ANCHORING THE DRIFT: GATED-ANCHOR ON-POLICY DISTILLATION FOR FAITHFUL RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Small vision-language models are attractive for deploying clinical report generation on local hardware, yet they need supervision from a stronger model. On-policy distillation supplies dense token-level supervision on the student's own rollouts, and a growing family of methods improves it by reshaping the teacher Kullback-Leibler term. Every repair in that family acts on the teacher term itself, so what it can deliver stays bounded by what the teacher's own distribution contains, and wherever the teacher is unreliable the student inherits that failure; on hard studies a second failure compounds it, since the student drifts and leaves the teacher scoring an already-wrong prefix. We present GA-OPD (Gated-Anchor On-Policy Distillation), which answers both with one construction. A ground-truth anchor, a teacher-forced cross-entropy term on the reference report, never passes through the teacher and therefore stays valid where the teacher's own distribution is misaligned with the reference; a per-token drift gate routes it in by attenuating the on-policy term exactly where the student prefix has diverged from the reference. The mixing rule is what carries the result. In the regime where the teacher fails, the same anchor entering at a constant weight reaches 7.22 RadGraph and the drift-routed anchor reaches 19.36, against 7.15 for plain on-policy distillation; GA-OPD exceeds three state-of-the-art variants of this family there by 1.92 to 12.12 points, and two of them collapse together with plain on-policy distillation. Stratifying by the teacher's own per-sample faithfulness, an axis independent of the compared arms, the advantage grows with teacher failure. On the full three-corpus chest X-ray test set of 4,676 studies the same 4B student attains the highest CheXbert micro-F1 among all arms, 59.34, matching a frozen 30B medical teacher while serving on a single accelerator.

1 INTRODUCTION

Radiology report generation asks a model to turn one or more chest X-ray images into a free-text description of clinical findings (Chen et al., 2020; 2021; Miura et al., 2020; Tanida et al., 2023). Because the output is read by clinicians, the quality that matters is clinical faithfulness rather than surface fluency, and it is measured by structured metrics such as CheXbert label agreement, RadGraph entity and relation overlap, and RaTEScore (Smit et al., 2020; Jain et al., 2021; Zhao et al., 2024), whose agreement with radiologist judgment has itself been the subject of dedicated study (Yu et al., 2023; Ostmeier et al., 2024). Large medical vision-language models write such reports well (Chen et al., 2024; Bannur et al., 2024), yet their size makes local deployment inside a hospital difficult (Zambrano Chaves et al., 2024). A small student model can be deployed, but it needs a stronger source of supervision before it becomes clinically reliable.

On-policy distillation (OPD) is an appealing source. The student samples a report itself, a frozen teacher scores the tokens of that rollout, and a divergence term pulls the student toward the teacher (Agarwal et al., 2023; Gu et al., 2023). The supervision is dense, defined on the student's own distribution, and needs only images rather than human-written references. A growing family of methods improves this recipe by reshaping the teacher term: restricting the divergence to the teacher's top-k support and renormalizing it, gating tokens by teacher entropy, or rewriting the divergence as a

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.

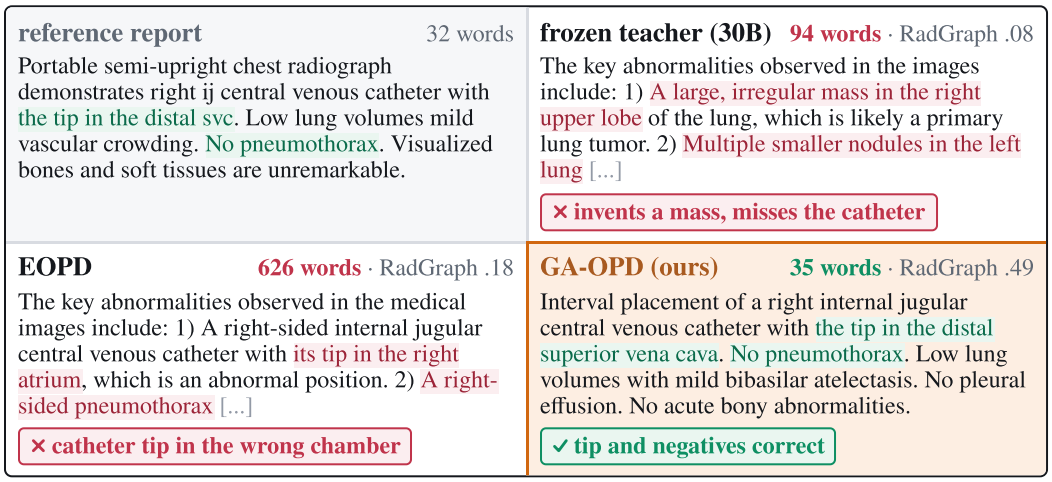


Figure 1: One CheXpert study, on which the frozen teacher scores below the supervised student. The teacher misses the catheter and invents a lung mass; the arm whose loss is a divergence over it inherits the enumeration formula and amplifies 94 words to 626, with the catheter tip in the wrong chamber; the drift-gated anchored objective writes at reference length and gets the tip and the negatives right. Table 2 gives the split as a whole.

policy gradient with a clipped reward (Fu et al., 2026; Jin et al., 2026; Ko et al., 2026; 2024; Wang et al., 2025).

These methods modify how teacher feedback is used, but teacher-only objectives provide no direct correction from reference reports when that feedback is unreliable. Prior work identifies several challenges: teacher feedback can lose discriminative power along student-generated prefixes, a stronger teacher does not necessarily yield a better student, and distillation can amplify repetitive generation (Liu et al., 2026b; Wang et al., 2026b; Luo et al., 2026). Two distinct, potentially interacting sources of unreliable supervision motivate our approach. First, student-generated prefixes can depart from those on which the teacher provides useful guidance, leaving local next-token supervision insufficient to correct earlier errors (Jiang et al., 2026; Fu et al., 2026). Second, the teacher's outputs can be misaligned with the target reporting task even before such deviations occur. We study this second case through a controlled teacher-instruction mismatch on CheXpert. Reweighting teacher feedback may mitigate its effects, but does not itself introduce a reference-based corrective signal. These observations motivate combining supervision on reference prefixes with a mechanism that adjusts the strength of teacher feedback along student-generated rollouts.

We introduce **GA-OPD** (Gated-Anchor On-Policy Distillation), which combines reference supervision with adaptive attenuation of teacher feedback on student-generated rollouts. A reference anchor applies teacher-forced cross-entropy to the reference report, providing a supervision source independent of the teacher's predictions and the sampled rollout prefix. A per-token drift gate, computed from the bag-of-tokens overlap between student and reference prefixes, attenuates the on-policy distillation term when lexical agreement is low. The anchor retains a fixed coefficient, while the gate adjusts the strength of teacher supervision relative to it. An additional teachability gate, adopted from prior analysis, modulates the teacher term using the teacher–student probability margin on the teacher's preferred token. Figure 2 summarizes the objective, and Figure 1 illustrates its behavior under the controlled teacher-instruction mismatch. In this setting, GA-OPD improves RadGraph from 7.22 for the ablation retaining the anchor and teachability gate but omitting the drift gate to 19.36. Across the combined test set, it achieves the highest CheXbert micro-F1 among the evaluated methods, although its gains are not uniform across metrics or teacher-agreement strata.

Our contributions are as follows.

- **Reference-guided weighting of on-policy supervision.** We propose a prefix-overlap gate that adjusts token-level distillation strength while retaining teacher-forced supervision on reference reports. The construction supplements teacher feedback with a reference objec-

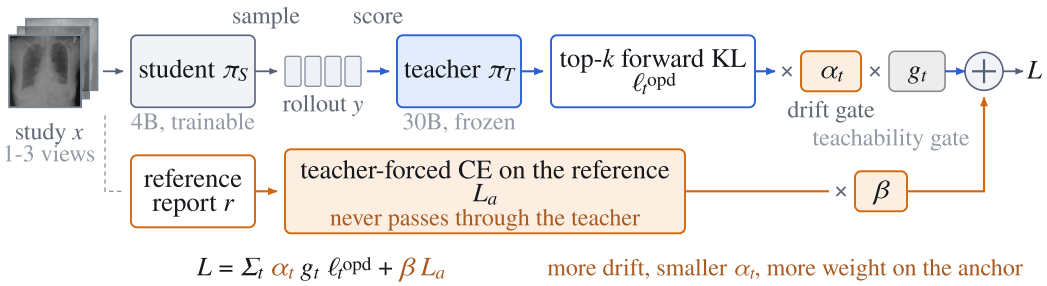


Figure 2: Overview of GA-OPD. The student samples a report and the frozen teacher scores it with a top-k forward Kullback-Leibler term that two per-token gates attenuate; in parallel a ground-truth anchor supplies teacher-forced supervision that never passes through the teacher. The drift gate sets the balance between the two sources, and it is the component the ablation isolates. The radiograph is from an IU-Xray test study.

tive and provides a simple mechanism for attenuating distillation when rollout prefixes have low reference overlap.

- **Evidence of robustness under controlled teacher mismatch.** On the 234-study CheXpert subset with a teacher-instruction mismatch, GA-OPD achieves 19.36 RadGraph, compared with 7.15 for plain OPD and 17.81 for continued supervised fine-tuning. Component ablations and training trajectories characterize how the evaluated configurations differ in reference agreement, output length, and repetition.
- **Characterization of aggregate gains and trade-offs.** Across 4,676 studies from three chest X-ray corpora, the 4B student achieves the highest CheXbert micro-F1 among the evaluated methods on both label sets. Stratification by the frozen teacher's reference agreement reveals larger relative gains on teacher-hard studies and losses on teacher-easy studies, identifying where the method helps and where it trades away performance.

