DIST-RRG: DISTRIBUTIONAL CHANGE MODELING ON STATISTICAL MANIFOLDS FOR LONGITUDINAL RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal radiology reports describe the current study in relation to earlier visits. Pointwise change representations retain estimated state but discard the uncertainty attached to each reading. We present Dist-RRG, a distributional framework that represents visits, between-visit changes, and prior reports as diagonal Gaussians. A distributional difference module retains both visits' variances; a trajectory-conditioned, Bhattacharyya-style dissimilarity scores priors; and moment-matched aggregation preserves within-prior uncertainty and between-prior disagreement. Aggregated log-variance modulates the decoder through FiLM. On Longitudinal MIMIC-CXR, micro-averaged CheXbert clinical-efficacy F1 is 0.509 with one prior and 0.516 with up to five priors (0.515 ± 0.003 over three training seeds). Under a matched protocol — same split, evaluation script, and averaging — the strongest re-trained open-source longitudinal baseline reaches 0.482 CE-F1, below Dist-RRG (P1) with a paired interval that excludes zero. The evaluation focuses on whether the propagated second moment carries useful information: progression calibration has 15-bin ECE 0.041 versus 0.063 without moment-matched aggregation; aggregated variance detects unreliable targets with AUROC 0.712; and retaining the least uncertain 70% raises CE-F1 to 0.553. At a fixed checkpoint, permuting variance vectors across patients reduces New-transition F1 from 0.412 to 0.389. These analyses support a bounded conclusion: propagated variance carries patient-specific information about prediction difficulty. They do not establish diagnostic confidence or superiority to other uncertainty estimators.

1 INTRODUCTION

Most sentences in a chest radiograph report are comparative. "Increasing left basilar atelectasis" and "unchanged right costophrenic blunting" describe the current study only in relation to an earlier one, so a generator that captions the current image alone solves a different task from the one radiologists perform. Recent work has responded by conditioning on prior visits (Zhu et al., 2023; Wang et al., 2024; Liu et al., 2025c; Yun et al., 2025; Liu et al., 2025b;a).

These methods share a representational commitment that we think is the binding constraint. A visit becomes a feature vector, a patch-difference map, or a discrete progression label, and change becomes the difference of two such objects — carrying the estimated state but not the confidence attached to it. Architectures that also hand the decoder the absolute prior and current embeddings may encode some of that confidence implicitly, but none has a structural mechanism that isolates uncertainty, accumulates it when two readings are subtracted, and keeps it separate from the estimated change when several priors are combined. The clinically relevant distinction therefore has no dedicated channel: consecutive studies read confidently as unchanged and consecutive studies that were both equivocal yield the same change representation. A decoder that cannot tell these apart has no basis for hedging on the second and committing on the first, and its reports assert more than the evidence supports.

The remedy is to keep the second moment. If each visit is a distribution rather than a point, the between-visit change is also a distribution whose variance accumulates both readings' uncertainty — the same mean shift is weaker evidence when either visit was uncertain. Two consequences follow:

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.

prior scoring should distinguish changes in mean from changes in variance, which information geometry motivates; and combining attended priors should preserve their disagreement, which the law of total variance does exactly. We are careful about which part is the contribution: subtracting means is still linear, and our criticism of prior work is that it discards the second moment and measures change with a Euclidean metric, not that subtraction is too simple a model of disease dynamics. Nonlinearity enters through the manifold on which parameters are compared and the trajectory-dependent reweighting of that comparison.

Dist-RRG. Three components instantiate this. A *distributional difference module* (DDM) forms a Gaussian change distribution per prior and per latent factor, with variance summing across the two visits. *Distributional prior guidance* (DPG) scores each prior with a Bhattacharyya-style dissimilarity, reweighted per dimension by a report-trajectory attention mask (RTAM) and corrected by a gated random-Fourier term. *Moment-matched distributional aggregation* (MMDA) combines the attended priors by matching the first two moments of the mixture they induce, a forward-KL projection onto the Gaussian family, so aggregate variance never falls below the attention-weighted average and strictly exceeds it when the attended means disagree. The result conditions a decoder through a gated temporal-fusion pathway, and the construction reduces exactly to a single-visit reporter when no priors are available.

What the uncertainty is good for. Preserving variance is easy to prove; that the preserved variance means anything is not, and we treat this as the empirical question of the paper, answered with calibration, selective prediction, and a direct intervention on the variance pathway (§4.4). The claim is correspondingly narrow — the variance is an operational signal of prediction difficulty, not an established clinical measurement, which would need multi-reader evaluation we do not have.

Contributions.

1. A distributional formulation of longitudinal report generation in which visits, changes, and prior reports are diagonal Gaussians on a shared statistical manifold, with the second moment propagated end to end (§3.2–§3.5).
2. Variance-sensitive temporal aggregation: an RTAM-reweighted Gaussian dissimilarity for scoring priors and MMDA for combining them, with the exact second-order relation between the surrogate and the Fisher–Rao metric, including its anisotropy at the deployed coefficient (Prop. 1).
3. An evaluation beyond corpus-level n-gram and CheXbert scores: matched reruns of open-source baselines with paired intervals, three-seed variability, a CheXbert-independent metric, transition F1 against a copy-prior control, per-class calibration, selective prediction, and a variance intervention (§4.2–§4.5).

2 Related Work

Longitudinal report generation, and why priors stay pointwise. Single-visit reporting is well developed — memory-driven and cross-modal transformers (Chen et al., 2020; 2021), warm-started decoders (Nicolson et al., 2023), frozen LLM decoders (Wang et al., 2023), clinically structured variants (Xiang et al., 2024; Yan et al., 2026; Cheng & Subramanian, 2026) — and is orthogonal to us, since our per-visit encoder is frozen and could be any of them. Prefilling feeds a prior image and report through a hierarchical memory-driven decoder (Zhu et al., 2023); HERGen processes whole patient sequences with a group-causal transformer (Wang et al., 2024); HC-LLM constrains an LLM decoder with time-shared and time-specific terms (Liu et al., 2025c); MLRG adds multi-view longitudinal contrastive pretraining (Liu et al., 2025b); Diff-RRG extracts disease-wise patch-difference maps and hands discrete progression labels to the decoder (Yun et al., 2025); and the concurrent TIM decouples spatial pathology from temporal progression and iteratively refines the prior and current reports against each other (Dong et al., 2026). Two methods model progression more richly than a single feature subtraction: RECAP reasons over a dynamic disease-progression graph built from observation and progression predictions on two consecutive radiographs (Hou et al., 2023), and DDaTR propagates difference-aware temporal residuals through every stage of the visual encoder (Song et al., 2025). Across all of these a prior is a point — a token, a feature map, a graph node, or a label — and change is a displacement between points, even where that displacement is computed

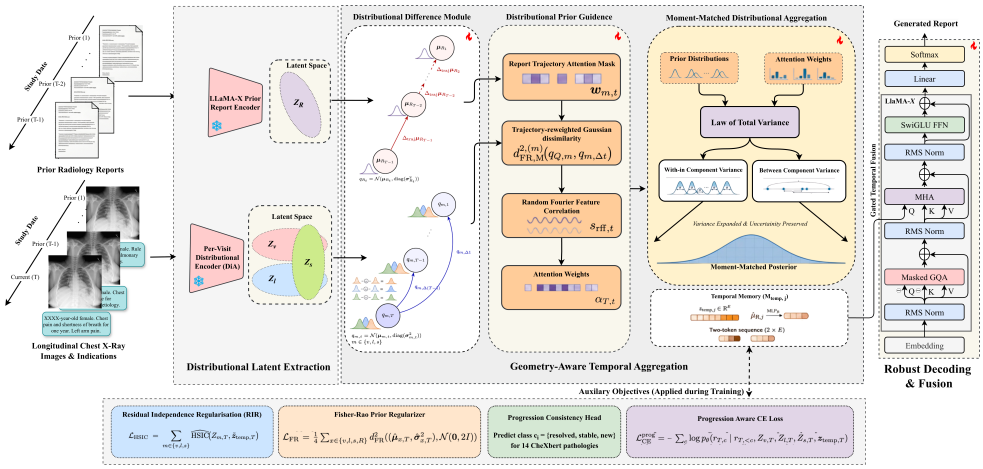


Figure 1: Dist-RRG. Prior reports and per-visit inputs become diagonal Gaussians (left). DDM forms a change distribution per prior whose variance sums both visits' uncertainty, DPG scores priors with a trajectory-reweighted Gaussian dissimilarity, and MMDA combines them by the law of total variance so between-prior disagreement is retained (centre); the aggregated moments form a two-token memory conditioning the decoder through gated fusion (right).

at several scales or reasoned over symbolically; all but HERGen operate on a single current–prior pair. The consequence is not limited expressiveness but a lost second moment: the representation of "stable" is identical whether both readings were confident or both equivocal. That is the gap Dist-RRG targets, and the mechanism that closes it is what Table 2 ablates.

Geometry on distributions. Fisher–Rao distance is the canonical Riemannian distance on parametric families (Rao, 1945; Amari & Nagaoka, 2000), and Čencov's theorem (Čencov, 1982; Campbell, 1986) makes the Fisher metric the unique invariant choice up to scale. Exact geodesics on the multivariate normal model exist only in special cases (Skovgaard, 1984; Pinele et al., 2020), so practical systems use approximations with characterised error (Atkinson & Mitchell, 1981; Nielsen, 2024). We follow that practice, and §3.4 states exactly where our surrogate departs from the metric. What differs here is what sits on the manifold — per-visit posteriors and change distributions — and that report trajectories drive the attention.

Uncertainty in report generation. Uncertainty-aware decoding exists for single-study reporting (Wang et al., 2025), and disentangled variational encoders are used for multimodal clinical fusion with missing inputs (Shaik et al., 2026; Yao et al., 2024; Shi et al., 2019). Our per-visit encoder is one of these (Shaik et al., 2026), frozen and used strictly as a distributional encoder. What is new is not that a variational encoder emits variance but that the variance survives *temporal* aggregation, and that we test whether what survives is useful rather than inferring it from accuracy.

3 Method

3.1 Setup

Figure 1 gives an overview. A patient record $\mathcal{D}_i = \{(V_t, L_t, R_t)\}_{t=1}^{T_i}$ pairs a radiograph, an indication, and a findings section at irregular intervals. Writing T for the target visit, the usable history is the most recent $K = \min(T-1, H)$ priors with $H = 5$, each carrying availability indicators $a_t = (a_t^V, a_t^L, a_t^R)$. The model is a conditional generator $p_\theta(R_T \mid \{q_{m,T}\}_m, \mathcal{H}_T)$ with two requirements: **(R1)** at $K = 0$ the forward pass and objective coincide numerically with the underlying single-visit reporter, and **(R2)** both stay well defined for any availability configuration; both hold by construction (§3.6). Per-visit posteriors come from a frozen DiA-gnostic VLVAE (Shaik et al., 2026), used only as a distributional encoder: it factorises $q_\phi(Z_v, Z_l, Z_s \mid V_t, L_t)$ into visual, indication, and mixture-

of-experts shared factors, each a diagonal Gaussian $q_{m,t} = \mathcal{N}(\mu_{m,t}, \mathrm{diag}\sigma^2_{m,t})$, $m \in \{v, l, s\}$; we consume those six parameter tensors and nothing else.

