ABN	lung lesion	109	75.2	0	–	39	51.3	76.9
ABN	edema	460	98.9	33	6.1	169	61.5	62.1
ABN	consolidation	180	78.9	28	0.0	66	71.2	72.7
ABN	pneumonia	222	91.4	33	12.1	83	34.9	36.1
ABN	atelectasis	847	92.7	0	–	294	21.4	22.1
ABN	pneumothorax	18	88.9	3	0.0	7	71.4	85.7
ABN	pleural effusion	634	94.8	6	16.7	211	58.3	60.7
ABN	pleural other	10	10.0	0	–	4	0.0	100.0
ABN	fracture	0	–	6	0.0	2	0.0	0.0
ABN	support devices	777	92.3	29	3.4	262	38.2	39.7
ABN	no finding	159	88.7	0	–	51	70.6	78.4

C.11 Additional message and training-objective ablations

These exploratory ablations use report-training seed 42, the shared Stage-A initialization, and three RG4 or five ABN Stage-B epochs. Every configuration is evaluated at the earliest checkpoint maximizing the validation composite in Equation (11). All evaluated configurations are reported under this same selection rule; final-epoch checkpoints are not substituted according to test performance. These experiments do not add independent seeds to Table 1.

Message content. Removing all eight residual atoms reduces the RG4 layout from 202 to 74 bits. The style extension adds one 64-way ID learned from training-report clusters, increasing the budget to 208 bits. Adding a three-way, section-conditioned sentence-count bucket gives 210 bits. Training uses reference-derived attribute labels; inference predicts both attributes from the available evidence. These variants do not introduce prior-report style or template fields, and preserve the main model's prior clinical-state fields.

Training objectives. The necessity objective is a training-only hinge with weight .05 and margin .05, applied every fourth optimizer update. It compares concept-token NLL under intact and single-state-counterfactual programs, holding other IDs fixed and recomputing changes. A separate configuration combines this objective with four residual atoms (138 bits); this is not a loss-only ablation.

On top of the style+length message, residual style/template supervision has weight .3, with unavailable template labels masked. The residual clinical-state adversary uses weight .1 and gradient reversal at the quantized residual input, ramping to full strength over the first quarter of training. Its feature extractor is separate from the narrative head. We evaluate these objectives separately and jointly; weak adversarial prediction does not establish statistical independence. State-weighted token NLL is a separate variant with weights $[1, 1.5, 1, 1.25]$ for blank/positive/negative/uncertain.

Calibration variants. Two variants refit negative-versus-blank thresholds while holding positive-endpoint thresholds fixed: one maximizes negative-state F1 and the other matches the negative-state rate. Neither changes the main model's reported calibration policy.

C.12 VALIDATION-SELECTED ABLATION RESULTS

Table 30: Additional ablations at validation-selected checkpoints, seed 42. RG4 has 6,215 test reports; ABN has 844. E: selected Stage-B epoch. Bold marks the highest displayed RG4 score, including rounded ties, not statistical significance. All scores retain their original scale.

Task	Configuration	E	Micro-F1	B4	RL	RG
RG4	FCSL	2	0.639	0.177	0.369	0.242
	No residuals	2	0.636	0.131	0.326	0.198
	+style	3	0.636	0.180	**0.374**	**0.250**
	+style+length	3	0.638	**0.182**	0.373	0.246
	+necessity	2	**0.640**	0.173	0.367	0.242
	+necessity, 4 residuals	3	0.633	0.173	0.366	0.237
	Style+length +residual supervision	2	0.637	0.178	0.372	0.246
	Style+length +residual GRL	3	0.639	**0.182**	0.373	0.246
	Style+length +both objectives	2	0.638	0.178	0.373	0.247
	+state-weighted NLL	3	0.631	0.177	0.370	0.243
	Negative-F1 calibration	2	0.636	0.180	0.369	0.245
	Negative-rate calibration	2	0.639	0.177	0.369	0.243
ABN	+necessity	3	0.531	0.050	0.163	0.068

Removing residuals substantially reduces BLEU-4 while only slightly changing micro-F1. The style+length extension improves BLEU-4 by 0.57 percentage points before rounding, without improving micro-F1. Residual gradient reversal barely changes its language scores, whereas state-weighted NLL lowers micro-F1 to 0.631. The necessity objective gives a small RG4 micro-F1 increase with lower BLEU-4; combining it with fewer residuals lowers both. These results characterize distinct content and objective choices rather than establish a uniformly better configuration. None of the extensions changes the main method or the matched-interface comparisons.