2 Related Work

On-policy distillation and its divergence-side refinements. On-policy distillation trains a student on its own samples and takes per-token supervision from a frozen teacher (Agarwal et al., 2023; Gu et al., 2023). A growing body of work reports that this supervision is not automatically reliable. The teacher's next-token distribution loses discriminative power as the student prefix lengthens, so an early deviation compounds along the trajectory (Liu et al., 2026b; Jiang et al., 2026); a stronger teacher does not guarantee a better student (Wang et al., 2026b); and the student can inflate its output with repeated text (Luo et al., 2026); surveys collect both the failures and the repairs (Zhang, 2026; Armandpour et al., 2026). What we observe in Section 4.2 agrees with that literature, in a task where the quantity at risk is clinical faithfulness rather than answer accuracy. The repairs share a structure: truncating the divergence to the teacher's top-k support and renormalizing it, gating tokens by teacher entropy, writing the divergence as a clipped policy gradient, reweighting positions by how far the current policy has moved from the one that produced the rollout, or replacing one teacher with several (Fu et al., 2026; Jin et al., 2026; Ko et al., 2026; 2024; Zheng & Jiang, 2026; Wang et al., 2026a). Each of them acts on the teacher term, so what it can deliver stays bounded by what the teacher's distribution contains. We keep the rollout-and-score structure and the rollout itself, and move a reference term into the update path instead, weighted per token by measured drift; where drift-side remedies remove unreliable teacher signal from the drifted region (Jiang et al., 2026; Fu et al., 2026), we inject an external ground-truth term exactly there.

Distillation mixed with ground-truth supervision. Combining a teacher divergence with a cross-entropy term on references is the classical form of knowledge distillation (Hinton et al., 2015), and sequence-level distillation for text generation routinely interpolates the teacher objective with token-level negative log-likelihood on the reference (Kim & Rush, 2016; Wen et al., 2023). Equation 8 with both gates held at one is exactly that objective. Our delta is not the mixture but the mixing rule:

measured drift, not a fixed coefficient, sets the balance between the two sources. Drift also gates elsewhere, on a different quantity: blockwise gating rescales the on-policy loss by how stale the behavior policy that produced the rollout has become (Zheng & Jiang, 2026), which never involves the reference, whereas ours measures how far the rollout has moved from the reference itself. It also relates to exposure-bias work, which schedules free-running against teacher-forced inputs rather than reweighting per token by a measured quantity (Bengio et al., 2015; Ranzato et al., 2016; Ross et al., 2011). Section 4.5 tests the mixing rule directly: the same mixture with the rule removed recovers nothing where the teacher fails.

Gated supervision and reinforcement learning for reporting. Joint divergence-and-reward training, preference-based distillation, and dual-path routing by a correctness verifier all mix two supervision sources (Xu et al., 2025; Zhang et al., 2025b; Zheng et al., 2026); in report generation, reinforcement learning against clinical rewards optimizes faithfulness directly and without a teacher (Zhang et al., 2025a; Miura et al., 2020; Delbrouck et al., 2022; Xiao et al., 2025). All of them route supervision by verifier outcomes, trajectory rewards, or fixed weights, and none routes a ground-truth term by measured prefix drift. For the teachability gate we adopt an existing observation that the teacher's discriminative lead decays along the suffix (Liu et al., 2026a) rather than claiming it.

3 METHOD

3.1 SETUP AND NOTATION

Let x denote the images of a study and $r = (r_1, \dots, r_M)$ the reference report. The student π_S samples a rollout $y = (y_1, \dots, y_T) \sim \pi_S(\cdot \mid x)$, and a frozen teacher π_T scores the tokens of that rollout. For brevity we write $\pi(\cdot \mid t)$ for $\pi(\cdot \mid x, y_{<t})$. Writing $\mathcal{K}_t$ for the teacher's top-k token set at step t, the on-policy term is a top-k truncated forward Kullback-Leibler term over that token set,

$$\ell_t^{\mathrm{opd}} = \sum_{v \in \mathcal{K}_t} \pi_T(v \mid t) \, \log \frac{\pi_T(v \mid t)}{\pi_S(v \mid t)}. \tag{1}$$

Since π_T is frozen and only ever evaluated, all information carried by ℓ^{opd} is bounded by the teacher. This is the ceiling our method addresses.

3.2 DRIFT-GATED GROUND-TRUTH ANCHOR

The anchor is a teacher-forced cross-entropy term on the reference report,

$$\mathcal{L}_{\mathrm{a}} = -\sum_{t=1}^{M} \log \pi_S(r_t \mid x, r_{<t}). \tag{2}$$

Two properties matter. First, $\mathcal{L}_{\mathrm{a}}$ never queries the teacher, therefore it can carry supervision the teacher's distribution does not provide. Second, it is teacher-forced on the reference prefix, therefore it is unaffected by whatever the student has already written.

Applying the anchor uniformly would trade away the benefit of on-policy training on the many studies where the teacher is reliable. We therefore gate the on-policy term by how far the student has already drifted from the reference. Let $c_u(w)$ count the occurrences of token id w in a sequence u, and define the bag-of-tokens overlap of two sequences as

$$\mathrm{ov}(u, v) = \frac{2 \sum_w \min\big(c_u(w),\, c_v(w)\big)}{|u| + |v|} \in [0, 1], \tag{3}$$

the harmonic mean of the two directional overlap rates. The drift gate at response position t is

$$\alpha_t = \mathrm{ov}\big(y_{\le t},\; r_{\le \min(t, M)}\big). \tag{4}$$

Four details fix the operator. It is computed on the student tokenizer's ids, whose vocabulary the teacher shares. The multiset intersection is maintained incrementally as the rollout is decoded, so the gate costs $O(1)$ per token and $O(T)$ per rollout, with no recomputation from scratch at any position. It is evaluated in the rollout worker and enters the objective as a constant, so no gradient

flows through it and the rollout expectation is not differentiated through. Finally α_1 is 1 when $y_1 = r_1$ and 0 otherwise, and the gate applied at the position that predicts y_t is computed from prefixes that include y_t.

A large α_t means the student still tracks the reference and the teacher signal is worth following. A small α_t means the student has drifted, the teacher is being asked to continue an already wrong prefix, and its signal is worth less. Since α_t multiplies the on-policy term while the anchor carries a constant weight, the relative influence of the teacher-independent anchor rises exactly on those reports where the student has drifted. We call this behavior conditional insurance: the anchor is cheap where on-policy training already works, yet it takes over where it does not. Throughout, drift-gated anchoring names this construction: the drift gate does not multiply the anchor, it sets the balance between the teacher-dependent term and the anchor as a function of measured drift. An anchor entering at a constant weight does not have this property, and Section 4.5 shows it does not recover faithfulness where the teacher fails.

3.3 Properties of the Drift Gate

Once the rollout passes the reference length the gate admits a closed bound. For $t \geq M$ the second argument of Eq. 4 is the full reference, so the multiset intersection is at most M and

$$\alpha_t \;\leq\; \frac{2M}{t+M}, \tag{5}$$

with equality only when the rollout has already covered every reference token. A rollout that keeps generating therefore loses its teacher-side supervision at rate $O(1/t)$ whatever it writes, while the weight on the anchor is untouched: the longer it runs past the reference, the more of its update comes from the reference alone.

The two ends of the gate are the two objectives it interpolates. With both gates held at one, Eq. 8 is the classical mixture of a teacher divergence with a reference cross-entropy; as $\alpha_t \to 0$ the update at that token is reference supervision alone. The interpolation runs per token along one rollout, not a per-report choice between two losses. The overlap itself is order-free by design: an order-sensitive comparison would score a finding stated in another order as drift, which in a report it is not.

3.4 Teachability Gate

The teacher's advantage over the student is not uniform along a report. Late tokens often retain a non-zero advantage while losing the local contrast that makes dense feedback useful (Liu et al., 2026a). We adopt a second per-token gate that measures the teacher's lead on its own preferred token,

$$v_t^* = \arg\max_v \pi_T(v \mid t), \qquad m_t = \pi_T(v_t^* \mid t) - \pi_S(v_t^* \mid t), \tag{6}$$

$$g_t = \mathrm{clamp}(m_t/s,\, 0,\, 1)\,, \tag{7}$$

with scale s. The lead is read on the teacher's own preferred token rather than on each model's separate maximum, and the clamp sets $g_t = 0$ whenever the teacher holds no lead there, including when m_t is negative. Where the teacher holds a clear lead the on-policy term is preserved; where it does not, the term is attenuated. We adopt this component from prior analysis and report its contribution separately in the ablation.

3.5 Training Objective

The complete objective combines the twice-gated on-policy term on rollouts with the anchor on references,

$$\mathcal{L} = \mathbb{E}_{y\sim\pi_S}\left[\sum_{t=1}^{T} \alpha_t\, g_t\, \ell_t^{\mathrm{opd}}\right] + \beta\, \mathcal{L}_{\mathrm{a}}, \tag{8}$$

where β sets the anchor strength.

In implementation the rollout worker augments a batch of N studies into N rollout samples and N teacher-forced reference samples, of which the trainer consumes the first N. An anchored arm therefore optimizes on 16 rollouts and their 16 references per step, against 32 rollouts for an unanchored

Table 1: In-domain test results over all 4,676 studies, in percent. Parentheses name each arm's components, and bold marks where GA-OPD leads every other arm.

Method	BLEU-1	ROUGE-L	CheXbert-5	CheXbert-14	RadGraph	RaTEScore
Zero-shot student (4B)	7.22	9.42	39.35	37.44	10.08	51.12
Supervised fine-tuning	29.55	24.97	56.24	49.87	22.25	55.99
Supervised fine-tuning, budget-matched	29.38	24.99	56.13	49.84	22.27	56.05
Plain OPD (ℓ^{opd})	29.25	24.83	55.30	48.70	23.64	57.55
+ teachability gate ($g_t\ell^{\mathrm{opd}}$)	29.44	24.62	54.81	47.97	23.87	57.88
+ drift gate + anchor ($\alpha_t\ell^{\mathrm{opd}}+\beta\mathcal{L}_{\mathrm{a}}$)	31.98	25.13	56.45	49.68	23.28	57.79
+ teachability gate + anchor ($g_t\ell^{\mathrm{opd}}+\beta\mathcal{L}_{\mathrm{a}}$)	27.03	24.28	55.92	48.79	22.57	56.68
Revisiting-OPD	25.86	23.94	55.32	47.89	23.02	57.21
EOPD	24.74	24.75	57.18	50.19	24.66	58.07
REOPOLD	32.71	25.14	58.13	51.57	23.12	56.58
GA-OPD (ours) ($\alpha_t g_t\ell^{\mathrm{opd}}+\beta\mathcal{L}_{\mathrm{a}}$)	31.40	**25.26**	**59.34**	**51.81**	23.30	57.46
Teacher (30B, frozen)	33.58	25.18	57.40	50.86	23.18	56.88

arm, at the same batch of 32 sequences; anchored arms see fewer rollouts and studies per step, not more data. The two terms are routed by a per-sample flag, so the anchor never consumes a teacher forward pass. Both terms enter one per-token loss tensor, averaged by a token mean over the batch's response mask, so every supervised token carries equal weight. Two properties follow. At $\beta = 1$ the anchor's share of the objective for one study is exactly $M/(M + T)$, fixed by the reference and rollout lengths; no coefficient is tuned to set it. And since α_t multiplies the on-policy term alone, the balance between teacher-dependent and teacher-independent supervision shifts per token with measured drift; β, fixed at one throughout, only scales the anchor as a whole. This is the mixing rule the ablation in Section 4.5 isolates; Algorithm 1 in Appendix B states one optimization step.