3.2 Distributional difference module

For every prior $t < T$ and factor m, DDM forms the between-visit change distribution. Treating the two visits' posteriors as independent,

$$q_{m,\Delta t} = \mathcal{N}\Big(\underbrace{\mu_{m,T} - \mu_{m,t}}_{\text{mean shift}},\ \mathrm{diag}\big(\underbrace{\sigma^2_{m,T} + \sigma^2_{m,t}}_{\text{variance sum}}\big)\Big). \tag{1}$$

The variance accumulates, so a fixed mean shift is weaker evidence when either visit was uncertain — the "stable and confident" versus "stable but equivocal" distinction that patch-difference maps (Yun et al., 2025) and token-level difference losses (Liu et al., 2025c) collapse by construction.

Consecutive studies are correlated, so Eq. (1) is an approximation whose direction is known: under positive cross-visit covariance it *over*-states change variance, conservative rather than over-confident (Remark 2, Appendix B); §4.5 sweeps an assumed ρ.

3.3 Prior-report encoder

To place prior reports on the same manifold, LX-PRE runs a frozen LLaMA-X stack (Su et al., 2024; Shazeer, 2020; Zhang & Sennrich, 2019) causally over the tokens of R_t with a learned null cross-attention memory, takes the final-layer hidden state at the last non-padding token (under a causal mask, the only position that has seen the whole report; no mean pooling), and maps it to $(\mu_{R,t}, \log \sigma^2_{R,t})$ through four linear projections. Only the null memory and those projections are trained; the KV cache is saved and restored around each pass so encoding priors does not disturb generation (Appendix J).

3.4 Distributional prior guidance

DPG produces one attention weight $\alpha_{T,t}$ per prior $t < T$, scoring each prior on four distributional blocks $\mathcal{B} = \{v, l, s, R\}$. Throughout, the block distribution $q_{b,t}$ of prior t is the *change* $q_{b,\Delta t}$ of Eq. (1) for $b \in \{v, l, s\}$ and the *absolute* report state $\mathcal{N}(\mu_{R,t}, \mathrm{diag}\sigma^2_{R,t})$ of §3.3 for $b = R$; this convention fixes what DPG scores and what MMDA aggregates (§3.5) and regularises (§3.7).

A variance-sensitive Gaussian dissimilarity. DPG uses a Bhattacharyya-style score whose local quadratic form can be related to the Fisher information metric (Amari & Nagaoka, 2000). It is a computational scoring rule, rather than an exact Fisher–Rao geodesic. For $q_A = \mathcal{N}(\mu_A, \mathrm{diag}\sigma^2_A)$, $q_B = \mathcal{N}(\mu_B, \mathrm{diag}\sigma^2_B)$ and $\bar{\sigma}^2_d = \frac{1}{2}(\sigma^2_{A,d} + \sigma^2_{B,d})$,

$$d^2_\kappa(q_A, q_B) \ = \ \sum_{d=1}^{D} \left[\frac{(\mu_{A,d} - \mu_{B,d})^2}{\bar{\sigma}^2_d} \ + \ \kappa \log \frac{\bar{\sigma}^2_d}{\sigma_{A,d}\,\sigma_{B,d}} \right], \qquad \kappa > 0. \tag{2}$$

Both terms are dimensionless; the logarithmic term is non-negative by the AM–GM inequality.

Proposition 1 (Surrogate geometry)**.** *d^2_κ is symmetric, non-negative, and zero iff $q_A = q_B$. To second order it equals the squared geodesic distance of $g_\kappa = \mathrm{diag}(\sigma^{-2}, \frac{\kappa}{2}\sigma^{-2})$ on each (μ_d, σ_d) half-plane. Since the Fisher metric is $g_{\mathrm{FR}} = \mathrm{diag}(\sigma^{-2}, 2\sigma^{-2})$, we have $g_\kappa = g_{\mathrm{FR}}$ iff $\kappa = 4$. For $\kappa \neq 4$, d^2_κ matches the exact squared Fisher–Rao distance to leading order along pure-mean displacements but scales pure-variance displacements by $\kappa/4$, so for $\kappa < 4$ it is a local* lower *bound in that direction.*

The proof is in Appendix A. The match at $\kappa = 4$ is local and does not establish global geodesic equality or invariance of the reweighted attention. All results use $\kappa = 1$, which under-weights variance displacements fourfold relative to Fisher–Rao; κ is a design coefficient, not a derived constant, and a $\kappa = 4$ training comparison is not evaluated.

Report-trajectory attention mask. The conditioning signal is not a prior report's content but the direction in which the reports have been moving. With t^- the most recent slot before t having

$a^R_{t^-} = 1$ and $\mu^{\varnothing}_R$ a learned null used when none exists,

$$\Delta_{\mathrm{traj}}\mu_{R,t} \;:=\; \mu_{R,t} - \mu_{R,t^-}, \qquad \mu_{R,t^-} := \mu^{\varnothing}_R \text{ if } t^- = \varnothing, \tag{3}$$

is defined at every slot including the earliest. A low-rank projection maps it to a per-dimension gate $w_{m,t,d} = \mathrm{sig}((\Delta_{\mathrm{traj}}\mu_{R,t})^\top U_m V_m[:,d] + b_{m,d}) \in (0,1)$. The mask modulates sensitivity to each parameter dimension. It reweights Eq. (2) per dimension,

$$d^{2,(m)}_{\kappa,\mathrm{M}}\big(q_{Q,m}, q_{m,\Delta t}\big) = \sum_{d=1}^{D_m} w_{m,t,d}\left[\frac{(\mu_{Q,m,d} - \mu_{\Delta,m,t,d})^2}{\bar\sigma^2_{m,t,d}} + \kappa \log \frac{\bar\sigma^2_{m,t,d}}{\sigma_{Q,m,d}\,\sigma_{\Delta,m,t,d}}\right], \tag{4}$$

against a learned query prototype $q_{Q,m}$. For fixed $w > 0$, the score has the local quadratic form of the positive definite metric $\mathrm{diag}(w)\, g_\kappa$. Because w depends on the prior being scored, the attention rule is a trajectory-conditioned dissimilarity, not a distance in a single fixed metric (Appendix B). The report block is unmasked ($w \equiv 1$) and, since R_T is the generation target and has no change distribution, is scored on the absolute prior state, $d^{2,(R)}_{\kappa,\mathrm{M}} = d^2_\kappa(q_{Q,R}, q_{R,t})$, against its own learned prototype.

Block combination, random-Fourier correction, and attention. Learned block weights $\tilde w = \mathrm{softmax}(\ell_{\mathrm{blk}})$ combine the four blocks under availability, renormalised by the present mass, $d^2_{\kappa,\mathrm{M}} = \sum_b a^b_t \tilde w_b d^{2,(b)}_{\kappa,\mathrm{M}} \,/\, \sum_b a^b_t \tilde w_b$, so scores stay comparable across priors with different availability. A gated random-Fourier term (Rahimi & Recht, 2007) adds a learnable kernel similarity, $s_{\mathrm{rff},t} = \phi(x_t)^\top W_\phi \phi(x_Q)$, with ϕ a frozen random feature map on $x_t = [\mu_{m,\Delta t};\, \log\sigma^2_{m,\Delta t}]_{m\in\{v,l,s\}}$, the stacked parameters of the three change blocks (an unavailable block contributes its learned null parameters; x_Q stacks the prototypes likewise), W_ϕ zero-initialised, and gate $\gamma = \mathrm{sig}(g_\phi)$ with a learnable scalar g_ϕ initialised at -5 (Table 13). With learnable temperature τ and visit mask m_t,

$$s_{T,t} = -\frac{d^2_{\kappa,\mathrm{M}}(q_Q, q_t)}{2\tau} + \gamma\, s_{\mathrm{rff},t}, \qquad \alpha_{T,t} = \frac{\exp(s_{T,t})\, m_t}{\sum_{t'} \exp(s_{T,t'})\, m_{t'}} \quad (K \geq 1). \tag{5}$$

With $W_\phi = 0$ at initialisation, the random-Fourier correction is inactive; the attention weights depend on the Gaussian dissimilarity.

3.5 Moment-matched distributional aggregation

Attention weights and the per-block distributions $q_{b,t}$ of §3.4 induce a Gaussian mixture for each block. MMDA replaces it with the diagonal Gaussian $\hat q_{b,T}$ matching its first two moments,

$$\hat\mu_{b,T} = \sum_t \alpha_{T,t}\,\mu_{b,t}, \qquad \hat\sigma^2_{b,T} = \underbrace{\sum_t \alpha_{T,t}\,\sigma^2_{b,t}}_{\text{within-prior}} + \underbrace{\sum_t \alpha_{T,t}\big(\mu_{b,t} - \hat\mu_{b,T}\big)^2}_{\text{between-prior}}. \tag{6}$$

Proposition 2 (Forward-KL projection and variance preservation)**.** *Eq.* (6) *is the unique minimiser over diagonal Gaussians q of the forward (inclusive) divergence $D_{\mathrm{KL}}(p_{\mathrm{mix}} \,\|\, q)$, where $p_{\mathrm{mix}} = \sum_t \alpha_{T,t}\mathcal{N}(\mu_{b,t}, \mathrm{diag}\sigma^2_{b,t})$. Consequently $\hat\sigma^2_{b,T} \geq \sum_t \alpha_{T,t}\sigma^2_{b,t}$, with equality if and only if all attended means coincide.*

Moment matching is the standard M-projection under forward KL (Minka, 2005); its role here is to retain the second moment during temporal attention, not to introduce a new projection theorem. Attention spread over disagreeing priors strictly increases aggregated variance, keeping retrieval ambiguity visible downstream.

For $m \in \{v,l,s\}$, $\hat q_{m,T}$ is a distribution over the change from the attended history, not over the history itself; the two views differ only by sign. Since $\mu_{m,T}$ is common to every change block and $\sum_t \alpha_{T,t} = 1$, the implied aggregated prior state $\bar\mu_{m,T} := \mu_{m,T} - \hat\mu_{m,T} = \sum_t \alpha_{T,t}\mu_{m,t}$ is exact and the between-prior term of Eq. (6) is unchanged (Eq. (8)). We aggregate changes, so the no-change reference of §3.7 concerns $\hat q_{m,T}$ alone, and recover $\bar\mu_{m,T}$ where a prior state is needed.