D PARAMETER ACCOUNTING

We distinguish nominal backbone size, all registered parameters, and Stage-B trainable parameters. The latter two sum $\mathrm{numel}(p)$ over all parameters or those with `requires_grad=True`. Per-model counts include the language model, online encoders, and conditioning modules, excluding offline labelers/scorers, optimizer state, non-parameter buffers, and replica duplication. They are not checkpoint sizes or FLOPs.

Table 31: Exact RG4 Stage-B trainable parameters by component. Component sums equal the recorded trainable totals for each of seeds 17/42/73.

Component	D+section	FCSL
Online encoder tuning	28,357,632	28,357,632
Decoder LoRA	10,822,656	10,822,656
Continuous fusion	5,065,217	0
Context-type embeddings	768	0
Clinical program	0	19,029,560
Structured prefix	0	263,168
Layer-local writes	0	21,274,632
Trainable total	44,246,273	79,747,648
Frozen total	919,622,720	919,757,888
System total	963,868,993	999,505,536

Online tuning updates the final four visual-encoder blocks. The decoder base is frozen, with LoRA parameters updated alongside the program and access modules. RG4 FCSL updates 7.97871% of

its registered system; D+section updates 4.59049%. The FCSL–D+section difference is 35,501,375 trainable parameters (1.80× the control's trainable count), not a reduction. The 0.8B-versus-7B literature comparison instead concerns nominal decoder scale: $0.8/7 = 11.43\%$, approximately one ninth, with an 88.57% reduction. It does not quantify full-system compression, training-time savings, or hardware efficiency. Stage-A pretraining and offline supervision are outside this Stage-B parameter comparison, and no complete wall-clock, memory, or energy benchmark is claimed.

ABN FCSL has 999,466,112 total and 79,708,224 trainable parameters (7.97508%); its clinical-program component has 18,990,136 trainable parameters, 39,424 fewer than RG4. Other trainable components are unchanged. D+section has the same 963,868,993 total and 44,246,273 trainable parameters on both datasets. Thus the compact-backbone and partial-adaptation statements hold for both tasks, whereas the extra conditioning capacity relative to D remains explicit.

Component counts are consistent across all twelve main runs and are provided in the supplement. LLM-RG4's 7B backbone is specified in its official implementation: `https://github.com/zh-wang-med/llm-rg4`.

Table 32: Parameter counts for the RG4 seed-42 interface and scale controls. The first four rows use a 0.8B backbone. Total parameters include frozen modules, unlike the trainable count; removal of frozen quantizer parameters changes Full/continuous totals without changing trainable capacity.

Model	Total parameters	Trainable parameters	Trainable (%)
Full	999,505,536	79,747,648	7.98
Continuous	999,370,368	79,747,648	7.98
Prefix-only	978,230,904	58,473,016	5.98
Write-only	999,242,368	79,484,480	7.95
FCSL (4B)	4,569,665,856	196,549,696	4.30
D+section (4B)	4,439,657,473	66,676,481	1.50

The five-fold nominal backbone difference is not a compute or memory benchmark. Full and continuous share trainable capacity, while prefix-only is smaller; discreteness does not establish a parameter advantage over these controls.

E Qualitative Evaluation Details

Selection and alignment. Figure 3 contains two outcome-selected examples from different comparisons, not a shared-checkpoint generation/editing experiment. Panel (a) uses the main seed-42 FCSL/D+section runs. Among 6,215 paired test cases, SN findings-only rows with 18–75 reference tokens and 15–85 tokens in both outputs yielded 566 candidates; keyword groups and lexical overlap narrowed these to 35 for review. The displayed example illustrates reference-supported generation without an additional diagnostic hypothesis.