4 Experiments

4.1 Setup

Data. We train on a pool of three chest X-ray report corpora, MIMIC-CXR, IU-Xray and CheXpert (Johnson et al., 2019; Demner-Fushman et al., 2016; Irvin et al., 2019), comprising 462,634 training studies. Free-text findings and impression sections for the CheXpert studies are taken from the CheXpert Plus release (Chambon et al., 2024); MIMIC-CXR and CheXpert Plus are distributed under credentialed access with data use agreements, IU-Xray under open access. We hold out a validation set of 3,156 studies from that pool, grouped at the patient level with a fixed seed, and verify that no validation patient appears in any test split. The in-domain test set contains 4,676 studies: 3,852 from MIMIC-CXR, 590 from IU-Xray and 234 from CheXpert.

Models and training. The student is Qwen3-VL-4B-Instruct (Bai et al., 2025), a general-domain vision-language model; the teacher is Hulu-Med-30A3 (Jiang et al., 2025), a 30B sparse mixture-of-experts medical model fine-tuned from the student's own family, so the two share a vocabulary; it stays frozen throughout. All distillation arms start from the same supervised fine-tuned checkpoint, share an identical distillation training set, and run under an identical optimization budget; they differ in the distillation loss. Training runs on four actor and four teacher devices (Kwon et al., 2023; Sheng et al., 2025). Appendix B.6 lists every hyperparameter.

Model selection. Every checkpoint is selected on the held-out validation set by peak RadGraph, and the test set is evaluated once for the selected checkpoint only.

Difficulty stratification. Several analyses below stratify test studies by difficulty, indexed by the frozen teacher's own per-sample RadGraph, an axis that depends on neither of the arms it compares. Because the axis is the teacher's agreement with the reference while the anchor is a term on that reference, we residualize the axis against the supervised student's per-sample RadGraph and keep the residual, which isolates the teacher failure not shared with a teacher-free student. Appendix C.1 shows studies from each end of the axis.

Table 2: Every arm on the CheXpert split of 234 studies. Here words is the mean report length, ≥500w the fraction of studies whose report exceeds 500 words, rep-4 the within-report 4-gram repetition rate averaged over studies, uniq. the number of distinct reports among the 234, and dist-2 the ratio of unique bigrams to all bigrams. Appendix B.1 gives the per-arm details and Appendix A every metric on every split.

	Clinical and lexical agreement			Text statistics				
Method	RadGraph	CheXbert-5	ROUGE-L	words	≥500w	rep-4	uniq.	dist-2
Reference	—	—	—	36.1	.000	.002	204	.478
Teacher (30B, frozen)	6.90	30.33	9.95	118.2	.030	.025	207	.064
Supervised FT	17.57	46.05	23.57	28.9	.004	.007	179	.121
Supervised FT, budget-matched	17.81	45.61	22.96	29.2	.004	.007	182	.124
Plain OPD (ℓ^{opd})	7.15	35.66	5.01	455.3	.611	.403	182	.037
+ teachability gate ($g_t\ell^{\mathrm{opd}}$)	7.56	32.67	4.89	465.3	.615	.416	178	.035
+ drift gate + anchor	17.69	47.11	22.22	59.0	.034	.053	192	.073
+ teachability gate + anchor	7.22	24.02	4.89	494.9	.650	.440	185	.029
Revisiting-OPD	7.50	30.39	4.48	520.7	.688	.469	180	.029
EOPD	7.23	31.08	2.89	615.1	.897	.602	169	.027
REOPOLD	17.44	46.84	22.70	26.4	.000	.006	183	.138
GA-OPD (ours)	**19.36**	**51.23**	22.56	43.3	.013	.022	197	.101

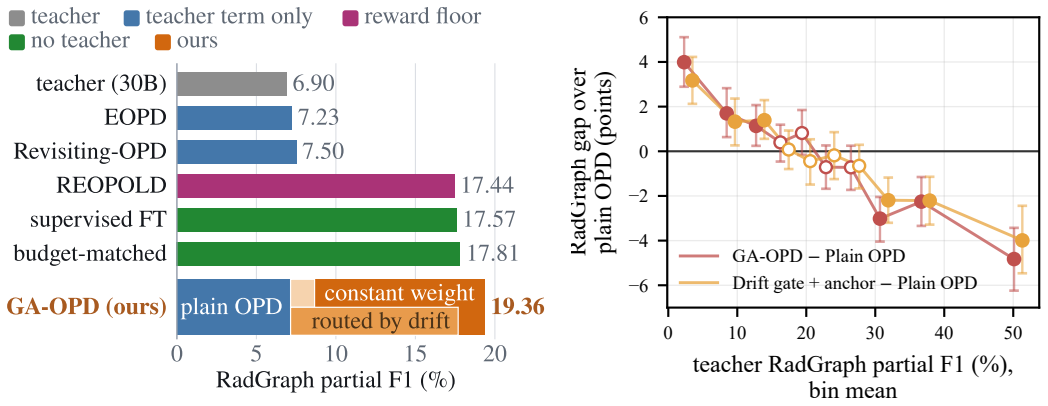
Metrics and baselines. We report BLEU-1 and ROUGE-L for surface quality, together with CheXbert micro-F1 over the 5-label and 14-label sets, RadGraph partial F1, and RaTEScore for clinical faithfulness (Smit et al., 2020; Jain et al., 2021; Zhao et al., 2024). RadGraph is a per-study score, so it carries every per-study comparison, including the difficulty axis below; CheXbert micro-F1 pools label decisions across the corpus and carries the aggregate one. We compare against plain on-policy distillation, three component ablations, and three state-of-the-art variants reproduced under our training configuration with only the loss replaced. Two supervised rows appear throughout: the checkpoint every distillation arm starts from, and a budget-matched control that continues it on references alone, with no teacher and no rollout. Revisiting-OPD, EOPD and REOPOLD denote the objectives of Fu et al. (2026), Jin et al. (2026) and Ko et al. (2026) respectively.

4.2 The Regime Where the Teacher Fails

We first isolate the studies this paper is about, using the CheXpert split, on which the frozen teacher is itself unreliable: its RadGraph of 6.90 falls far below the 17.57 of the supervised student. Table 2 and Figure 3a report it. Plain on-policy distillation and the two baselines that only reshape the teacher term all collapse to the teacher's level, their ROUGE-L falling from the 23.57 of the supervised student to between 2.89 and 5.01; REOPOLD, the one baseline carrying a reward floor, is the exception and holds. GA-OPD reaches 19.36 RadGraph, above every other arm, including the teacher and the supervised student.

The budget-matched supervised control, which sees no teacher and no rollout, reaches 17.81 here, while every arm whose teacher term carries neither a drift gate nor a reward floor scores below half of that. In this regime the teacher term is a liability of more than ten points, and the drift gate is what converts it back into an asset: routed by drift, the same objective clears the teacher-free control and reaches the best score of any arm.

That degradation has a specific form, characterized in the right half of Table 2. The teacher-only arms and the ungated-anchor ablation write 455 to 615 words against a 36-word reference, exceed 500 words on 61 to 90 percent of studies, and repeat 40 to 60 percent of their own 4-grams. This is length inflation driven by repetition, the pathology reported for on-policy distillation on reasoning benchmarks (Luo et al., 2026; Wang et al., 2026b), appearing here on a generation task where the quantity at risk is clinical faithfulness. The arms do not simply copy it from the teacher. The teacher is verbose, at 118 words, but it is not repetitive, at a 4-gram repetition rate of .03; the arms inherit its opening formula and repeat it, driving 4-gram repetition more than tenfold above the teacher's. Nor is the pathology a property of the objectives: on MIMIC-CXR, where the teacher is reliable,



(a) (b)

Figure 3: (a) Decomposition of the gain on the 234 studies where the teacher is unreliable. Color marks the mechanism, not the method: objectives with only a teacher term remain at the teacher's own level, arms with a reward floor or no teacher near the supervised student. The bottom bar overlays the arms built on a plain OPD base: the anchor at a constant weight, 7.22, the anchor routed by measured drift, 17.69, and the full objective, 19.36. (b) Ten equal-mass bins of the frozen teacher's per-sample RadGraph. Filled markers denote bins that separate from zero; Appendix B.8 gives the procedure. Both drift-gated arms trace the same curve.

every distillation arm stays between 57 and 77 words at repetition rates of .03 to .06 (Appendix A). It appears only where the teacher's output distribution is misaligned with the reference, and a better-shaped divergence over that distribution does not remove it; Appendix C.1 gives verbatim examples. Both drift-gated arms instead write at reference scale, GA-OPD at 43.3 words against 36.1 with a 4-gram repetition rate of .022, and GA-OPD produces 197 distinct reports over the 234 studies, more than any other student arm and close to the 204 of the references.

GA-OPD exceeds Revisiting-OPD by 11.86 points of RadGraph, EOPD by 12.12 and REOPOLD by 1.92, all three significant after a Benjamini-Hochberg correction over the declared comparison family (adjusted $p \leq 0.032$). Appendix B.8 gives the full interval estimates and the procedure.

The separation holds along the whole training trajectory: on the CheXpert portion of the validation set, GA-OPD stays between 20.7 and 22.1 at all ten scored checkpoints and every arm carrying a teacher term without a drift gate between 7.3 and 10.4 (Appendix B.7).