3.6 TEMPORAL MEMORY, DECODING, AND THE EMPTY-HISTORY BRANCH

We draw $z_{b,T} = \hat{\mu}_{b,T} + \hat{\sigma}_{b,T} \odot \varepsilon_b$, $\varepsilon_b \sim \mathcal{N}(0, I)$, at training and take the mean at inference. The stacked sample $z_{\text{temp},T} := [z_{b,T}]_{b \in \mathcal{B}}$ is the temporal latent; its law $\hat{q}_T = \prod_b \mathcal{N}(\hat{\mu}_{b,T}, \text{diag}\hat{\sigma}^2_{b,T})$ is the *trainable temporal posterior* of §3.7, trainable because its moments pass through every Stage-B module that feeds the attention while the per-visit moments beneath them stay frozen. The aggregated log-variances are not merely carried along: they modulate the temporal token through FiLM (Perez et al., 2018), the pathway we intervene on in §4.4. With u_T stacking the four $\log(\hat{\sigma}^2_{b,T} + \epsilon)$,

$$M_{\text{temp}}[1] = \big(1 + \gamma_\psi(u_T)\big) \odot \text{MLP}_{\text{temp}}\big(z_{\text{temp},T}\big) + \beta_\psi(u_T), \qquad M_{\text{temp}}[2] = \text{MLP}_R(\hat{\mu}_{R,T}), \tag{7}$$

giving a two-token memory $M_{\text{temp}} \in \mathbb{R}^{2 \times E}$, which enters every decoder layer as a gated residual cross-attention $h' = h + \text{sig}(g)\,\text{MHA}(\text{Norm}(h), \text{Norm}(M_{\text{temp}}))$ with per-layer gate initialised at $g = -1$. For $K = 0$, define the forward pass to bypass the temporal residual entirely; no attention weights or aggregated moments are evaluated. The memory and the temporal latent are set to zero ($M_{\text{temp}} = 0$, $z_{\text{temp},T} = 0$) and every history-dependent loss is gated off. This explicit bypass makes the branch identical to the single-visit reporter (Prop. 4) without relying on attention over zero inputs, which biases can make nonzero. For missing inputs, learned null distributions and availability-renormalised scores keep the computation defined; the factorised variational treatment assumes observed and missing blocks are conditionally independent given the remaining latent state, a modelling approximation rather than a claim about the clinical missingness process (Appendix C).

3.7 TRAINING OBJECTIVE

Training has two stages. Stage A fits the per-visit encoder with its own objective $\mathcal{L}_{\text{enc}}$ (Shaik et al., 2026). Stage B freezes it and trains the temporal pathway, the decoder, and the auxiliary heads; $\mathcal{L}_{\text{enc}}$ is then constant and only the terms below carry gradient.

Terms. The likelihood is a progression-aware cross-entropy on the target report. Three auxiliary terms shape the temporal pathway; all carry $\mathbf{1}[K \geq 1]$ and their equations are in Appendix D. *(i)* A *surrogate prior* pulls aggregated blocks toward a no-change reference using the same dissimilarity as the attention scores. The reference differs by block type: change blocks are differences of two visits and go toward $\mathcal{N}(0, 2I)$, the report block is an absolute state and goes toward $\mathcal{N}(0, I)$. *(ii)* A *progression-consistency* head predicts a change class in $\{\text{resolved}, \text{stable}, \text{new}\}$ for each of 14 CheXbert findings. A relational class cannot be predicted from the aggregated history alone, so the head consumes the current shared mean $\mu_{s,T}$, the implied aggregated prior state $\bar{\mu}_{s,T}$ of Eq. (8), and the aggregated change $\hat{\mu}_{s,T}$; Appendix D says why a trajectory-level input is paired with a single-step label. It uses CheXbert supervision the baselines do not, and §4.3 examines its conditional incremental effect with a CheXbert-independent metric. *(iii) Residual independence* between the temporal latent and the per-visit factors is regularised — not guaranteed — with HSIC (Gretton et al., 2005; Sriperumbudur et al., 2010), since linear orthogonality is too weak when the coupling is nonlinear.

Total. $\mathcal{L} = \lambda_{\text{CE}}\mathcal{L}^{\text{prog}}_{\text{CE}} + \lambda_{\text{E}}\mathcal{L}_{\text{ELBO}} + \lambda_{\text{FR}}\mathcal{L}_{\text{FR}} + \lambda_{\text{PC}}\mathcal{L}_{\text{PC}} + \lambda_{\text{H}}\mathcal{L}_{\text{HSIC}}$ with $(1.0,\ 10^{-2},\ 0.1,\ 0.3,\ 0.3)$, where $\mathcal{L}_{\text{ELBO}}$ is the closed-form KL from the temporal posterior $\hat{q}_T$ (§3.6) to the block references of *(i)* (Eq. (10)); it is what makes the likelihood a variational bound (Prop. 5), whereas *(i)* is an auxiliary penalty in the attention's geometry. Weights come from a coarse search on the validation split and were then fixed for every protocol and ablation in §4; no per-row retuning was performed.

4 EXPERIMENTS

4.1 SETUP

Data. We use Longitudinal MIMIC-CXR (Johnson et al., 2019) with the patient-level split of Zhu et al. (2023). The test split has 266 patients and 2,324 study visits; every study but each patient's first has a predecessor, giving $2{,}324 - 266 = 2{,}058$ targets and, at 14 CheXbert findings each, 28,812 target–pathology decisions. Roughly 45% of test cases lack an indication.

Table 1: Results on Longitudinal MIMIC-CXR. Longitudinal baselines use P1; Dist-RRG also uses P2 (up to five priors). S = secondary benchmark comparison; O = original-paper result; M = matched rerun, retrained from released code on our split and scored with our evaluation script; R = our run (Appendix F). S and O rows retain their source CE-F1 averaging conventions; only M and R rows share one convention (micro-averaged) and are ranked against each other. RG-F1 = RadGraph-F1, reported only where we hold generated reports. Cross-method ROUGE-L is omitted because it depends on sentence segmentation; within-model ROUGE-L is in Table 2. N-gram overlap alone cannot validate clinical correctness (Yu et al., 2023).

Method	Source	BLEU				Clinical efficacy			RG-F1
		B@1	B@2	B@3	B@4	Prec	Rec	F1	
Single-visit input									
AoANet (Huang et al., 2019)	S	0.272	0.168	0.112	0.080	0.437	0.249	0.317	–
R2Gen (Chen et al., 2020)	S	0.308	0.190	0.126	0.089	0.457	0.290	0.355	–
R2GenCMN (Chen et al., 2021)	S	0.328	0.203	0.135	0.096	0.521	0.396	0.449	–
CvT2DistilGPT2 (Nicolson et al., 2023)	S	0.363	0.225	0.151	0.109	0.485	0.310	0.378	–
R2GenGPT (Wang et al., 2023)	S	0.345	0.202	0.128	0.087	0.331	0.253	0.287	–
GMoD (Xiang et al., 2024)	S	0.378	0.234	0.155	0.107	0.496	0.429	0.460	–
Longitudinal input, single prior (protocol P1)									
Prefilling (Zhu et al., 2023)	S	0.343	0.210	0.141	0.100	0.506	0.364	0.423	–
HERGen (Wang et al., 2024)	S	0.389	0.242	0.163	0.117	0.421	0.289	0.295	–
HC-LLM (Liu et al., 2025c)	S	0.404	0.247	0.164	0.116	0.488	0.415	0.448	–
Diff-RRG (Yun et al., 2025)	S	0.405	0.251	0.169	0.120	0.528	0.430	0.474	–
TIM (Dong et al., 2026)	O	0.430	0.265	0.179	0.124	0.563	0.505	0.511	–
Prefilling (Zhu et al., 2023)	M	0.348	0.215	0.145	0.103	0.511	0.380	0.435	0.165
HERGen (Wang et al., 2024)	M	0.392	0.248	0.168	0.121	0.435	0.245	0.312	0.120
Diff-RRG (Yun et al., 2025)	M	0.411	0.255	0.172	0.123	0.535	0.439	0.482	0.198
Dist-RRG (P1, single prior)	R	0.421	0.263	0.181	0.128	0.575	0.436	0.509	0.214
Dist-RRG (P2, variable history)	R	0.425	0.268	0.185	0.131	0.582	0.441	0.516	0.227

Protocols. **P1** pairs each target with its immediately preceding study only, the setting published longitudinal baselines report; **P2** uses up to $H = 5$ priors.

Metrics. NLG quality is BLEU-$\{1..4\}$ (Papineni et al., 2002). Within-model ablations add ROUGE-L (Lin, 2004) and BERTScore F1 (Zhang et al., 2020). Clinical efficacy uses CheXbert over 14 findings (Smit et al., 2020b;a). For Dist-RRG and every matched rerun, precision and recall are pathology-wise averages and CE-F1 is micro-averaged over all target–pathology decisions. Because CheXbert also supervises our progression head, we add RadGraph-F1 (RG-F1) (Jain et al., 2021; Delbrouck et al., 2022), an entity–relation overlap score with no CheXbert component (Appendix E). Progression-transition F1 scores stable/new/resolved labels derived from generated and reference reports (Appendix E); calibration uses 15-bin ECE on the progression head's probabilities. Intervals come from a patient-level bootstrap over the 266 test patients, so studies from one patient are not treated as independent; a difference between two systems on the same targets uses the paired version of that bootstrap. The full model and the two ablations our conclusions rest on (MMDA disabled, Euclidean attention) are trained with three seeds; seed 0 is the fixed checkpoint used by every fixed-checkpoint analysis.

Implementation. Per-factor latent 256, RTAM rank 32, 512 random Fourier features, $\kappa = 1$, $H = 5$, two-token memory, 12-layer decoder; AdamW at 5×10^{-5}, 10% warm-up, cosine decay, 25 epochs, batch 4; 114M trainable parameters on one NVIDIA A40. Appendix J gives every setting.

4.2 Comparison with prior methods

Dist-RRG reaches micro CE-F1 0.509 under P1 and 0.516 under P2. On the same 2,058 targets the paired patient-level bootstrap puts the P2−P1 difference at 0.007 with 95% interval $[0.001, 0.014]$, a modest but measurable gain from longer history; across three training seeds the P2 CE-F1 is 0.515 ± 0.003 (Table 5).

The matched block is the ranking evidence. Each rerun uses the authors' released code and source hyperparameters, with model selection on our validation split only (Appendix F). Under the same

Table 2: Component-level ablation on Longitudinal MIMIC-CXR under variable history ($K \leq 5$). The first row is the frozen per-visit encoder; subsequent rows incrementally add each component. Single-run scores; three-seed variability and paired intervals for complete-model contrasts are in Table 5. RG-F1 has no CheXbert component.

LX-PRE	DDM	DPG	MMDA	PC	RIR	B@1	B@4	ROUGE$_L$	BERT-F1	CE-F1	RG-F1
×	×	×	×	×	×	0.381	0.102	0.331	0.835	0.428	0.153
✓	×	×	×	×	×	0.388	0.107	0.338	0.841	0.436	0.161
✓	✓	×	×	×	×	0.395	0.112	0.345	0.848	0.452	0.174
✓	✓	✓	×	×	×	0.402	0.117	0.352	0.854	0.468	0.189
✓	✓	✓	✓	×	×	0.409	0.122	0.359	0.861	0.485	0.201
✓	✓	✓	✓	✓	×	0.418	0.126	0.368	0.866	0.502	0.215
✓	✓	✓	✓	✓	✓	**0.425**	**0.131**	**0.378**	**0.871**	**0.516**	**0.227**

split, preprocessing, and micro-averaged CheXbert scoring, Dist-RRG (P1) exceeds the strongest matched rerun, Diff-RRG, by 0.027 CE-F1 (paired 95% interval $[0.012, 0.041]$) and by 0.016 RG-F1 ($[0.005, 0.028]$); both intervals exclude zero. Each rerun is a single training run, so baseline seed variance is unmeasured; the CE-F1 gap is about nine times our own across-seed standard deviation. TIM, the strongest published result and the BLEU-1 leader, is concurrent work and is not among the reruns, so this is not a state-of-the-art claim: published values, including TIM's 0.511, keep their source conventions and are not ranked against our rows. MLRG (Liu et al., 2025b) and DDaTR (Song et al., 2025) report on other splits and are not tabulated (Appendix F).