Panel (b) uses the validation-selected Full, Continuous, and Prefix-only models from the shared-request evaluation in Appendix C.4. We reviewed 83 SN, readable-length requests with all three intact reports aligned to the target and joint success only for Full, then compared complete three-model texts for six candidates. The selected pneumothorax reversal clearly expresses the requested target while retaining coexisting findings. All three models preserve their thirteen non-target binary labels, so the example does not show fewer non-target changes for Full. Selection is post hoc and outcome-dependent; neither panel estimates prevalence, clinical correctness, or a causal effect of discretization.

Images and reports are matched to saved test cases; no new reports are generated. Authorization is described in Appendix B.5. The GT in panel (b) is the original reference, not an edited counterfactual target.

E.1 Complete texts for Figure 3

Ellipses in the main figure mark omitted text. The following reports preserve all clinical words, normalizing only leading section labels and punctuation spacing.

Panel (a): report generation. *GT (RG4 reference).* there is enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis. central catheter remains in place.

D+section. there is enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis at the bases. in the appropriate clinical setting supervening pneumonia would have to be considered. a right subclavian catheter extends to the lower portion of the svc.

FCSL. there is enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis at the bases. central catheter remains in place.

Panel (b): requested pneumothorax reversal. *GT (original case).* the heart is massively enlarged. small bilateral pleural effusions are present. diffuse interstitial opacities with perihilar predominance are likely secondary to mild interstitial pulmonary edema. no focal consolidations are present concerning for pneumonia. there is no pneumothorax. the mediastinal contours are within normal limits. impression: mild interstitial pulmonary edema. massive cardiomegaly. small bilateral pleural effusions.

FCSL (Full), before edit. a right internal jugular central venous catheter tip terminates in the mid svc. the heart size is moderately enlarged. the aorta is tortuous and diffusely calcified. there is mild pulmonary edema with small bilateral pleural effusions. bibasilar airspace opacities likely reflect compressive atelectasis. no pneumothorax is identified. there are no acute osseous abnormalities. impression: mild congestive heart failure with small bilateral pleural effusions and bibasilar atelectasis.

FCSL (Full), after edit. a right internal jugular central venous catheter tip terminates in the mid svc. the heart size is moderately enlarged. the aorta is tortuous and diffusely calcified. there is mild pulmonary edema with small bilateral pleural effusions. bibasilar airspace opacities likely reflect compressive atelectasis. there is a small right apical pneumothorax. no acute osseous abnormality is seen. impression: mild congestive heart failure with small bilateral pleural effusions and bibasilar atelectasis. small right apical pneumothorax.

Continuous, before edit. moderate cardiomegaly is present. mild pulmonary edema is present. small bilateral pleural effusions are present. there is no focal consolidation or pneumothorax. a right-sided hemodialysis catheter terminates in the right atrium. impression: mild pulmonary edema and small bilateral pleural effusions are present. moderate cardiomegaly is present.

Continuous, after edit. a right-sided dialysis catheter terminates in the right atrium. moderate cardiomegaly is present. there is mild pulmonary edema. small bilateral pleural effusions are present. there is no pneumothorax. the visualized osseous structures are unremarkable. impression: moderate cardiomegaly with mild pulmonary edema and small bilateral pleural effusions.

Prefix-only, before edit. a right internal jugular central venous catheter tip terminates in the mid svc. moderate cardiomegaly is present. mild pulmonary edema is present. small bilateral pleural effusions are present with associated bibasilar atelectasis. no pneumothorax is identified. there are no acute osseous abnormalities. impression: mild pulmonary edema and small bilateral pleural effusions are present with associated bibasilar atelectasis.

Prefix-only, after edit. a right internal jugular central venous catheter tip terminates in the mid svc. the heart is moderately enlarged. there is mild pulmonary edema. small bilateral pleural effusions are present. bibasilar patchy opacities likely reflect atelectasis. no pneumothorax is identified. there are no acute osseous abnormalities. impression: mild pulmonary edema and small bilateral pleural effusions. bibasilar atelectasis.