4.3 A 4B Student at the Level of Its 30B Teacher

Averaged over all 4,676 studies, GA-OPD improves on the supervised checkpoint every distillation arm starts from along all six axes at once: +1.85 BLEU-1, +0.29 ROUGE-L, +3.10 and +1.94 CheXbert micro-F1, +1.05 RadGraph and +1.47 RaTEScore (Table 1). The rest trade one axis against another: plain on-policy distillation buys 1.39 RadGraph at a cost of 0.94 CheXbert-5, and the strongest divergence-reshaping baseline buys 2.41 RadGraph at a cost of 4.81 BLEU-1. GA-OPD also reaches the highest CheXbert micro-F1 of any arm on both label sets, 59.34 and 51.81, and exceeds the frozen 30B teacher on both. On aggregate RadGraph the distillation arms fall within 2.1 points of one another, a spread produced on a test set of which 95 percent lies outside the split where the teacher fails. The teacher is a sparse mixture-of-experts model, 128 experts with 8 active per token, so parameter counts do not state the deployment benefit; the student's weights occupy 8.3 GB and the model serves on one accelerator at 8.8 GiB of device memory, against 58 GB and two accelerators at 29.4 GiB each for the teacher.

4.4 The Advantage Grows With Teacher Failure

The effect is not a property of one corpus but a gradient over a continuous axis. On the axis defined in the setup we estimate a difference-in-differences of the arm gap between the teacher-hard and the

Table 3: Difference-in-differences of RadGraph on the raw and residualized teacher-difficulty axes. The tercile column contrasts the hardest against the easiest tercile, the decile column the worst against the best decile. Entries marked † do not separate from zero; Appendix B.8 gives the intervals.

	Raw axis		Residualized	
Gap over plain OPD	terc.	dec.	terc.	dec.
In-domain, 4,676 studies				
GA-OPD	$+5.10$	$+8.84$	$+5.91$	$+9.89$
Drift gate + anchor	$+4.42$	$+7.22$	$+4.68$	$+8.14$
MIMIC-CXR, 3,852 studies				
GA-OPD	$+1.00$	$+1.29^{\dagger}$	$+1.91$	$+1.66^{\dagger}$
Drift gate + anchor	$+0.63^{\dagger}$	$-0.12^{\dagger}$	$+1.10$	$+0.05^{\dagger}$

teacher-easy end, which also removes any uniform shift from references hard for every arm (Efron & Tibshirani, 1994; Dror et al., 2020). Table 3 reports it. The advantage of GA-OPD over plain OPD grows with teacher failure, from $+1.91$ within MIMIC-CXR, where the teacher rarely fails, to $+5.91$ and $+9.89$ once the full in-domain set including CheXpert is considered. Residualizing strengthens every estimate relative to the raw axis, so the effect is not an artifact of difficulty that a teacher-free student would share; Appendix B.8 reports which contrasts separate from zero.

Figure 3b shows the same effect on ten equal-mass bins of the teacher axis: the gap over plain OPD falls from $+3.99$ points in the lowest bin to -4.83 in the highest and changes sign near the middle of the axis. At the low end this is a reversal rather than a level shift, plain OPD falling below even the supervised student while GA-OPD stays above it.

4.5 Only the Routed Anchor Recovers Faithfulness

Holding the drift gate at a constant while keeping the anchor and the teachability gate isolates the routing rule from the term it routes (Figure 3a). Against the 7.56 of the teachability gate alone, an anchor added without the drift gate scores 7.22 in the teacher-failure regime; the same anchor added with it scores 19.36. The two arms also score 24.02 against 51.23 on CheXbert-5 and 4.89 against 22.56 on ROUGE-L, so an ungated anchor leaves the length collapse of Table 2 intact. The recovery is therefore contingent on the drift gate routing the anchor in, and not on the presence of a reference term in the objective. The remaining ablations quantify the other components. Removing the anchor and the drift gate together, which leaves the teacher term under the teachability gate alone, costs 11.80 points, from 19.36 to 7.56, returning the objective to the level of plain OPD; removing the teachability gate alone costs 1.66 points, from 19.36 to 17.69, a real yet secondary component. Replacing our objective with either divergence-reshaping variant costs 11.86 and 12.12 points, so neither reshaping the teacher term nor adding the anchor ungated substitutes for the routed anchor.

The rescue also does not depend on the direction of the divergence: with plain OPD and GA-OPD retrained in the reverse direction it is 12.45 RadGraph points against 12.21 forward, and the aggregate scores of both arms agree to within 0.31 points (Appendix B.3).

5 Conclusion

We showed that the recent family of on-policy distillation methods for report generation shares a ceiling its own design cannot lift, since all of its supervision passes through the teacher. Where a medical teacher is itself unreliable, which is also where prefix drift is most damaging, methods that reshape the teacher term collapse together with plain OPD. GA-OPD adds a ground-truth anchor that bypasses the teacher, and routes that anchor by the measured drift of the student prefix: conditional insurance, decisive where the teacher cannot help and negligible where it can. Indexing difficulty by the teacher's own faithfulness shows the advantage growing with teacher failure. We read this as a design rule: where a distillation teacher is unreliable on part of the input distribution, the remedy is not a better-shaped divergence but a supervision source the teacher does not provide, together with a rule that decides when to spend it.

AI USE STATEMENT

The vision-language models trained and evaluated in this paper (the Qwen3-VL-4B student, the Hulu-Med-30A3 teacher and every baseline arm) and the models that compute the evaluation metrics (CheXbert, RadGraph and RaTEScore) are objects of study and measurement instruments, specified in Section 4.1, and are not uses of generative AI as an assistant. No generative AI tool was used to generate, annotate, clean or judge any data. Under recommended disclosure, generative AI tools were used to improve the readability of the writing. The authors reviewed all AI-assisted content and take full responsibility for the final content of this work.

ETHICS STATEMENT

All corpora are de-identified chest radiographs released for research use, MIMIC-CXR and CheXpert Plus under credentialed access with data use agreements; we attempted no re-identification. The system is a research artifact, not a clinical tool. The metrics we optimize measure agreement with a reference report and do not certify clinical safety (Yu et al., 2023; Ostmeier et al., 2024); any deployment would require prospective evaluation, radiologist oversight of every report, and monitoring for automation bias.

REPRODUCIBILITY STATEMENT

Section 3 defines the objective and both gates, and Algorithm 1 in Appendix B gives the training loop. Section 4.1 specifies the corpora, the patient-level validation split, the models, the selection rule and the metrics. Appendix B lists the full hyperparameters, the checkpoint-selection procedure, the compute environment, the prompt templates and the interval estimates, and Appendix A reports the results on every split. We will release the code upon publication.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. *arXiv preprint arXiv:2306.13649*, 2023.

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc N. M. Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking on-policy distillation: Where it helps, where it hurts, and why. *arXiv preprint arXiv:2605.10889*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pp. 1171–1179, 2015.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

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)*, 2021. doi: 10.18653/v1/2021.acl-long.459.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jenia Jitsev, Sergios Gatidis, Jean-Benoit Delbrouck, Akshay S. Chaudhari, and Curtis P. Langlotz. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360, 2022.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. doi: 10.1093/jamia/ocv080.

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. *Statistical Significance Testing for Natural Language Processing*. Springer International Publishing, 2020. ISBN 9783031021749. doi: 10.1007/978-3-031-02174-9.

Bradley Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994. ISBN 9780429246593. doi: 10.1201/9780429246593.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*, 2019.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH 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.

Li Jiang, Haoran Xu, Yichuan Ding, and Amy Zhang. Trajectory-refined distillation. *arXiv preprint arXiv:2606.08432*, 2026.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, Fudong Zhu, Kenli Li, Weidi Xie, Jimeng Sun, Jian Wu, and Zuozhu Liu. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016. doi: 10.18653/v1/D16-1139.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. *arXiv preprint arXiv:2309.06180*, 2023.

Kaiyuan Liu, Ziyuan Zhuang, Yang Bai, Bing Wang, Rongxiang Weng, and Jieping Ye. Prefix teach, suffix fade: Local teachability collapse in strong-to-weak on-policy distillation. *arXiv preprint arXiv:2605.13643*, 2026a.

Yanjiang Liu, Jie Lou, Xinyan Guan, Yuqiu Ji, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Xing Yu, and Yaojie Lu. Your teacher can't help you here: Combating supervision fidelity decay in on-policy distillation. *arXiv preprint arXiv:2605.30833*, 2026b.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying opd: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026.

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*, 2020.

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*, 2024. doi: 10.18653/v1/2024.findings-emnlp.21.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2016.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297. ACM, 2025. doi: 10.1145/3689031.3696075.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52729.2023.00718.

Jianze Wang, Ying Liu, Jinlong Chen, Xuchun Hu, Qilong Zhang, Yu Cao, Jun Wang, Hua Yang, Yong Xie, and Qianglong Chen. MAD-OPD: Breaking the ceiling in on-policy distillation via multi-agent debate. *arXiv preprint arXiv:2605.01347*, 2026a.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations. *arXiv preprint arXiv:2607.13399*, 2026b.

Yuan Wang, Shujian Gao, Jiaxiang Liu, Songtao Jiang, Haoxiang Xia, Xiaotian Zhang, Zhaolu Kang, Yemin Wang, and Zuozhu Liu. Beyond n-grams: A hierarchical reward learning framework for clinically-aware medical report generation, 2025. URL https://arxiv.org/abs/2512.02710.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. doi: 10.18653/v1/2023.acl-long.605.

Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology report generation via multi-objective preference optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8664–8672, 2025.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. Kdrl: Post-training reasoning llms via unified knowledge distillation and reinforcement learning. *arXiv preprint arXiv:2506.02208*, 2025.

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*, 2023. doi: 10.1016/j.patter.2023.100802.

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. Towards a clinically accessible radiology foundation model: Open-access and lightweight, with automated evaluation. *arXiv preprint arXiv:2403.08002*, 2024.

Bowen Zhang. A formula-driven survey and research agenda for on-policy distillation. *arXiv preprint arXiv:2606.22793*, 2026.