4.3 Which components matter

Table 2 measures additions in a fixed order. With LX-PRE, DDM, and DPG held fixed, replacing naive averaging with moment matching raises CE-F1 from 0.468 to 0.485. This conditional contrast does not identify MMDA's contribution in the complete model; Table 5 does. Disabling MMDA in the complete model gives CE-F1 0.489 (paired Δ -0.027, 95% interval $[-0.042, -0.015]$), and the other complete-model removals and substitutions reduce CE-F1 by 0.006–0.013, each with a paired interval on the same 2,058 targets. The intervals for disabling MMDA, Euclidean attention, and removing RTAM exclude zero; for the first two, which have three training seeds, the seed-mean gap to the full model also exceeds the across-seed standard deviation, while RTAM was trained once and is supported against test-sampling variability only. Removing the random-Fourier correction or the surrogate prior gives intervals that include zero, so their contributions are not distinguishable from test-set noise.

Auxiliary supervision. The progression-consistency head uses CheXbert transition labels, and CheXbert also evaluates clinical efficacy. Adding PC before RIR increases CE-F1 from 0.485 to 0.502; over the same step RG-F1, which uses no CheXbert component, moves from 0.201 to 0.215. Because the RG-F1 gain ($+0.014$) is comparable to the CE-F1 gain ($+0.017$), the PC gain is not specific to CheXbert; both are single-run differences.

4.4 Is the preserved variance worth anything?

Prop. 2 guarantees variance survives aggregation; whether it carries information is separate. Calibration, selective prediction, and variance intervention examine this question at a fixed checkpoint on the full test set; Figure 2 (Appendix G) plots all three.

Calibration. Full Dist-RRG attains a 15-bin ECE of 0.041 against 0.063 for the matched MMDA-disabled ablation, a 35% relative reduction (Table 10); over three training seeds the aggregate ECE is 0.042 ± 0.002 versus 0.064 ± 0.004 (Table 5). The MMDA-disabled model reaches CE-F1 0.489, 0.027 below the full model, so moment matching improves calibration without costing clinical efficacy. Since 79.5% of transition labels are stable, an aggregate ECE could improve through the majority class alone; Table 10 shows it does not, with one-vs-rest ECE improving on all three classes and macro ECE going $0.081 \rightarrow 0.052$.

Ranking unreliable predictions. Define $u_i = D_s^{-1} \sum_d \hat{\sigma}^2_{s,T,i,d}$, the dimension-normalised variance of the aggregated shared-change block, and call a target *unreliable* when at least one actionable (*new* or *resolved*) transition is wrong. As a detection score u gives AUROC 0.712 and AUPRC 0.536 at prevalence 0.328 (Table 12); in selective prediction (Table 11) the lowest-uncertainty $90/80/70\%$ raise CE-F1 to $0.529/0.541/0.553$ and cut the actionable error rate to $0.292/0.253/0.216$, thresholds from the global ranking, nothing refitted.

Intervention on the variance pathway. Correlation cannot rule out that the channel helps merely by existing. Holding the checkpoint, aggregated means, attention weights, and inputs fixed, we modify only the log-variances entering the FiLM layer of Eq. (7): *shuffling* permutes variance vectors across patients (five fixed permutations), preserving the marginal while destroying patient correspondence; *clamping* substitutes the validation mean. The cost lands where it should (Table 9): New F1 falls $0.412 \rightarrow 0.389 \rightarrow 0.376$, Resolved F1 $0.730 \rightarrow 0.703 \rightarrow 0.691$, ECE rises $0.041 \rightarrow 0.056 \rightarrow 0.061$.

4.5 Where the accuracy comes from

Aggregate CE-F1 hides where a longitudinal model earns its score, since the transition distribution is dominated by "no change". Table 3 adds the control that makes it interpretable: a copy-prior predictor reaches 0.887 stable F1 and zero on both actionable classes, so a model can look strong on the majority class while detecting no progression.

	Stable	New	Resolved
Copy-prior control	0.887	0.000	0.000
Prefilling (M, P1)	0.865	0.210	0.450
HERGen (M, P1)	0.820	0.115	0.280
Diff-RRG (M, P1)	0.881	0.345	0.612
Dist-RRG (P1)	0.889	0.385	0.684
Dist-RRG (P2)	**0.894**	**0.412**	**0.730**

Table 3: Progression-transition F1 over $28{,}812$ labels ($79.5/10.3/10.2\%$ stable/new/resolved); M rows are the matched reruns of Table 1. Distinct from the presence-level CE-F1.

Dist-RRG exceeds the control on stable and is non-trivial on new (0.412) and resolved (0.730), the easier class since a finding present in the prior gives the model something to reference. At equal history (P1) it exceeds the strongest matched rerun, Diff-RRG, by 0.040 on new and 0.072 on resolved transitions; this is a whole-system comparison without intervals and does not isolate DDM. Appendix L shows a case in which the five-visit history, unlike the most recent prior alone, avoids an unsupported change claim.

Robustness. CE-F1 degrades gradually as inputs are withheld, reaching 0.493 with every prior report removed, and falls from 0.524 under 90 days of follow-up to 0.498 beyond two years — tolerance, not adaptation, since elapsed time is not an input. Rare-finding macro-F1 rises from 0.288 to 0.319, and an assumed cross-visit correlation up to $\rho = 0.75$ costs 0.012 CE-F1 (Appendix G).

5 Discussion

Dist-RRG preserves the second moment through temporal attention, and fixed-checkpoint calibration, selective prediction, and variance intervention show a modest operational contribution; none of this establishes clinical uncertainty or superiority to entropy, single-visit variance, dropout, ensembles, or the between-prior term alone. Calibration is scored against the CheXbert labels that supervise the progression head, which RG-F1 does not address. Matched reruns cover Prefilling, HERGen, and Diff-RRG as single runs; TIM, HC-LLM, MLRG, DDaTR, and RECAP remain uncompared on our split. Seeds cover the full model and two ablations; complete-model PC and RIR removals, $\kappa = 4$ training, and a most-recent-prior PC head are absent. Evaluation uses one institution and automated labels; MS-CXR-T, CheXpert Plus (Appendix H), GREEN, and a reader study (Appendix I) remain necessary. The claim is variance utility within this model and accuracy under one matched protocol, not a clinical guarantee.

AI USE STATEMENT

We used generative AI tools for editing assistance (grammar, spelling, LaTeX organization), to redraft specific passages for clarity, and to assist in planning additional evaluation experiments. We did not use generative AI tools to generate research ideas, to design the method, to produce or analyse experimental results, or to write code that produced any reported number. All LaTeX, figures, and tables were reviewed line by line by the authors against the underlying experiment logs. We take responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work uses MIMIC-CXR-JPG (Johnson et al., 2019) under the PhysioNet Credentialed Health Data Use Agreement. The data are de-identified and were collected previously; no human-subjects research was conducted by the authors and no IRB approval was required. Two risks deserve naming. First, MIMIC-CXR draws from a single institution, and models trained on it inherit its demographic and practice-pattern composition; performance on under-represented groups is not characterised here. Second, a system that drafts comparative findings can entrench a prior report's errors, since the copy-prior control in Table 3 shows how much of the transition distribution is reachable by repetition alone. Dist-RRG is a drafting aid for a radiologist who reviews and signs the report. It is not validated for autonomous use, and the uncertainty it produces should not be presented to a clinician as a calibrated statement of diagnostic confidence (§5).

REPRODUCIBILITY STATEMENT

The per-visit encoder, the three temporal modules, the training objective, and every hyperparameter needed to retrain the model are specified in §3, §4.1, and Appendix J; Table 13 lists the fixed configuration, selected once by a coarse search on the validation split and held fixed across every protocol and ablation. Appendix K states the Stage-B training step in full, including the availability masks and the empty-history bypass. All five propositions are stated with their assumptions and proved in Appendices A–C; the two modelling assumptions they rest on — the cross-visit independence approximation behind Eq. 1 and the conditional independence of observed and missing blocks — are stated explicitly in Remark 2 and Remark 5. Experiments use the public Longitudinal MIMIC-CXR split of Zhu et al. (2023); MIMIC-CXR-JPG is distributed under the PhysioNet Credentialed Health Data Use Agreement, so we will share the split definition and the preprocessing code, not the images or reports themselves. Appendix D states every Stage-B objective term in closed form, including the temporal-posterior KL and the projection removal in the likelihood. Appendix E gives the target-counting convention, the complete mapping from CheXbert label pairs to stable/new/resolved transition labels (Table 4), the 15-bin ECE procedure, the paired bootstrap, and the RadGraph-F1 configuration; Appendix F documents the provenance of every published value in Table 1 and the protocol of every matched rerun. Anonymised code, the evaluation script for each table, the fixed checkpoint (seed 0) used for all uncertainty analyses, and the configuration and code revision of each matched baseline rerun will be made available to reviewers and area chairs through an anonymous repository when the discussion period opens. The full model and the MMDA-disabled and Euclidean-attention ablations are trained with three seeds; the remaining ablations are single runs. The inference-mask seeds in Table 6 and the variance permutations in Table 9 quantify perturbation variability and are distinct from training-seed variability.

REFERENCES

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

Colin Atkinson and A. F. S. Mitchell. Rao's distance measure. *Sankhyā: The Indian Journal of Statistics, Series A*, 43:345–365, 1981.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision–language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

L. Lorne Campbell. An extended Čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98(1):135–141, 1986.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449, 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)*, pp. 5904–5914, 2021.

Sheng Cheng and Devika Subramanian. Rethinking Radiology Report Generation: From Narrative Flow to Topic-Guided Findings. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jean-Benoit Delbrouck, Pierre Chambon, Khaled Saab, Amin Shirasteh, Curtis Langlotz, and Daniel Rubin. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.319.

Yiheng Dong, Yi Lin, Shilong Huang, Xiyan Yang, and Xin Yang. TIM: Temporal Decoupling with Iterative Mutual-Refinement Model for Longitudinal Radiology Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6951–6961, 2026.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory (ALT)*, pp. 63–77, 2005.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. RECAP: Towards precise radiology report generation via dynamic disease progression reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2134–2147, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.140.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4634–4643, 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. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Alistair EW 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.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Chen Liu, Wenfang Yao, Kejing Yin, William K Cheung, and Jing Qin. Multimodal Disease Progression Modeling via Spatiotemporal Disentanglement and Multiscale Alignment. *arXiv preprint arXiv:2510.11112*, 2025a.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest X-ray report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10348–10359, 2025b.

Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. HC-LLM: Historical-constrained large language models for radiology report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 5595–5603, 2025c.

Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, 2005.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest X-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633, 2023.

Frank Nielsen. Approximation and bounding techniques for the Fisher–Rao distances between parametric statistical models. In *Handbook of Statistics*, volume 51, pp. 67–116. Elsevier, 2024. doi: 10.1016/bs.host.2024.06.003.

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.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Julianna Pinele, João E. Strapasson, and Sueli I. R. Costa. The Fisher–Rao distance between multivariate normal distributions: special cases, bounds and applications. *Entropy*, 22(4):404, 2020.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NeurIPS*, 2007.

C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

Nagur Shareef Shaik, Teja Krishna Cherukuri, Adnan Masood, and Dong Hye Ye. DiA-gnostic VLVAE: Disentangled Alignment-Constrained Vision Language Variational AutoEncoder for Robust Radiology Reporting with Missing Modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8814–8823, 2026.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32, 2019.