Kai Zhang, Christopher Malon, Lichao Sun, and Martin Renqiang Min. Editgrpo: Reinforcement learning with post-rollout edits for clinically accurate chest x-ray report generation. *arXiv preprint arXiv:2509.22812*, 2025a.

Songming Zhang, Xue Zhang, Tong Zhang, Bojie Hu, Yufeng Chen, and Jinan Xu. Aligndistil: Token-level language model alignment as adaptive policy distillation. *arXiv preprint arXiv:2503.02832*, 2025b.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845*, 2024.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. Scope: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *arXiv preprint arXiv:2604.10688*, 2026.

Table 4: MIMIC-CXR split, 3,852 studies. CheXbert-5 and CheXbert-14 are micro-F1 over the 5-label and 14-label sets; RadGraph is partial F1.

	Clinical and lexical agreement						Text statistics				
Method	BLEU-1	ROUGE-L	CheXbert-5	CheXbert-14	RadGraph	RaTEScore	words	≥500w	rep-4	uniq.	dist-2
Reference	—	—	—	—	—	—	81.6	.000	.012	2341	.099
Zero-shot	7.99	9.97	41.01	38.76	10.01	50.82	394.7	.099	.055	3852	.057
Supervised FT	29.55	24.46	57.07	49.33	21.93	55.54	58.7	.006	.026	2634	.018
Supervised FT, budget-matched	29.40	24.60	56.97	49.34	21.99	55.75	57.3	.004	.025	2634	.020
Plain OPD	30.74	24.74	56.64	49.39	22.72	56.72	63.4	.012	.037	2769	.017
+ teachability gate	31.29	24.68	56.34	48.55	23.18	57.27	63.4	.010	.036	2479	.015
+ drift gate + anchor	31.95	24.64	57.23	49.72	22.86	57.24	73.9	.021	.056	2810	.018
+ teachability gate + anchor	31.33	24.70	57.47	49.46	22.80	57.02	68.4	.018	.043	2697	.017
Revisiting-OPD	32.15	24.21	56.84	48.55	22.77	56.83	73.2	.021	.051	2676	.013
EOPD	33.59	24.82	58.56	50.76	24.13	57.52	76.6	.024	.056	2905	.013
REOPOLD	32.88	24.67	58.97	51.24	22.66	56.06	72.0	.017	.043	2900	.016
GA-OPD	31.18	24.63	60.07	51.53	22.78	56.82	61.7	.006	.032	2888	.019
Teacher (30B)	32.98	25.25	58.88	52.35	22.93	56.68	61.1	.001	.016	3213	.044

Table 5: IU-Xray split, 590 studies. CheXbert-5 and CheXbert-14 are micro-F1 over the 5-label and 14-label sets; RadGraph is partial F1.

	Clinical and lexical agreement						Text statistics				
Method	BLEU-1	ROUGE-L	CheXbert-5	CheXbert-14	RadGraph	RaTEScore	words	≥500w	rep-4	uniq.	dist-2
Reference	—	—	—	—	—	—	28.2	.000	.001	398	.224
Zero-shot	4.39	7.70	12.41	22.31	12.62	55.66	326.6	.015	.043	590	.106
Supervised FT	29.64	28.82	36.07	60.33	26.16	60.51	29.0	.002	.005	88	.040
Supervised FT, budget-matched	29.63	28.36	36.26	60.03	25.87	59.77	28.3	.002	.005	89	.043
Plain OPD	31.35	33.25	35.05	59.12	36.25	69.97	32.8	.003	.007	89	.031
+ teachability gate	29.53	32.10	34.17	59.49	34.80	68.96	38.1	.000	.006	119	.028
+ drift gate + anchor	27.04	29.48	39.22	49.17	28.19	63.71	35.4	.003	.008	128	.034
+ teachability gate + anchor	28.77	29.22	31.87	60.10	27.16	61.71	30.4	.002	.006	114	.037
Revisiting-OPD	28.89	29.91	32.84	58.23	30.80	66.54	33.9	.003	.009	126	.036
EOPD	28.41	32.94	42.11	60.99	35.03	69.23	38.9	.005	.010	125	.029
REOPOLD	25.28	29.19	40.00	59.98	28.35	61.93	37.4	.007	.011	120	.032
GA-OPD	30.45	30.50	42.48	58.40	28.26	63.49	31.3	.000	.005	123	.042
Teacher (30B)	33.94	30.75	40.00	61.38	31.29	65.59	26.5	.000	.001	191	.059

Liwen Zheng and Haiyun Jiang. Blockwise policy-drift gating for on-policy distillation. *arXiv preprint arXiv:2606.24084*, 2026.

A Per-Split Results

Tables 4, 5 and 6 give every arm on each split: the six metrics of the main paper as percentages, and five statistics of the generated text. Words is the mean report length, ≥500w the fraction of studies whose report exceeds 500 words, rep-4 the within-report 4-gram repetition rate averaged over studies, uniq. the number of distinct reports, and dist-2 the corpus-level ratio of unique bigrams to all bigrams; lengths are whitespace-delimited words on the generated text. The three splits separate cleanly. On MIMIC-CXR and IU-Xray, where the frozen teacher's output distribution matches the reference genre, every arm but the zero-shot student writes between 26 and 77 words, and every arm repeats at most 6 percent of its 4-grams. On CheXpert, where it does not, every arm whose teacher term carries neither a drift gate nor a reward floor writes 455 to 615 words and repeats 40 to 60 percent of its 4-grams, while the arms carrying a drift-gated anchor, the reward-floor baseline and the two supervised controls stay at reference length. GA-OPD is also the only distillation arm whose bigram diversity on MIMIC-CXR exceeds that of the supervised checkpoint it starts from. Text statistics for the zero-shot student on CheXpert were not retained.

B Additional Training and Evaluation Details

Table 6: CheXpert split, 234 studies. This is the regime analyzed in the main paper: the teacher scores below the supervised student on every axis.

	Clinical and lexical agreement						Text statistics				
Method	BLEU-1	ROUGE-L	CheXbert-5	CheXbert-14	RadGraph	RaTEScore	words	≥500w	rep-4	uniq.	dist-2
Reference	—	—	—	—	—	—	36.1	.000	.002	204	.478
Zero-shot	2.60	4.54	21.38	27.42	4.67	44.61	—	—	—	—	—
Supervised FT	21.68	23.57	46.05	48.00	17.57	52.04	28.9	.004	.007	179	.121
Supervised FT, budget-matched	21.37	22.96	45.61	47.47	17.81	51.67	29.2	.004	.007	182	.124
Plain OPD	2.01	5.01	35.66	27.66	7.15	39.81	455.3	.611	.403	182	.037
+ teachability gate	1.97	4.89	32.67	28.42	7.56	40.06	465.3	.615	.416	178	.035
+ drift gate + anchor	15.79	22.22	47.11	49.45	17.69	51.97	59.0	.034	.053	192	.073
+ teachability gate + anchor	1.85	4.89	24.02	22.37	7.22	38.34	494.9	.650	.440	185	.029
Revisiting-OPD	1.77	4.48	30.39	26.43	7.50	39.78	520.7	.688	.469	180	.029
EOPD	1.35	2.89	31.08	28.79	7.23	39.05	615.1	.897	.602	169	.027
REOPOLD	20.86	22.70	46.84	48.40	17.44	51.67	26.4	.000	.006	183	.138
GA-OPD	20.92	22.56	51.23	49.83	19.36	52.78	43.3	.013	.022	197	.101
Teacher (30B)	7.69	9.95	30.33	16.16	6.90	38.15	118.2	.030	.025	207	.064

Algorithm 1 One GA-OPD optimization step

input studies $\{x^{(i)}\}$ with references $\{r^{(i)}\}$, student π_S, frozen teacher π_T, top-k, anchor weight β, scale s
for each study i in the batch **do**
  sample rollout $y^{(i)} \sim \pi_S(\cdot \mid x^{(i)})$
  maintain the running bag-of-tokens intersection of $y^{(i)}_{\le t}$ and $r^{(i)}_{\le \min(t,M)}$; set α_t by Eq. 4
  emit a teacher-forced sample with input $(x^{(i)}, r^{(i)})$ and flag $\texttt{is_sft} = 1$; flag the rollout with $\texttt{is_sft} = 0$
end for
keep the first N of the $2N$ emitted samples
score rollout tokens under π_T, keeping its top-k probabilities; set g_t by Eq. 7
per token: $\ell_t = (1 - \texttt{is_sft})\, \alpha_t g_t\, \ell_t^{\mathrm{opd}} + \texttt{is_sft}\, \beta\, (-\log \pi_S(r_t \mid x, r_{<t}))$
average ℓ_t over the response mask of the batch and take one optimizer step on π_S

B.1 Checkpoint Selection

Every distillation arm stores a checkpoint every 300 steps, the two-phase baseline additionally at its phase switch, step 20, and supervised fine-tuning at the end of its single epoch, step 1,807; every stored checkpoint is evaluated on the held-out validation set of 3,156 studies. The selected checkpoint is the peak of validation RadGraph. The selected steps are 1,807 for supervised fine-tuning, 300 for the budget-matched supervised control, 1,800 for plain OPD, 1,200 for the teachability-gate ablation, 2,400 for both the drift-gate-plus-anchor and the teachability-gate-plus-anchor ablations, 300 for GA-OPD, 1,500 for Revisiting-OPD and EOPD, and 20 for REOPOLD.

REOPOLD, which uses a two-phase schedule, entered an absorbing state during its second phase: rollout perplexity fell to one and remained there, mean response length dropped from 96 tokens to about 6, and the loss became constant at the negative of its reward floor. All of its second-phase checkpoints score far below its first-phase checkpoint on validation, so the selection rule places its best checkpoint at step 20. We report this because it affects how that baseline should be read, not because we consider the run invalid: the collapse is a property of the published objective under our configuration.

B.2 Alternative Checkpoints

Revisiting-OPD and EOPD peak on validation CheXbert at a different step than on validation RadGraph. Evaluated at those alternative steps they reach 58.54 and 58.18 CheXbert micro-F1 with 23.40 and 23.02 RadGraph, which does not change any conclusion of the main paper.

B.3 Direction of the Divergence

The aggregate RadGraph of the two reverse-direction runs is 23.63 for the plain method against 23.64 in the forward direction, and 23.61 for GA-OPD against 23.30. On the CheXpert split the gap between the two arms is 12.45 reversed against 12.21 forward. The two runs use different estimators: the plain arm uses a sampled estimator, while the anchored arm uses a normalized top-k form. Only these two arms were retrained in the reverse direction, and the per-sample stratified analysis was not recomputed for them, so the main paper restricts that comparison to the regime-level rescue.

B.4 Number of Runs and Uncertainty

Each arm corresponds to a single training run with a fixed seed, and no arm was rerun under a second seed. The patient-level split of the training pool into training and validation studies uses seed 42. Training keeps the framework defaults, which set both the sharded-optimizer seed and the data-loader seed to 42, and no arm overrides them. Evaluation carries no randomness, since decoding is greedy at temperature zero. All reported confidence intervals are paired bootstrap intervals over test studies with 2,000 resamples drawn under seed 0, and therefore quantify sampling uncertainty of the evaluation set rather than variance across training seeds. We state this explicitly because the two are often conflated.

B.5 Compute Environment

Experiments ran on a single node running Ubuntu 22.04.5 with kernel 6.8, equipped with Intel Xeon Platinum 8558 processors totaling 192 logical cores and eight NVIDIA H20-3e accelerators of 141 GB each, using PyTorch 2.11 with CUDA 13.0, a vLLM 0.23 inference backend for rollout generation and teacher scoring, and FlashAttention 2.8. Rollout and teacher scoring occupied four devices each.

B.6 Full Hyperparameters

All distillation arms: batch of 32 sequences per optimization step, actor learning rate 1×10^{-6}, 3,000 optimization steps, maximum prompt length 5,632 tokens, maximum response length 512 tokens, teacher top-k of 64, anchor weight $\beta = 1$, teachability scale $s = 0.3$, bf16 precision with fused attention and padding removal, four actor and four teacher devices. The optimizer is AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay 0.01, at a constant learning rate with no warmup, and gradients are clipped at global norm 1.0. All parameters are trained; no adapter or low-rank tuning is used. Training rollouts are sampled at temperature 1.0 with one rollout per prompt; the per-token loss is clamped at 10.0 and teacher log-probabilities at -10.0 for numerical stability, and losses are aggregated by a token mean over the response mask. Supervised fine-tuning: batch size 256, learning rate 2×10^{-5}, one epoch, warmup ratio 0.03, weight decay 0.01. Budget-matched supervised control: same initialization as every distillation arm, batch of 32 sequences per optimization step, learning rate 1×10^{-6}, 3,000 optimization steps, loss a teacher-forced cross-entropy on the reference report, with no teacher process and no rollout. Evaluation: greedy decoding, temperature zero, at most 1,024 generated tokens, per-image pixel budget taken from the preprocessed inputs.

Clinical metrics are computed with fixed implementations at pinned versions: `f1chexbert` 0.0.2 for CheXbert micro-F1, `radgraph` 0.1.18 for RadGraph partial F1, and `RaTEScore` 0.6.0.

B.7 Per-Step Validation Curves

Every stored checkpoint of every arm was scored on the full held-out validation set of 3,156 studies, of which 1,001 come from CheXpert, 2,005 from MIMIC-CXR and 150 from IU-Xray. Table 7 lists validation RadGraph on the CheXpert portion at each step, and Figure 4 plots that portion alongside the full set. GA-OPD peaks on the full validation set at its first stored checkpoint, step 300, at 24.72, and no later checkpoint exceeds it. On the CheXpert portion the reading is stable across all ten steps: the two drift-gated arms stay between 20.7 and 22.1, every arm that carries a teacher term without a drift gate stays between 7.3 and 10.4, and the budget-matched supervised control, which carries no teacher term at all, stays between 20.3 and 20.9. Adding a teacher term costs about twelve points on this portion at every matched step, the ungated anchor of the teachability-gate-plus-anchor arm does

Table 7: Validation RadGraph partial F1 on the CheXpert portion of the validation set ($n = 1{,}001$) at every stored checkpoint. REOPOLD is omitted because its selection falls at step 20; its values are 20.20 at step 20 and 2.27, 0.53 and 0.01 at steps 300, 600 and 900.

Method	300	600	900	1200	1500	1800	2100	2400	2700	3000
GA-OPD	21.55	22.07	21.24	21.69	21.31	20.73	20.76	21.21	22.03	21.14
drift gate + anchor	21.32	21.63	21.38	21.71	21.02	20.99	20.94	20.92	22.03	21.48
Supervised, budget-matched	20.42	20.54	20.68	20.69	20.30	20.59	20.86	20.49	20.31	20.59
Plain OPD	8.69	8.49	8.30	8.09	8.20	8.52	8.38	8.26	8.77	8.74
+ teachability gate	9.17	8.79	8.81	8.68	8.36	8.35	8.55	7.59	7.64	8.14
+ teachability gate + anchor	9.50	7.72	7.59	7.40	8.29	8.76	8.20	8.31	7.39	8.34
Revisiting-OPD	9.52	8.97	8.42	8.14	8.68	8.85	8.76	8.28	8.18	8.22
EOPD	10.34	8.44	8.25	7.61	8.49	8.60	8.61	8.58	8.41	8.65

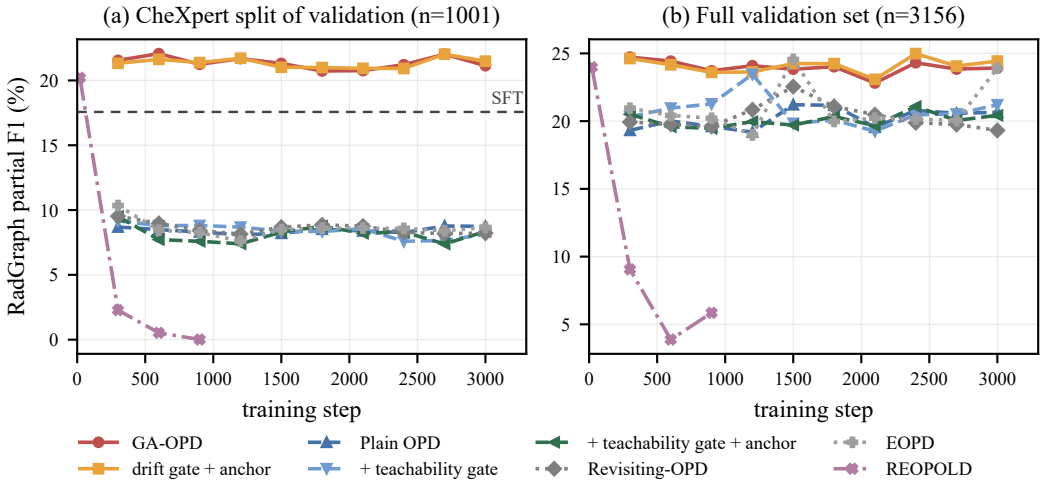


Figure 4: Validation RadGraph against training step, on the CheXpert portion of the validation set (a) and on the full set (b). The two drift-gated arms hold their level from the first stored checkpoint; every arm carrying an ungated teacher term sits about twelve points below them on (a) at every step.

not recover them, and only the drift-gated arms return to and exceed the level reached without any teacher.

B.8 Interval Estimates and Comparison Family

Table 8 gives 95% paired bootstrap intervals with 2,000 resamples for every arm on the load-bearing metric and on CheXbert micro-F1 over the 5-label set, on the full in-domain test set and on the CheXpert split. The declared comparison family is the seven contrasts of GA-OPD against every other arm on the CheXpert split together with the same seven on the aggregate. Under a Benjamini-Hochberg correction over those fourteen, the seven CheXpert contrasts have adjusted p-values between 0.001 and 0.032, and on the aggregate the contrasts against the supervised student, the teachability-gate arm and EOPD have adjusted p-values of 0.001 while the remainder are above 0.04. The family was fixed before the teachability-gate-plus-anchor ablation and the budget-matched supervised control were run, and we do not enlarge it after the fact: those two arms are reported with point intervals only, computed under the same procedure of 2,000 paired resamples under seed 0, and the adjusted p-values above are unchanged.

The margins quoted in the main text carry the following intervals, all from that same procedure and computed from unrounded per-sample scores, so a margin can differ by 0.01 from the difference of the rounded values printed in the tables. On the CheXpert split GA-OPD exceeds Revisiting-OPD by 11.86 points of RadGraph ($[10.10,\ 13.61]$), EOPD by 12.12 ($[10.32,\ 13.99]$) and REOPOLD by 1.92

Table 8: Point estimates with 95% paired bootstrap intervals over test studies, 2,000 resamples under seed 0. These intervals quantify evaluation-set sampling and not variance across training seeds.

	In-domain (4676)		CheXpert (234)	
Method	RadGraph	CheXbert-5	RadGraph	CheXbert-5
Supervised FT	22.25 [21.88, 22.59]	56.24 [54.98, 57.38]	17.57 [15.84, 19.35]	46.05 [40.62, 51.67]
Supervised FT, budget-matched	22.27 [21.92, 22.61]	56.13 [54.86, 57.28]	17.81 [16.06, 19.65]	45.61 [39.91, 50.92]
Plain OPD	23.64 [23.25, 24.02]	55.30 [54.09, 56.47]	7.15 [6.39, 8.00]	35.66 [30.13, 40.70]
+ teachability gate	23.87 [23.49, 24.25]	54.81 [53.59, 55.90]	7.56 [6.68, 8.48]	32.67 [27.25, 37.96]
+ drift gate + anchor	23.28 [22.90, 23.62]	56.45 [55.28, 57.49]	17.69 [16.03, 19.35]	47.11 [40.60, 53.01]
+ teachability gate + anchor	22.57 [22.19, 22.93]	55.92 [54.74, 57.12]	7.22 [6.40, 8.07]	24.02 [17.92, 30.32]
Revisiting-OPD	23.02 [22.65, 23.39]	55.32 [54.17, 56.59]	7.50 [6.65, 8.35]	30.39 [24.60, 36.20]
EOPD	24.66 [24.26, 25.05]	57.18 [56.05, 58.30]	7.23 [6.38, 8.09]	31.08 [25.40, 36.72]
REOPOLD	23.12 [22.77, 23.49]	58.13 [56.95, 59.24]	17.44 [15.66, 19.32]	46.84 [41.04, 52.23]
GA-OPD	23.30 [22.96, 23.65]	59.34 [58.31, 60.41]	19.36 [17.56, 21.09]	51.23 [45.64, 56.94]

($[0.57,\ 3.36]$), none of the three intervals containing zero. Within the component ablation, removing the teachability gate costs 1.66 ($[0.20,\ 3.07]$) and removing the anchor together with the drift gate costs 11.80 ($[10.04,\ 13.63]$). The difference-in-differences of the main paper use the same paired bootstrap; entries marked there with a dagger have 95% intervals containing zero and all others exclude it, and residualizing the axis moves two MIMIC-CXR contrasts from intervals containing zero to intervals excluding it.