Lene Theil Skovgaard. A Riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics*, 11:211–223, 1984.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y 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 (EMNLP)*, pp. 1500–1519, 2020a.

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. In *EMNLP*, 2020b.

Shanshan Song, Hui Tang, Honglong Yang, and Xiaomeng Li. DDaTR: Dynamic difference-aware temporal residual network for longitudinal radiology report generation. *IEEE Transactions on Medical Imaging*, 44(12):5345–5357, 2025.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

N. N. Čencov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, 1982.

Fuying Wang, Shenghui Du, and Lequan Yu. HERGen: Elevating radiology report generation with longitudinal data. In *European Conference on Computer Vision*, pp. 183–200. Springer, 2024.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1(3):100033, 2023.

Ziao Wang, Sixing Yan, Kejing Yin, Xiaofeng Zhang, and William K Cheung. CURV: Coherent Uncertainty-Aware Reasoning in Vision-Language Models for X-Ray Report Generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhipeng Xiang, Shaoguo Cui, Caozhi Shang, Jingfeng Jiang, and Liqiang Zhang. GMoD: Graph-driven Momentum Distillation Framework with Active Perception of Disease Severity for Radiology Report Generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15005 of *Lecture Notes in Computer Science*, pp. 295–305. Springer, 2024.

Sixing Yan, Ziao Wang, Kejing Yin, William K Cheung, Ka Chun Cheung, and Simon See. Learning Self-Critiquing Mechanisms for Region-Guided Chest X-Ray Report Generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16416–16424, 2024.

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, et al. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4(9):100802, 2023.

Hannah Yun, Junyeong Maeng, Eunsong Kang, and Heung-Il Suk. Diff-RRG: Longitudinal disease-wise patch difference as guidance for LLM-based radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 152–161. Springer, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *NeurIPS*, 2019.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M Summers, and Zhiyong Lu. Utilizing longitudinal chest X-rays and reports to pre-fill radiology reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 189–198. Springer, 2023.

A GEOMETRY OF THE SURROGATE DISTANCE

Proposition 1 (Surrogate geometry). *d_κ^2 is symmetric, non-negative, and zero iff $q_A = q_B$. To second order it equals the squared geodesic distance of $g_\kappa = \mathrm{diag}(\sigma^{-2}, \frac{\kappa}{2}\sigma^{-2})$ on each (μ_d, σ_d) half-plane. Since the Fisher metric is $g_{\mathrm{FR}} = \mathrm{diag}(\sigma^{-2}, 2\sigma^{-2})$, we have $g_\kappa = g_{\mathrm{FR}}$ iff $\kappa = 4$. For $\kappa \neq 4$, d_κ^2 matches the exact squared Fisher–Rao distance to leading order along pure-mean displacements but scales pure-variance displacements by $\kappa/4$, so for $\kappa < 4$ it is a local* lower *bound in that direction.*

Proof. **Symmetry** is immediate since $\bar{\sigma}_d^2$ and $\sigma_{A,d}\sigma_{B,d}$ are symmetric in (A, B). **Non-negativity and identity of indiscernibles**: the mean term is a square over a positive quantity; the logarithmic term is non-negative because $\bar{\sigma}_d^2 = \frac{1}{2}(\sigma_{A,d}^2 + \sigma_{B,d}^2) \geq \sigma_{A,d}\sigma_{B,d}$ by the arithmetic–geometric mean inequality, with equality iff $\sigma_{A,d} = \sigma_{B,d}$. The sum vanishes iff every term vanishes, i.e. iff $\mu_A = \mu_B$ and $\sigma_A = \sigma_B$. Both terms are dimensionless.

Second-order form. Fix a coordinate d and drop it. Write $\mu_A = \mu + h/2$, $\mu_B = \mu - h/2$, $\sigma_A^2 = \theta + \delta/2$, $\sigma_B^2 = \theta - \delta/2$ with $\theta = \sigma^2 > 0$, so that $\bar{\sigma}^2 = \theta$. The mean term is h^2/θ exactly. For the logarithmic term,

$$\kappa \log \frac{\theta}{\sqrt{(\theta + \delta/2)(\theta - \delta/2)}} = -\frac{\kappa}{2} \log\left(1 - \frac{\delta^2}{4\theta^2}\right) = \frac{\kappa\,\delta^2}{8\theta^2} + O(\delta^4),$$

giving $d_\kappa^2 = h^2/\theta + \kappa\delta^2/(8\theta^2) + O(\|(h, \delta)\|^4)$.

Now consider the metric $g_\kappa = \mathrm{diag}(\sigma^{-2}, \frac{\kappa}{2}\sigma^{-2})$ in the coordinates (μ, σ). Its squared geodesic distance between nearby points equals the squared line element to second order, $h^2/\sigma^2 + \frac{\kappa}{2}(\Delta\sigma)^2/\sigma^2$. With $\sigma = \sqrt{\theta}$,

$$\Delta\sigma = \sqrt{\theta + \delta/2} - \sqrt{\theta - \delta/2} = \sqrt{\theta}\Big[\big(1 + \tfrac{\delta}{2\theta}\big)^{1/2} - \big(1 - \tfrac{\delta}{2\theta}\big)^{1/2}\Big] = \frac{\delta}{2\sqrt{\theta}} + O(\delta^3),$$

where the remainder is $O(\delta^3)$ rather than $O(\delta^2)$ because the expansion is odd in δ: every even-order term of the two square roots cancels. Squaring therefore gives $(\Delta\sigma)^2 = \delta^2/(4\theta) + O(\delta^4)$ with no $O(\delta^3)$ cross-term, and the variance contribution is $\frac{\kappa}{2} \cdot \frac{\delta^2}{4\theta} \cdot \frac{1}{\theta} = \kappa\delta^2/(8\theta^2)$, matching the expansion above. Hence d_κ^2 is the squared geodesic distance of g_κ to second order.

Relation to Fisher–Rao. For $p(x; \mu, \sigma) = \mathcal{N}(x; \mu, \sigma^2)$ the score components are $\partial_\mu \log p = (x - \mu)/\sigma^2$ and $\partial_\sigma \log p = -1/\sigma + (x - \mu)^2/\sigma^3$, giving $I(\mu, \sigma) = \mathrm{diag}(\sigma^{-2}, 2\sigma^{-2}) = g_{\mathrm{FR}}$. Comparing with g_κ, the $\mu\mu$ entries agree for every κ and the $\sigma\sigma$ entries agree iff $\kappa/2 = 2$, i.e. $\kappa = 4$. For a pure variance displacement the exact Fisher–Rao expansion is $\delta^2/(2\theta^2)$ while the surrogate gives $\kappa\delta^2/(8\theta^2)$, a ratio of $\kappa/4$; for $\kappa < 4$ the surrogate is therefore a local lower bound in that direction. The product structure of the diagonal Gaussian manifold makes squared distances additive across coordinates, so the per-coordinate statement extends to the sum in Eq. (2). □

B MASKED DISTANCE, AGGREGATION, AND REDUCTION

Proposition 3 (Local geometry of the fixed-mask score). *Fix weights $w \in (0, 1)^D$. Then $d_{\kappa,\mathrm{M}}^2(q_A, q_B) = \sum_d w_d e_d^2$, where e_d^2 is the d-th bracket of Eq.* (2)*, is symmetric, non-negative, and zero iff $q_A = q_B$; and it is the squared geodesic distance to second order of the metric* $\mathrm{diag}(w)\, g_\kappa$*, which is positive definite because $w_d > 0$.*

Proof. Each $e_d^2 \geq 0$ by Prop. 1, and $w_d > 0$ preserves both non-negativity and the equality condition. Diagonal rescaling by strictly positive factors preserves positive definiteness and smoothness of the metric tensor, and the second-order expansion of its geodesic reproduces $\sum_d w_d e_d^2$ by the same computation as in Appendix A. □

Remark 1. In Dist-RRG, $w = w_{m,t}$ is produced from the report trajectory (Eq. (3)) and therefore depends on which prior is being scored. Prop. 3 is a statement about the geometry induced for a fixed w; across priors the mask changes, so the scores are comparisons under trajectory-conditioned metrics rather than distances in one fixed metric. We state this rather than describing the attention as computing a single global distance. Because w is zero-initialised through $\mathrm{sig}(0) = 0.5$, the model starts from a uniform mask and any departure is learned.

Proposition 2 (Forward-KL projection and variance preservation)**.** *Eq.* (6) *is the unique minimiser over diagonal Gaussians q of the forward (inclusive) divergence $D_{\mathrm{KL}}(p_{\mathrm{mix}} \,\|\, q)$, where $p_{\mathrm{mix}} = \sum_t \alpha_{T,t}\, \mathcal{N}(\mu_{b,t}, \mathrm{diag}\sigma^2_{b,t})$. Consequently $\hat{\sigma}^2_{b,T} \geq \sum_t \alpha_{T,t}\sigma^2_{b,t}$, with equality if and only if all attended means coincide.*

Proof. Let $p_{\mathrm{mix}} = \sum_t \alpha_t\, \mathcal{N}(\mu_t, \mathrm{diag}\sigma_t^2)$ and let $q = \mathcal{N}(\nu, \mathrm{diag}s^2)$ range over diagonal Gaussians. Diagonal Gaussians form an exponential family with sufficient statistics (x, x^2) per coordinate, so $D_{\mathrm{KL}}(p_{\mathrm{mix}}\|q) = -H(p_{\mathrm{mix}}) - \mathbb{E}_{p_{\mathrm{mix}}}[\log q]$ and only the second term depends on q. Minimising it is maximising $\mathbb{E}_{p_{\mathrm{mix}}}[\log q]$, which for an exponential family is uniquely achieved by matching the expected sufficient statistics, i.e. $\nu = \mathbb{E}_{p_{\mathrm{mix}}}[X]$ and $s^2 = \mathrm{Var}_{p_{\mathrm{mix}}}(X)$ coordinatewise. The tower rule gives $\mathbb{E}[X] = \sum_t \alpha_t\mu_t$ and, with J the mixture-component index (distinct from K, the number of available priors), the law of total variance gives $\mathrm{Var}(X) = \mathbb{E}_J[\mathrm{Var}(X \mid J)] + \mathrm{Var}_J(\mathbb{E}[X \mid J]) = \sum_t \alpha_t\sigma_t^2 + \sum_t \alpha_t(\mu_t - \hat{\mu})^2$, which is Eq. (6). The second statement follows by subtracting the within-prior term: the remainder is a weighted second moment, non-negative, and zero iff $\mu_t = \hat{\mu}$ for every t with $\alpha_t > 0$. □

Remark 2 (On the independence assumption)**.** Consecutive studies are correlated, so Eq. (1) is an approximation, but its direction is determined: $\mathrm{Var}(Z_{m,T} - Z_{m,t}) = \sigma^2_{m,T} + \sigma^2_{m,t} - 2\mathrm{Cov}(Z_{m,T}, Z_{m,t})$, so under positive cross-visit covariance the independent form *over*-states change variance and is conservative rather than over-confident. We use only the sign, not the magnitude, and do not claim the inflation is a calibrated safety margin; Table 8 sweeps an assumed ρ.