B.9 Prompt Templates

Prompts are taken verbatim from the source corpora rather than written by us: preprocessing copies the user turn of each source record unchanged. Each prompt is preceded by one `<image>` placeholder per image in the study.

MIMIC-CXR, IU-Xray and CheXpert draw on a shared pool of eleven paraphrases of one request, and the pool is identical across the training, validation and test splits of all corpora. A representative instance is

> Could you compile a medical report that summarizes the findings from this medical image?

Public chest X-ray corpora do not provide a split on which a strong medical teacher fails. On the CheXpert split we therefore replace the instruction the teacher receives with a different task specification, so that its output distribution is misaligned with the reference reports:

> Based on the given sequence of medical images, outline the key abnormalities, indicating the anatomical regions affected, any abnormal structures, densities, or masses, and propose potential diagnoses according to these observations.

The teacher receives this instruction both when it scores rollouts in training and when it is evaluated as an arm. Every other arm receives the corpus instruction unchanged in training and in evaluation, and no other split is affected.

The consequence is visible in the outputs. The frozen teacher answers the instruction, emitting an enumeration of abnormalities followed by candidate diagnoses on 232 of the 234 CheXpert studies and on none of the 3,852 MIMIC-CXR or 590 IU-Xray studies. On the CheXpert split the teacher's output distribution is therefore misaligned with the reference while the students' is not.

C Case Analysis

C.1 On the Difficulty Axis

Three studies illustrate the teacher-difficulty axis of Section 4, with long outputs truncated at 55 words and their full length stated. The easy study sits near the top of the axis: the frozen teacher scores 73.7 RadGraph on it, writes a faithful report, and every arm agrees. The two hard studies sit

near the bottom, where the teacher scores 17.8 and 12.8 RadGraph, hallucinates a tumor, and the teacher-only arm copies the hallucination.

Easy study, teacher RadGraph 73.7.

Prompt. Could you create a clinical report summarizing the key outcomes from this medical image?



MIMIC-CXR, 2 views

reference report 33 words

PA and lateral views of the chest pain . The lungs are clear . Cardiomediastinal silhouette is normal . No acute osseous abnormalities detected . Stent is identified in the upper abdomen .

frozen teacher (30B) **17 words** · RadGraph 73.7

The lungs are clear . The cardiomediastinal silhouette is normal . No acute osseous abnormalities identified .

plain OPD **17 words** · RadGraph 73.7

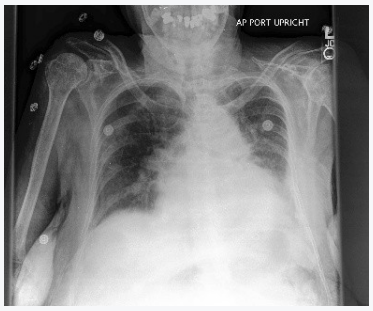
The lungs are clear . The cardiomediastinal silhouette is normal . No acute osseous abnormalities identified .

GA-OPD (ours) **18 words** · RadGraph 57.1

The lungs are clear . The cardiomediastinal silhouette is within normal limits . No acute osseous abnormalities .

Hard study, teacher RadGraph 17.8.

Prompt. Could you compile a medical report that summarizes the findings from this medical image?



CheXpert, 1 view

reference report 24 words

Portable upright ap and lateral views of the chest demonstrate low lung volumes. Slightly improved pulmonary edema. Continued retrocardiac opacity and left pleural effusion.

frozen teacher (30B) **87 words** · RadGraph 17.8

The key abnormalities observed in the images include: 1) A large, irregular mass in the left lower lobe of the lung, which is likely a primary lung tumor. 2) Multiple smaller nodules in the right lung, which could represent metastatic disease. 3) Enlarged lymph nodes in the mediastinum, suggesting lymphadenopathy. 4) A small pleural effusion [...]

plain OPD **730 words** · RadGraph 21.1

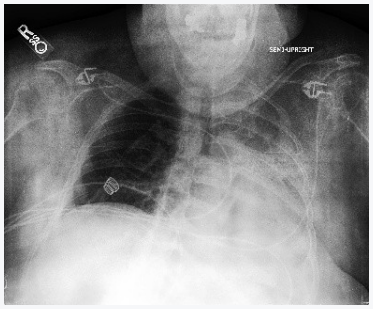
The key abnormalities observed in the medical images include: 1) A large mass in the left lower lobe of the lung, which is causing a shift of the mediastinum to the right. This mass is likely malignant given its size and location. 2) A small amount of pleural effusion in the left lower lobe, which [...]

GA-OPD (ours) **47 words** · RadGraph 42.9

Interval removal of left upper extremity PICC line. No pneumothorax. Interval increase in left pleural effusion. Interval increase in left lower lobe opacity, which may represent atelectasis or consolidation. Interval increase in right lower lobe opacity, which may represent atelectasis or consolidation. Interval increase in pulmonary edema.

Hard study, teacher RadGraph 12.8.

Prompt. Could you assemble a medical report to summarize the results captured in this image?



CheXpert, 1 view

reference report 41 words

No lines and tubes. Low lung volumes with persistent near complete opacification of the left hemithorax consistent with large pleural effusion and associated atelectasis. The faint right mid lung opacity seen on the prior exam is not visualized on today's study.

frozen teacher (30B)

110 words · RadGraph 12.8

The key abnormalities observed in the images include: 1) A large, irregular mass in the left lung, which is likely a primary lung tumor. 2) Enlarged lymph nodes in the mediastinum, suggesting lymphadenopathy. 3) Multiple small nodules in both lungs, indicative of metastatic disease. 4) A large mass in the liver, which could be a [...]

plain OPD **729 words** · RadGraph 16.2

The key abnormalities observed in the medical images include: 1) A large, dense mass in the left lung, which is likely a tumor, given its size and density. 2) A smaller, less dense mass in the right lung, which could be a metastatic lesion or a separate primary tumor. 3) Multiple small, round opacities in [...]

GA-OPD (ours) **46 words** · RadGraph 21.1

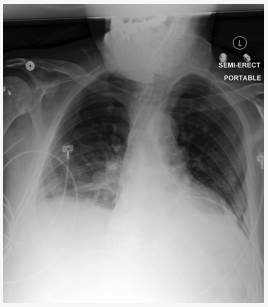
Interval removal of the left chest tube. No pneumothorax. Stable position of the right internal jugular central venous catheter. Stable appearance of the left upper extremity PICC line. Stable appearance of the left pleural effusion and left lower lobe atelectasis. Stable appearance of the right lung.

C.2 Prefix Drift

The examples above concern an unreliable teacher. The failure mode the drift gate addresses is a different one, and it is visible on MIMIC-CXR, where the teacher is reliable: the student reproduces the opening of the reference verbatim and then continues into content the reference does not contain. Both rollouts below are from plain OPD. For each study we give the drift gate α_t of Eq. 4 of the main paper, computed on the student tokenizer, which is the quantity the objective uses.

Study 1.

Prompt. Can you draft a medical summary based on the results of this medical image?



MIMIC-CXR, 1 view

reference report 103 words

As compared to the previous radiograph , the right PICC line has been pulled back . The line projects over the axillary vein . The newly placed Dobbhoff tube is curled in the pharynx . Both devices need to be repositioned . Borderline size of the cardiac silhouette . Partial left lower lobe atelectasis . [...]

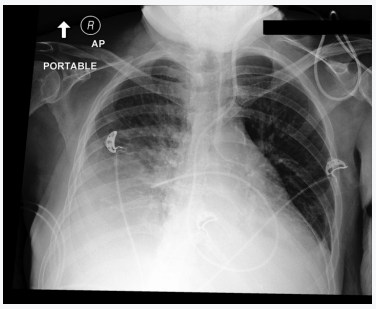
plain OPD **53 words** · RadGraph 5.3

As compared to the previous radiograph , the patient has been extubated and the nasogastric tube has been removed . The lung volumes have decreased , likely reflecting a decrease in ventilatory pressure . The pre-existing bilateral parenchymal opacities are unchanged in extent and severity . No new opacities . No pneumothorax .

The first nine tokens are identical to the reference and $\alpha_t = 1$ across them. The rollout then diverges: $\alpha_{10} = 0.90$, $\alpha_{15} = 0.73$, $\alpha_{20} = 0.55$, and $\alpha_{71} = 0.30$ at the end. The reference reports a retracted PICC line and a Dobbhoff tube curled in the pharynx, both needing repositioning; the rollout instead describes extubation and nasogastric tube removal, and every device finding of the reference is lost.

Study 2.

Prompt. Would you mind preparing a medical report that highlights the results of this diagnostic image?



MIMIC-CXR, 1 view

reference report 92 words

As compared to the previous radiograph , no additional line or monitoring devices visible on the current examination . The pre-existing Swan-Ganz catheter is in unchanged position . As on the previous report , it is noted that the the device needs to be pulled back by approximately 4 cm , as it is located [...]

plain OPD **55 words** · RadGraph 33.3

As compared to the previous radiograph , the patient has received a left-sided chest tube . The extent of the left pleural effusion has decreased . The right pleural effusion is unchanged . The lung volumes are low . The pre-existing parenchymal opacities are constant in appearance . Constant size of the cardiac silhouette .