Remark 3 (Aggregating changes versus prior states)**.** For $m \in \{v, l, s\}$ the attended blocks are the change distributions $q_{m,\Delta t}$ of Eq. (1), whose means share the common term $\mu_{m,T}$. With $\sum_t \alpha_{T,t} = 1$,

$$\bar{\mu}_{m,T} := \mu_{m,T} - \hat{\mu}_{m,T} = \sum_t \alpha_{T,t}\, \mu_{m,t}, \qquad \mu_{m,\Delta t} - \hat{\mu}_{m,T} = \bar{\mu}_{m,T} - \mu_{m,t}, \tag{8}$$

so the aggregated prior state is recovered exactly from the aggregated change, and the between-prior variance of Eq. (6) is identical under either convention. Only the sign of the mean distinguishes them; the paper aggregates changes and uses $\bar{\mu}_{m,T}$ wherever a prior state is required (Eq. (12)).

Proposition 4 (Exact single-visit reduction)**.** *Under the $K = 0$ branch of §3.6, the forward pass and the Stage-B objective of Dist-RRG coincide numerically with those of the underlying single-visit reporter.*

Proof. The empty-history branch bypasses temporal residual fusion at every decoder layer, so each layer equals its single-visit counterpart. All history-dependent penalties, Eqs. (10)–(13), are multiplied by $\mathbf{1}[K \geq 1]$ and vanish; in addition $z_{\mathrm{temp},T} = 0$ on this branch, so the projection removal in Eq. (9) is the identity and the likelihood conditions on $Z_{s,T}$ itself. Thus the forward pass and objective reduce to those of the single-visit reporter. No assertion about biased attention on zero inputs is needed. □

C MISSING MODALITIES

Proposition 5 (Validity under partial availability)**.** *Assume (i) each block $b \in \mathcal{B}$ has a prior $p(Z_b)$ and, when b is unavailable at prior t, its variational factor is set to a learned distribution $\mathcal{N}(m_b, \mathrm{diag}\sigma^2_{\varnothing,b})$ with $\sigma^2_{\varnothing,b} > 0$ that does not depend on the missing input, and (ii) the observed and missing blocks are conditionally independent given the remaining latent state. Let $\mathcal{O}$ be any availability configuration and $p_\theta(R_T \mid \mathcal{O}) = \int p_\theta(R_T \mid Z_\mathcal{O}, Z_{\bar{\mathcal{O}}})\, p(Z_{\bar{\mathcal{O}}})\, q_\phi(Z_\mathcal{O} \mid \mathcal{O})\, dZ$ the marginal likelihood in which every unavailable block $Z_{\bar{\mathcal{O}}}$ is integrated out under its prior. Then $-\big(\mathcal{L}^{\mathrm{prog}}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{ELBO}}\big)$ is an evidence lower bound on $\log p_\theta(R_T \mid \mathcal{O})$, and the Stage-B objective at $\lambda_{\mathrm{CE}} = \lambda_{\mathrm{E}} = 1$ equals the negative of that bound plus auxiliary penalties with non-negative weights.*

Proof. Under (i) the variational factor for a missing block is a fixed distribution independent of the unobserved data, so it is a legitimate element of the variational family for the marginal model in which that block is integrated out. Under (ii), integrating the missing block out of the generative model yields the marginal conditional $p_\theta(R_T \mid Z_\mathcal{O}) = \int p_\theta(R_T \mid Z_\mathcal{O}, Z_{\bar{\mathcal{O}}})\, p(Z_{\bar{\mathcal{O}}})\, dZ_{\bar{\mathcal{O}}}$, a mixture over the prior of the missing block. This marginal is in general not equal to the full

conditional at any fixed value of $Z_{\bar{\mathcal{O}}}$; it is the quantity the bound targets, and (ii) is what allows the prior, rather than a posterior over the missing block, to appear in the mixture. The temporal posterior $\hat{q}_T$ (§3.6) factorises over blocks and uses the null factor of (i) for every missing block, so Jensen's inequality applied to $\log \mathbb{E}_{\hat{q}_T}\big[\, p_\theta(R_T \mid Z_T, z_{\text{temp}})\, p(z_{\text{temp}})/\hat{q}_T(z_{\text{temp}})\big]$ gives $\log p_\theta(R_T \mid \mathcal{O}) \geq \mathbb{E}_{\hat{q}_T}[\log p_\theta(R_T \mid Z_T, z_{\text{temp}})] - D_{\text{KL}}(\hat{q}_T \,\|\, p)$, whose two terms are $-\mathcal{L}_{\text{CE}}^{\text{prog}}$ and $-\mathcal{L}_{\text{ELBO}}$ of Appendix D; the per-visit factors Z_T are drawn from the frozen q_ϕ, whose own KL is a Stage-A constant. The remaining Stage-B terms have non-negative weights, which proves the second statement. □

Remark 4 (Practical objective)**.** The deployed weights are $\lambda_{\text{CE}} = 1$ and $\lambda_{\text{E}} = 10^{-2}$, a β-weighting of the KL term in the sense of Higgins et al. (2017). With $\beta < 1$ the weighted objective is not itself a bound on $\log p_\theta(R_T \mid \mathcal{O})$; Prop. 5 identifies which terms constitute the bound and under which assumptions, and Stage B optimises the β-weighted surrogate of it.

Remark 5. Assumption (ii) is a modelling choice, not a fact about chest radiography: a missing indication is not independent of the finding it would have described. We state it because the bound depends on it, and §4.5 measures what actually happens empirically when inputs are removed.

D OBJECTIVE TERMS IN FULL

Stage A fits the per-visit encoder with its own objective (Shaik et al., 2026). Stage B freezes it, so those terms are constant and are excluded; only the terms below carry gradient.

Notation: $Z_{m,T} \sim q_{m,T}$ are the per-visit factors of the target, $z_{\text{temp},T} = [z_{b,T}]_{b\in\mathcal{B}}$ is the temporal latent with law $\hat{q}_T = \prod_b \hat{q}_{b,T}$, $\hat{q}_{b,T} = \mathcal{N}(\hat{\mu}_{b,T}, \text{diag}\hat{\sigma}^2_{b,T})$ (§3.6), and $z_{s,T}$ denotes the shared-change block of $z_{\text{temp},T}$, which lives in the same D_s-dimensional coordinate system as $Z_{s,T}$ because it is an aggregate of shared-factor differences (§3.4). The block priors are those of the surrogate prior below, $p(Z_b) = \mathcal{N}(0, s_b^2 I)$ with $s_b^2 = 2$ for $b \in \{v, l, s\}$ and $s_R^2 = 1$.

Progression-aware likelihood.

$$\mathcal{L}_{\text{CE}}^{\text{prog}} = -\sum_c \log p_\theta\big(r_{T,c} \mid r_{T,<c}, Z_{v,T}, Z_{l,T}, \tilde{Z}_{s,T}, z_{\text{temp},T}\big), \qquad \tilde{Z}_{s,T} = Z_{s,T} - \frac{\langle Z_{s,T},\, z_{s,T}\rangle}{\|z_{s,T}\|^2 + \epsilon}\, z_{s,T}, \tag{9}$$

where the conditioning on $z_{\text{temp},T}$ is through the memory M_{temp} of Eq. (7). The projection removal is a single Gram–Schmidt step: $\tilde{Z}_{s,T}$ is the component of the current shared factor orthogonal to the sampled shared-change direction $z_{s,T}$, with $\epsilon = 10^{-6}$ (Table 13) and gradient flowing through both arguments. No learned map is involved, and on the $K = 0$ branch, where $z_{\text{temp},T} = 0$, the operation is the identity. The tilde is used for this residualised quantity only; $z_{\text{temp},T}$ itself is never residualised.

Temporal-posterior KL.

$$\mathcal{L}_{\text{ELBO}} = \sum_{b\in\mathcal{B}} D_{\text{KL}}\big(\hat{q}_{b,T} \,\|\, p(Z_b)\big)\cdot\mathbf{1}[K \geq 1] = \frac{1}{2}\sum_{b\in\mathcal{B}}\sum_{d=1}^{D_b}\left[\frac{\hat{\sigma}^2_{b,T,d} + \hat{\mu}^2_{b,T,d}}{s_b^2} - 1 - \log\frac{\hat{\sigma}^2_{b,T,d}}{s_b^2}\right]\cdot\mathbf{1}[K \geq 1]\,. \tag{10}$$

This is the KL term of the bound in Prop. 5. The variational distribution is the moment-matched temporal posterior $\hat{q}_T$ rather than a separate inference network conditioned on R_T: its parameters are the outputs of MMDA, and they are trainable through every Stage-B module that feeds the attention (§3.6). The prior is the same no-change reference used by the surrogate prior, so the two penalties agree on the target and differ only in geometry (KL versus d_κ^2) and weight (10^{-2} versus 0.1).

Surrogate prior.

$$\mathcal{L}_{\text{FR}} = \frac{1}{|\mathcal{B}|}\Big[\sum_{m\in\{v,l,s\}} d_\kappa^2\big(\hat{q}_{m,T}, \mathcal{N}(0, 2I)\big) + d_\kappa^2\big(\hat{q}_{R,T}, \mathcal{N}(0, I)\big)\Big]\cdot\mathbf{1}[K \geq 1]\,. \tag{11}$$

Both arguments of d_κ^2 are distributions, as in Eq. (2). The references differ because, by the block convention of §3.4, $\hat{q}_{m,T}$ for $m \in \{v, l, s\}$ is an aggregate of *change* distributions: under standard-normal per-visit priors the difference of two visits has variance 2, and a zero mean is the no-change

state. Pulling $\hat{q}_{m,T}$ toward $\mathcal{N}(0, 2I)$ therefore regularises the aggregated change toward "no change" and leaves the implied aggregated prior state $\bar{\mu}_{m,T} = \mu_{m,T} - \hat{\mu}_{m,T}$ of Eq. (8) free to track the current visit; it does not pull historical memory toward zero. The report block $\hat{q}_{R,T}$ is an aggregate of absolute states and goes toward $\mathcal{N}(0, I)$.

Progression consistency.

$$\mathcal{L}_{\mathrm{PC}} = \frac{1}{14} \sum_{\ell=1}^{14} \mathrm{CE}\Big(g_\psi\big([\mu_{s,T};\, \bar{\mu}_{s,T};\, \hat{\mu}_{s,T}]\big)_\ell,\; c_{\ell,T} \Big) \cdot \mathbf{1}[K \geq 1]\,, \tag{12}$$

with $c_{\ell,T}$ the CheXbert transition label against the most recent available prior (Appendix E). The three arguments are the current shared mean, the implied aggregated prior state, and the aggregated change; by Eq. (8) the last two are $\mu_{s,T} - \hat{\mu}_{s,T}$ and $\hat{\mu}_{s,T}$, so this is the same set of quantities as "current, aggregated prior, difference". A relational class cannot be read from the aggregated history alone, which is why all three appear.

The input is trajectory-level while the label is single-step, and we pair them deliberately. The reference report's comparative statements are written against the previous study — "as compared to the previous radiograph" — so the most recent prior is the only frame in which a transition label is defined, whereas the decoder conditions on the attended aggregate. The head therefore asks whether the representation the decoder actually receives suffices to recover the transition the report will describe; it is a consistency constraint on the aggregate, not an estimator of a pairwise delta from a pairwise input. Two facts keep this well posed. Under P1 ($K = 1$) the aggregate and the most recent prior coincide, so input and label frames agree exactly. Under P2, attention is free to concentrate on the most recent prior, and the RTAM gate is computed from the report trajectory ending at that prior; when it instead spreads over disagreeing priors, MMDA reports the disagreement as variance, which is the case the calibration analysis of §4.4 examines. A head supervised from the most recent prior's change block alone is a natural variant and is not evaluated here (§5).

Residual independence.

$$\mathcal{L}_{\mathrm{HSIC}} = \sum_{m \in \{v,l,s\}} \widehat{\mathrm{HSIC}}\big(Z_{m,T}, z_{\mathrm{temp},T}\big) \cdot \mathbf{1}[K \geq 1]\,, \qquad \widehat{\mathrm{HSIC}}(X, Y) = \tfrac{1}{(N-1)^2} \mathrm{tr}(KHLH), \tag{13}$$

with N the minibatch size ($N = 4$, distinct from $|\mathcal{B}| = 4$ in Eq. (11)), $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$, and Gram matrices K, L from RBF kernels whose bandwidth is the median pairwise distance in the minibatch. The second argument is the temporal latent $z_{\mathrm{temp},T}$ as delivered to the decoder, without a tilde; the indicator is written explicitly so that Eqs. (10)–(13) carry the same gate and Prop. 4 needs no separate argument for this term.

E Evaluation protocol

Counting targets. The test split has 266 patients and 2,324 study visits. Each patient's earliest study has no preceding study, so the number of longitudinal targets is $2{,}324 - 266 = 2{,}058$. Each target contributes 14 CheXbert transition labels, giving 28,812 target–pathology decisions.

Transition labels. For each of the 14 findings we take the CheXbert label in the reference report of the target study and in the reference report of the most recent available prior, and map the pair to *stable*, *new*, or *resolved*. The same derivation is applied to the generated report. CheXbert emits one of four classes per finding — Positive, Negative, Uncertain, Blank — so there are 16 prior–current pairs. We first binarise each label with the same convention used for the presence-level CE-F1 of Table 1: Positive $\mapsto 1$, and Negative, Uncertain, and Blank $\mapsto 0$ (uncertain treated as absent). The transition is then read from the binary pair: $(0, 0)$ and $(1, 1)$ are *stable*, $(0, 1)$ is *new*, and $(1, 0)$ is *resolved*. Table 4 spells out all 16 cases; in particular, Uncertain→Positive is *new*, Positive→Uncertain is *resolved*, and any pair that stays within {Negative, Uncertain, Blank} is *stable*. The same map produces the training targets $c_{\ell,T}$ of Eq. (12). One-vs-rest F1 is computed per class. These are progression-task scores over the 28,812 transition labels and are not comparable to the 14-pathology CE-F1 of Table 1, which scores presence rather than change.

Table 4: Mapping of the 16 CheXbert (prior, current) label pairs to transition classes. Rows: label in the most recent available prior; columns: label in the current study. P = Positive, N = Negative, U = Uncertain, B = Blank.

Prior \ Current	P	N	U	B
P	stable	resolved	resolved	resolved
N	new	stable	stable	stable
U	new	stable	stable	stable
B	new	stable	stable	stable

Calibration. Progression probabilities are binned into 15 equal-width confidence bins. Aggregate ECE treats the three-class prediction directly; per-class ECE uses a one-vs-rest formulation with the same bins, and macro ECE is the unweighted mean over the three classes.

RadGraph-F1. RG-F1 compares entities and relations extracted by RadGraph (Jain et al., 2021) from the generated and reference findings sections, following Delbrouck et al. (2022); we report the entity–relation score averaged over targets. It involves no CheXbert component, so it is independent of the label source that supervises the progression head. The same extractor and settings score every M and R row.

Bootstrap. We resample the 266 test patients with replacement, retain every target belonging to each sampled patient, and recompute the metric on the resampled set; reported intervals are the 2.5–97.5 percentiles over 10000 resamples. Resampling at the patient level rather than the target level is necessary because targets from one patient share history and are not independent. For the P2 checkpoint this gives a marginal CE-F1 interval of $[0.496, 0.536]$; it quantifies test-set sampling variability of one system and is not the yardstick for a paired difference.

Paired bootstrap and training seeds. For a difference between two systems evaluated on the same targets, each resample draws one set of patients, scores both systems on it, and records the difference; the reported interval is the 2.5–97.5 percentile range of those 10000 differences. Pairing removes the between-patient variance shared by both systems, which the marginal interval of one system includes. Paired intervals are computed at seed 0. Training-seed variability is reported separately as the mean and standard deviation over seeds $\{0, 1, 2\}$, each trained with the fixed configuration of Table 13; the two sources of variability are not combined into one interval.

F BASELINE PROVENANCE

Table 1 distinguishes source results from our runs. S denotes values tabulated in secondary benchmark comparisons; O denotes the original TIM result; M denotes matched reruns; R denotes our P1/P2 runs. For S and O rows the input protocol specifies single-prior conditioning, but does not imply identical backbones, preprocessing, tokenizers, or clinical metric averaging. Published source values are retained without using them to establish a common-convention ranking.

Matched reruns. Each M row is retrained from the authors' official released code on the patient-level split of Zhu et al. (2023) under protocol P1, with the hyperparameters reported in its source paper, model selection on our validation split only, and no access to the test split before final scoring. Generated findings sections are scored with the same script as Dist-RRG: identical report preprocessing and BLEU tokenisation, CheXbert binarisation of Appendix E, micro-averaged CE-F1, and the same RG-F1 configuration. Each M row is a single training run. We use the latest available commits from the official repositories for Prefilling, HERGen, and Diff-RRG, with batch sizes 16, 8, and 16 respectively. These batch sizes match the default configurations reported in their source papers, utilizing gradient accumulation where necessary to fit the memory constraints of a single NVIDIA A40 GPU.

HC-LLM and TIM. HC-LLM (Liu et al., 2025c) reports BLEU-4 0.128 and CE-F1 0.357 for its R2GenGPT configuration. Table 1 instead uses the BiomedCLIP + BioMistral-7B comparison

in Diff-RRG (Yun et al., 2025), with BLEU-4 0.116 and CE-F1 0.448; TIM (Dong et al., 2026) tabulates the same row. These are secondary comparisons, not baseline reruns performed in this work. Neither HC-LLM nor TIM is rerun: TIM is concurrent work (§2), and both enter Table 1 only through published values.

Methods cited but not tabulated. MLRG (Liu et al., 2025b) reports on MIMIC-CXR, MIMIC-ABN, and Two-view CXR and conditions the current visit on multiple views; it does not report on the patient-level Longitudinal MIMIC-CXR split of Zhu et al. (2023), and its input protocol differs from P1 in the current-visit input as well as in the history. DDaTR (Song et al., 2025) evaluates on the standard MIMIC-CXR benchmark and on its own longitudinal subsets defined by the proportion of studies with a prior, not on the Zhu et al. (2023) split. RECAP (Hou et al., 2023) predates that split and reports on MIMIC-CXR and MIMIC-ABN. For none of the three does a published number exist that shares the split and the single-prior protocol of the P1 rows, so tabulating their published values would add a second axis of incomparability to the averaging-convention one the table already carries. A matched rerun of MLRG would additionally require removing its multi-view current input, which changes the method; these three remain uncompared on our split (§5).

Table 5: Complete-model removals and substitutions (P2). CE-F1 and Δ are at seed 0, the fixed checkpoint; the paired 95% interval is for Δ on the same 2,058 targets (Appendix E). The last two columns give mean $\pm$ s.d. over three training seeds where trained; other rows are single runs. "MMDA disabled" is the configuration of Table 10.

	Seed 0			Three seeds	
Configuration	CE-F1	Δ CE-F1	Paired 95% CI	CE-F1	ECE
Full Dist-RRG	0.516	–	–	0.515 ± 0.003	0.042 ± 0.002
MMDA disabled	0.489	-0.027	$[-0.042, -0.015]$	0.487 ± 0.004	0.064 ± 0.004
Euclidean attention	0.503	-0.013	$[-0.024, -0.003]$	0.501 ± 0.005	0.052 ± 0.003
w/o RTAM	0.507	-0.009	$[-0.018, -0.001]$	–	–
w/o RFF	0.510	-0.006	$[-0.014, 0.002]$	–	–
w/o FR prior	0.508	-0.008	$[-0.017, 0.001]$	–	–

G ADDITIONAL RESULTS

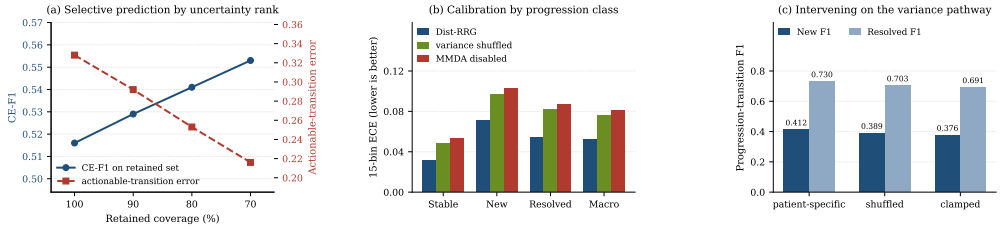


Figure 2: Operational evidence for the propagated variance (§4.4). (a) Deferring the most uncertain targets raises CE-F1 and lowers the actionable-transition error rate monotonically (Table 11). (b) Disabling MMDA degrades calibration on all three classes, not only the dominant stable one (Table 10). (c) Permuting the variance across patients, or clamping it, costs accuracy on exactly the two actionable classes (Table 9).

Rare findings. The five CheXbert pathologies with the lowest positive reference support were fixed a priori. Macro-F1 on that subset improves from 0.288 for the frozen per-visit baseline to 0.319 for full Dist-RRG.

Reference-report ambiguity. Dimension-normalised aggregated shared-factor variance correlates weakly with the CheXbert uncertain-finding count ($\rho = 0.274$, nominal $p = 1.9 \times 10^{-35}$) and an explicit hedging indicator ($\rho = 0.219$, nominal $p = 3.0 \times 10^{-23}$). These markers measure

Table 6: Input availability at inference (P2). Random-removal rows use five fixed inference-mask seeds and report mean $\pm$ s.d.; these are not training seeds.

Condition	CE-F1
Full input	0.516
No current indication	0.507
No prior indications	0.511
50% of prior reports removed	0.505 ± 0.003
All prior reports removed	0.493
50% of historical visits removed	0.501 ± 0.004

Table 7: Interval to the immediately preceding study. Median 12 days, maximum 1,644 days.

Interval	n	CE-F1
0–89 days	1,606	0.524
90–364 days	331	0.519
365–729 days	83	0.510
$\geq$ 730 days	38	0.498
Total	2,058	0.516

Table 8: Sensitivity to the independence approximation: assumed cross-visit correlation ρ applied elementwise as $\sigma^2_\Delta(\rho) = \sigma^2_T + \sigma^2_t - 2\rho\sigma_T\sigma_t$ before aggregation. This inference intervention does not estimate correlation.

ρ	CE-F1	New F1	Res. F1	ECE
0.00	0.516	0.412	0.730	0.041
0.25	0.514	0.408	0.725	0.043
0.50	0.510	0.399	0.714	0.048
0.75	0.504	0.386	0.698	0.054

documented report ambiguity, not diagnostic confidence; targets share patient histories, so the nominal p-values do not establish patient-independent significance.