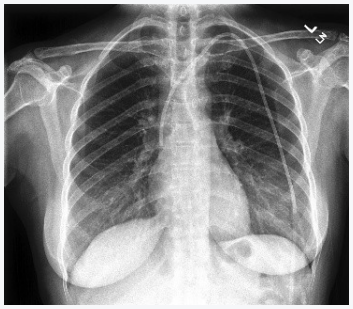
The first eight tokens are identical to the reference and $\alpha_t = 1$ across them; the gate then falls to $\alpha_{15} = 0.53$, $\alpha_{20} = 0.50$ and $\alpha_{67} = 0.30$. The reference states that no additional line is present and that the existing Swan-Ganz catheter must be withdrawn by about 4 cm; the rollout instead introduces a newly placed left-sided chest tube that the study does not contain, and the repositioning indication is absent.

C.3 Divergence-Reshaping Baselines Where the Teacher Fails

The two studies below are from the CheXpert split of Section 4.2. Among the 130 CheXpert test studies with a reference of 24 to 60 words that appear nowhere else in this paper, we rank each study by the RadGraph margin of GA-OPD over the best of plain OPD, Revisiting-OPD, EOPD and REOPOLD, and show two of the eight highest-ranked. Outputs longer than 55 words are truncated at 55 words and their full length is stated.

Study A, teacher RadGraph 0.0.

Prompt. Can you produce a summary document detailing the findings from this medical image?



CheXpert, 1 view

reference report 40 words

Interval removal of right-sided mediport and placement of a tunneled left ij with tip at the cavoatrial junction. No pneumothorax. Normal cardiomediastinal silhouette and pulmonary vasculature. No focal pulmonary opacities are seen. Pleural effusions are absent. Osseous structures are unchanged.

frozen teacher (30B) **95 words** · RadGraph 0.0

The key abnormalities observed in the medical images include: 1) A large, irregular mass in the right lung, which is likely a primary lung tumor. 2) Multiple smaller nodules in the left lung, which could represent metastatic disease. 3) Enlarged lymph nodes in the mediastinum, suggesting lymphatic spread of the disease. 4) A small pleural [...]

plain OPD **643 words** · RadGraph 0.0

The key abnormalities observed in the medical images include: 1) A 1.5 cm nodule in the right upper lobe, which is concerning for a primary lung malignancy. 2) A 1.2 cm nodule in the left lower lobe, which could represent a metastatic lesion or a separate primary tumor. 3) A 1.0 cm nodule in the [...]

Revisiting-OPD **682 words** · RadGraph 0.0

The key abnormalities observed in the medical images include: 1) A 1.5 cm nodule in the right upper lobe, which is a potential sign of a primary lung cancer or metastatic disease. 2) A 1.2 cm nodule in the left lower lobe, which could also indicate a primary lung cancer or metastatic lesion. 3) A [...]

EOPD **659 words** · RadGraph 0.0

The key abnormalities observed in the medical images include: 1) A 1.5 cm nodule in the right upper lobe, which is concerning for a primary lung malignancy. 2) A 1.2 cm nodule in the left lower lobe, which could also be a primary lung malignancy or a metastatic lesion. 3) A 1.0 cm nodule in [...]

REOPOLD **18 words** · RadGraph 12.5

Left subclavian central venous catheter with tip in the superior vena cava. No pneumothorax. No acute cardiopulmonary disease.

budget-matched SFT

16 words · RadGraph 12.9

Left subclavian central venous catheter with tip in the superior vena cava. No pneumothorax. Clear lungs.

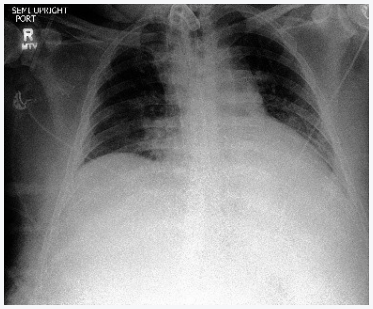
GA-OPD (ours) **30 words** · RadGraph 28.6

Stable position of left tunneled central venous catheter with tip overlying the cavoatrial junction. Normal heart size and pulmonary vascularity. No focal consolidation, pleural effusion, or pneumothorax. Bones are unremarkable.

The frozen teacher and the three arms that carry its term describe lung masses or nodules that the reference does not report, and each scores 0.0 RadGraph. REOPOLD and the budget-matched control name a left subclavian catheter in the superior vena cava, where the reference reports a tunneled left internal jugular line ending at the cavoatrial junction; REOPOLD scores 12.5. GA-OPD names a left tunneled central venous catheter at the cavoatrial junction and reports the normal heart, vasculature, lungs and pleura, and scores 28.6.

Study B, teacher RadGraph 12.5.

Prompt. Can you develop a medical summary of the findings from this diagnostic scan/image?



CheXpert, 1 view

reference report 27 words

Low lung volumes. Slight improvement in the aeration of the upper lobes bilaterally. Stable cardiomediastinal silhouette. Bibasilar consolidation, left greater than right, with small left pleural effusion.

frozen teacher (30B) **94 words** · RadGraph 12.5

The key abnormalities observed in the images include: 1) A large, irregular mass in the left lower lobe of the lung, which is likely a primary lung tumor. 2) Multiple smaller nodules in the right lung, which could represent metastatic disease. 3) Enlarged lymph nodes in the mediastinum, suggesting lymphadenopathy. 4) A small pleural effusion [...]

plain OPD **709 words** · RadGraph 14.3

The key abnormalities observed in the medical images include: 1) A large mass in the left lower lobe of the lung, which is causing compression of the left lower lobe bronchus. This mass is likely malignant given its size and location. 2) A small amount of pleural effusion in the left lower lobe, which is [...]

Revisiting-OPD **724 words** · RadGraph 7.7

The key abnormalities observed in the medical images include: 1) A large, irregular mass in the left upper lobe of the lung, which is likely a primary lung tumor. 2) A smaller, well-defined mass in the right lower lobe, which could be a secondary metastatic lesion. 3) A small, round opacity in the right upper [...]

EOPD **624 words** · RadGraph 7.8

The key abnormalities observed in the medical images include: 1) A left-sided pleural effusion, which is evident from the blunting of the left costophrenic angle and increased opacity in the left lower lung field. 2) A left-sided pneumothorax, indicated by the presence of air in the pleural space causing partial lung collapse. 3) A left-sided [...]

REOPOLD **26 words** · RadGraph 19.4

Interval placement of a TRACHEOSTOMY tube. Low lung volumes with bibasilar opacities, left greater than right, which may represent atelectasis or consolidation. Small left pleural effusion.

budget-matched SFT

26 words · RadGraph 19.4

Interval placement of a TRACHEOSTOMY tube. Low lung volumes with bibasilar opacities, left greater than right, which may represent atelectasis or consolidation. Small left pleural effusion.

GA-OPD (ours) **33 words** · RadGraph 39.0

TRACHEOSTOMY tube and right upper extremity PICC line are unchanged. Low lung volumes with bibasilar opacities, left greater than right, which may represent atelectasis or consolidation. Small left pleural effusion. Stable cardiomediastinal silhouette.

The frozen teacher, plain OPD and Revisiting-OPD report a lung mass, and EOPD a pneumothorax and a subclavian catheter; none of these is in the reference, and the four arms score 7.7 to 14.3 RadGraph. REOPOLD and the budget-matched control return the same 26 words, which match the bibasilar opacities and the left pleural effusion and add a tracheostomy tube placement; REOPOLD scores 19.4. GA-OPD reports the same findings together with the stable cardiomediastinal silhouette, names a tracheostomy tube and a PICC line that the reference does not mention, and scores 39.0.

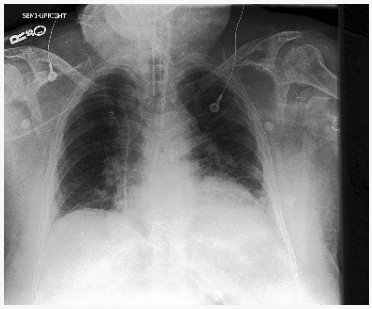
C.4 Component Ablation

The ablation of Section 4.5 compares GA-OPD with two arms that lack the drift gate. The study below is selected by the same rule, with the RadGraph margin of GA-OPD taken over the better

of OPD + teachability gate and OPD + teachability gate + anchor, and is one of the eight highest-ranked.

Study C, teacher RadGraph 7.3.

Prompt. Could you create a clinical report summarizing the key outcomes from this medical image?



CheXpert, 1 view

reference report 58 words

Interval placement of a right upper extremity PICC, with its tip 5 cm below the carina, in the right atrium. Endotracheal tube and feeding tube standard position and appearance. The lung volumes remain low, with mild degree of bibasilar opacities, likely atelectasis. No focal consolidation, and no significant pulmonary edema. Cardiac size at the upper limits of normal.

OPD + teachability gate

685 words · RadGraph 12.9

The key abnormalities observed in the medical images include: 1) A right-sided PICC line with its tip located in the right atrium, which is an abnormal position and may require repositioning. 2) A left-sided PICC line with its tip in the left brachiocephalic vein, which is also an abnormal position and may need adjustment. 3) [...]

OPD + teachability gate + anchor

655 words · RadGraph 8.8

The key abnormalities observed in the sequence of medical images include: 1) A large, irregularly shaped mass in the right upper lobe of the lung, which is likely a primary lung tumor. 2) A smaller, well-defined mass in the left lower lobe, which could be a metastatic lesion or a secondary primary tumor. 3) Multiple [...]

GA-OPD (ours) **56 words** · RadGraph 56.5

Interval placement of a right upper extremity PICC line with the tip in the right atrium. Interval placement of a feeding tube with the tip in the stomach. Low lung volumes with bibasilar opacities, which may represent atelectasis or consolidation. I have personally reviewed the images for this examination and agreed with the report transcribed [...]

The teachability-gated arm keeps the enumeration format of the teacher; it places a right-sided PICC line in the right atrium, as the reference does, adds a left-sided PICC line the reference does not contain, and scores 12.9 RadGraph. Adding the anchor without the drift gate leaves the enumeration in place and replaces the device findings with a lung mass, scoring 8.8. GA-OPD reports the PICC line in the right atrium, the feeding tube, the low lung volumes and the bibasilar opacities, and scores 56.5.