Table 9: The variance intervention (P2), numeric form of Fig. 2(c). Shuffling keeps the global variance distribution intact, so degradation cannot be attributed to reducing the amount of variance; it isolates whether the variance is patient-specific. Clamping removes patient-level variation entirely, using the validation-set mean to avoid test leakage.

Inference setting	CE-F1	New F1	Resolved F1	ECE
Patient-specific variance (Dist-RRG)	**0.516**	**0.412**	**0.730**	**0.041**
Variance shuffled across patients	0.507	0.389	0.703	0.056
Variance clamped to validation mean	0.503	0.376	0.691	0.061

Table 10: Calibration of the stable/new/resolved progression probabilities, 15-bin one-vs-rest ECE with the same bins for every class (seed 0). Disabling MMDA is the matched ablation that removes preservation of within-prior uncertainty and between-prior disagreement; its CE-F1 and three-seed ECE are in Table 5. Aggregate ECE scores the three-class prediction directly and differs from the mean of the one-vs-rest values.

Model	Stable	New	Resolved	Macro	Aggregate
Full Dist-RRG	**0.031**	**0.071**	**0.054**	**0.052**	**0.041**
Variance shuffled	0.048	0.097	0.082	0.076	0.056
MMDA disabled	0.053	0.103	0.087	0.081	0.063

Table 11: Selective prediction by uncertainty rank (numeric form of Fig. 2(a)). AUROC 0.712, AUPRC 0.536, positive prevalence 0.328.

Coverage	CE-F1	Actionable err.
100%	0.516	0.328
90%	0.529	0.292
80%	0.541	0.253
70%	0.553	0.216

Table 12: Aggregated variance as a detector of unreliable targets. A target is positive when at least one *new* or *resolved* transition is wrong.

Quantity	Value
AUROC	0.712
AUPRC	0.536
Positive prevalence	0.328
Targets scored	2,058

H Scope of the evaluation and planned external validation

Longitudinal MIMIC-CXR is a single-institution corpus, so the robustness results of §4.5 measure resilience to input availability and to irregular follow-up intervals *within* that corpus. They are not evidence of transfer across institutions or acquisition protocols, and we do not present them as such. No cross-dataset or cross-modality result is reported in this paper.

Two external evaluations are the natural next step, and we name them so the gap is explicit. MS-CXR-T (Bannur et al., 2023) carries expert-curated temporal progression labels on chest radiograph pairs, which matches the stable/new/resolved formulation of Table 3 directly and would test the progression head against human annotation rather than against CheXbert-derived labels. CheXpert Plus (Chambon et al., 2024) supplies reports from a different institution and would give a cross-site test of report generation under both zero-shot transfer and fine-tuning. Neither is evaluated here.

I Reader study protocol

The following is a proposed protocol, stated before data collection. Sample 200 target studies stratified by aggregated uncertainty decile and by whether the reference contains an actionable transition. Three board-certified radiologists, blinded to model identity, see the prior report, the current radiograph, and two candidate findings sections (reference and generated) in randomised order. Each rater records, per candidate: (a) whether each comparative statement is supported, contradicted, or unsupported by the images; (b) an error-severity grade for each unsupported statement; and (c) a forced-choice preference. We will report inter-rater agreement, per-category error rates, and the correlation between error severity and the model's uncertainty score u_i — the last is the test that would actually establish

clinical meaning for the variance, which §4.4 deliberately stops short of claiming. The study has not been run, and no reader-study result is claimed anywhere in this paper.

J IMPLEMENTATION DETAILS

Table 13: Fixed configuration. All values were selected by a coarse search on the validation split and then held fixed across every protocol, ablation, and training seed; no per-row retuning was performed.

Component	Setting
Per-factor latent dimension	256
LX-PRE	6 layers, 8 heads, hidden 512; trainable: null memory, 4 projections
RTAM	rank 32
Random Fourier features	512 features, Ω and b_ϕ frozen, W_ϕ zero-init, gate -5
Surrogate coefficient	$\kappa = 1$ (Prop. 1)
Temporal memory	2 tokens; fusion gates initialised at -1 (per decoder layer)
Report decoder	12 layers, 16 heads, hidden 1024
Max history	$H = 5$
Optimiser	AdamW, peak LR 5×10^{-5}, 10% warm-up, cosine decay
Schedule	25 epochs, batch size 4, gradient clipping 1.0
Objective weights	$(\lambda_{\mathrm{CE}}, \lambda_{\mathrm{E}}, \lambda_{\mathrm{FR}}, \lambda_{\mathrm{PC}}, \lambda_{\mathrm{H}}) = (1.0,\ 10^{-2},\ 0.1,\ 0.3,\ 0.3)$
Training seeds	$\{0, 1, 2\}$: full, MMDA disabled, Euclidean attn.; $\{0\}$: others
Numerical guards	log-variances clipped to $[-10, 10]$; $\epsilon = 10^{-6}$ inside every log
Trainable parameters	114M (per-visit encoder frozen)
Hardware	single NVIDIA A40; $\approx$ 366 GFLOPs and 0.18 s per sample at inference

K TRAINING STEP

Algorithm 1 specifies Stage-B optimization with availability masks and an explicit temporal-fusion bypass for an empty history.

Algorithm 1 Dist-RRG Stage-B training step for one patient record.

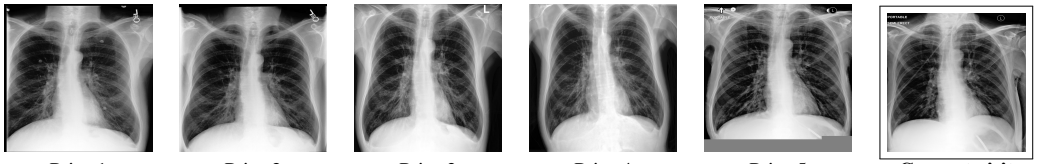
Require: Record $\{(V_t, L_t, R_t)\}_{t=1}^{T_i}$, availability $\{a_t\}$, frozen encoder q_ϕ, trainable parameters Θ.

for $T = 1$ to T_i **do**
  $\{q_{m,T}\}_{m\in\{v,l,s\}} \leftarrow q_\phi(V_T, L_T)$
  $K \leftarrow \min(T-1, H)$; apply availability dropout to $\{a_t\}_{t<T}$
  if $K = 0$ **then**
    $M_{\mathrm{temp}} \leftarrow \mathbf{0}$; $z_{\mathrm{temp},T} \leftarrow \mathbf{0}$ ▷ explicit branch, §3.6
  else
    for each retained prior t **do**
      $q_{m,\Delta t} \leftarrow \mathcal{N}(\mu_{m,T} - \mu_{m,t}, \mathrm{diag}(\sigma^2_{m,T} + \sigma^2_{m,t}))$ ▷ Eq. (1)
      $q_{R,t} \leftarrow$ LX-PRE(R_t) if a_t^R=1 else learned null
      compute $\Delta_{\mathrm{traj}}\mu_{R,t}$, then the RTAM gate $w_{m,t}$ ▷ Eqs. (3), (4)
      compute block-combined $d^2_{\kappa,\mathrm{M}}$ under availability and the random-Fourier score $s_{\mathrm{rff},t}$
    end for
    $\{\alpha_{T,t}\} \leftarrow$ softmax over retained priors ▷ Eq. (5)
    $(\hat{\mu}_{b,T}, \hat{\sigma}^2_{b,T}) \leftarrow$ MMDA over $\{q_{b,t}\}$; $\bar{\mu}_{s,T} \leftarrow \mu_{s,T} - \hat{\mu}_{s,T}$ ▷ Eqs. (6), (8)
    sample $z_{\mathrm{temp},T}$; build FiLM-modulated M_{temp}; form $\tilde{Z}_{s,T}$ ▷ Eqs. (7), (9)
  end if
  decode R_T; apply temporal residual fusion only if $K \geq 1$
end for
accumulate $\mathcal{L}$ (§3.7); backpropagate through Θ

L QUALITATIVE EVALUATION

Figure 3 compares one versus five priors for a patient with bronchiectasis and pseudomonas colonisation. The most recent prior suspects pneumonia. With that prior alone, Dist-RRG describes the opacity as "minimally decreased", an unsupported change. With five priors, it describes the chronic opacities as constant, matching the reference.

This case illustrates history sensitivity, not an isolated variance effect: additional report context also changes. No dimension-level latent analysis is available. Evidence for the variance pathway instead comes from the intervention in Table 9.



Prior 1 | Prior 2 | Prior 3 | Prior 4 | Prior 5 | **Current visit**

Attribute	**Clinical text**
Current indication	Bronchiectasis and pseudomonas, cough, persistent tachycardia, evaluation.
Historical context (prior indications and reports)	
Prior 1	**Ind:** [none provided] **Rpt:** The lungs remain hyperinflated, suggesting chronic obstructive pulmonary disease. No focal consolidation, pleural effusion, or evidence of pneumothorax is seen. The cardiac and mediastinal silhouettes are stable and unremarkable.
Prior 2	**Ind:** ___-year-old male with cough and fever. **Rpt:** No focal consolidation, pleural effusion, or pneumothorax is seen. Heart and mediastinal contours are within normal limits. Lungs are again noted to be hyperinflated.
Prior 3	**Ind:** [none provided] **Rpt:** Review of frontal and lateral views were remarkable for bilateral lower lung bronchiectasis with peribronchial opacities. In the right lower and medial lung, peribronchial opacities have improved... Lungs are mildly hyperinflated. Heart size, mediastinal and hilar contours are normal.
Prior 4	**Ind:** [none provided] **Rpt:** In comparison with the study of [DATE], there is little overall change in the peribronchial thickening and impaction with extensive bibasilar bronchiectasis... Hyperexpansion of the lungs is consistent with emphysema and the cardiac size is normal.
Prior 5	**Ind:** Known bronchiectasis and pseudomonas colonization, now hypoxia and cough with fevers. Assessment. **Rpt:** As compared to the previous radiograph, there is a subtle but new opacity at the right lung base... Given the clinical presentation, pneumonia must be suspected... Otherwise, the radiograph is unchanged, extensive overinflation with bronchiectasis but no pleural effusions.
Ground truth	**As compared to the previous radiograph there is no relevant change.** Moderate to severe overinflation with known areas of bronchiectasis and perifocal parenchymal opacities. **The opacities are unchanged in distribution and severity. Normal size of the cardiac silhouette.** Normal hilar and mediastinal structures.
Dist-RRG (K=1, Prior 5 only)	**As compared to the previous radiograph there is no relevant change.** The pre-existing parenchymal opacity at the right lung base **has minimally decreased in severity and extent**. No new parenchymal opacities. **Unchanged size of the cardiac silhouette.** Unchanged course of the left PICC line.
Dist-RRG (K=5, full history)	**As compared to the previous radiograph there is no relevant change.** The **pre-existing parenchymal opacities are constant in extent and distribution.** No new parenchymal opacities. No pleural effusions. **Normal size of the cardiac silhouette.**

Figure 3: Qualitative case. **(Top)** The patient's longitudinal sequence: five priors and the current study. **(Bottom)** With one prior, the model inherits Prior 5's suspicion of pneumonia and asserts a change the current study does not support. With the full trajectory it recovers the chronic, unchanged reading that matches the reference. Green marks statements consistent with the ground truth, red marks the unsupported change claim.