Rethinking Contrastive Targets in Radiology Language-Image Pretraining

Yunhe Gao, Ashwin Kumar, Jiaming Liu, Chong Wang, Maya Varma, Fuying Wang, Stefania L. Moroianu, Louisa Fay, Sergios Gatidis, Akshay S. Chaudhari, Curtis Langlotz

Center for Artificial Intelligence in Medicine and Imaging (AIMI)
Stanford University

Abstract

CLIP-style contrastive image–text pretraining is widely used in radiology data, where each image is paired with one report and every other report in the batch is treated as a negative. Under this recipe, two failures appear, neither of which is systematically addressed by existing contrastive learning theory: zero-shot evaluation peaks then degrades while training loss keeps decreasing, and training fails to learn from large batches. We hypothesize the cause is structural rather than radiology-specific: many unpaired reports share clinical findings with a given image, so CLIP's paired-identity target is systematically biased, not merely noisy. To test this, we construct *CLEVR-Radiology*, a synthetic testbed that preserves only this overlapping-fact structure and reproduces both failures. With evidence that the structure is sufficient to reproduce the failures, we develop a target-misspecification theory that explains both failures from a single deterministic gradient bias amplified by batch size, and use it to motivate a minimal compatibility-based correction, *CorLIP*: text is treated as positive whenever its clinical assertions agree with the image, with no architecture changes. Beyond benchmark improvements, CorLIP largely removes the post-peak degradation, makes large-batch training succeed where it previously collapsed, substantially amplifies the gains from more data, and is the only configuration that remains stable up to $B{=}2048$. The correction transfers from chest X-ray to 3D CT modalities.

1 Introduction

Contrastive language–image pretraining (CLIP) has emerged as the dominant paradigm for learning general-purpose visual representations from paired image–text data [1], and the same contrastive objective is widely used in radiology vision–language pretraining (VLP) across imaging modalities, including X-ray [2–7], CT [8–10], MRI [11] and multi-modality images [12, 13]. These systems usually inherit a paired-identity assumption from CLIP: within a batch, the paired text is the only positive for an image, and every other text is treated as a negative. This assumption is approximately correct for web image–caption data, where captions and image content are less likely to overlap, but its validity on radiology image–report data has not been systematically examined. Prior work tends to attribute medical CLIP's difficulties to a broad medical-vs-general domain gap [14–16]. We argue this gap has a precise and identifiable structure that helps explain why the standard recipe is systematically biased.

We systematically study the training dynamics of medical CLIP and document two failure modes that current contrastive theory does not address. First, the *rise-then-fall dynamic*: zero-shot performance rises, peaks, and then degrades while training loss decreases monotonically, even though a supervised classifier on the identical backbone shows no late-training drop, ruling out generic overfitting. Second,

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

large-batch collapse: training fails to learn (AUC near random, loss near $\log B$) at batch sizes that are routinely beneficial for general-domain CLIP [1, 17, 18]. Existing literature has noted pieces of these phenomena without systematic analysis [19–21, 10] or attributed the rise-then-fall pattern to overfitting [22], but neither failure has a unified explanation. Together, these upstream failures cap downstream performance and prevent medical CLIP from leveraging the growing scale of image–report data.

These observations point to a single suspect: the contrastive target. We hypothesize that both failures arise from a structural property of the data rather than from medical-specific appearance, language, or annotation. Specifically, radiology image–report pairs share what we call a *shared-schema fact-bundle structure*: anatomy is the same across images (lungs, heart, mediastinum, pleura), each anatomical region carries a finding state drawn from a small clinical vocabulary (e.g., effusion present or absent, cardiomegaly present or absent), and each report bundles assertions about many such states simultaneously. Under this structure, multiple unpaired reports can match a given image on most asserted findings, so paired identity is systematically biased rather than merely noisy. The hypothesis is falsifiable: any data sharing the same three ingredients should fail in the same way, regardless of domain. To test this, we construct *CLEVR-Radiology* [23], a synthetic testbed that retains the three ingredients but strips every radiology-specific factor such as image appearance, clinical jargon, annotation noise, and scale. Both failures recur identically on CLEVR-Radiology, supporting the hypothesis that the structure, rather than radiology-specific artifacts alone, can drive the failures. We then formalize this as a target-misspecification theory that derives both failures from a single deterministic gradient bias amplified by batch size.

Guided by the theory, we propose *CorLIP*, a minimal correction that replaces CLIP's paired identity with a fact-compatibility target: a text is treated as positive whenever its clinical assertions agree with the image. The correction is compatibility-based rather than similarity-based: a text need not summarize the whole image or be globally similar to the paired report; it only needs to assert facts that are true of the image. This contrasts with prior medical contrastive approaches that mitigate false negatives or improve representations through similarity computations, knowledge integration, or auxiliary alignment [24–26, 20]. Beyond benchmark improvements on chest X-ray classification and grounding, three insights follow: the rise-then-fall dynamic largely disappears; large-batch training, previously seen as harmful for medical CLIP, succeeds and remains stable up to $B{=}2048$; and additional training data delivers substantial rather than marginal gains. The correction generalizes across radiology modalities, from chest X-ray (CXR) to 3D CT.

Contributions. *(i) Diagnostic.* We identify two failures of medical CLIP (rise-then-fall, large-batch collapse) and isolate their cause as structural via *CLEVR-Radiology*, a controlled synthetic testbed. *(ii) Theory.* A target-misspecification theory derives both failures from a single deterministic gradient bias, with two verified predictions: a quantitative loss-floor match and a mechanistic shortcut heatmap drift. *(iii) Correction.* We propose *CorLIP*[1], a minimal compatibility-based correction, which requires no change to architecture or optimizer, and only changes the batch target matrix constructed with an offline fact parser. Beyond benchmark improvements on chest X-ray classification and grounding, CorLIP recovers scaling regimes that do not work with vanilla medical CLIP: large-batch training succeeds where it previously collapsed (stable up to $B{=}2048$), additional data delivers substantial rather than marginal gains, and the correction generalizes from chest X-ray to 3D CT modalities.

2 Observations

We document the training dynamics of vanilla medical CLIP on MIMIC-CXR [27] as a representative radiology pretraining corpus. We sweep over batch size (moderate $B{=}128$, large $B{=}1024$) and text granularity (full report, impression section, individual sentence), using a standard recipe (ViT-B + BERT [28, 29] with vanilla symmetric CLIP loss). Zero-shot 14-class mean AUC [30, 31] measures downstream quality; a supervised multi-label classifier on the same backbone serves as a non-contrastive control. Full setup is in Appendix E.

Two failure modes. At moderate batch ($B{=}128$), zero-shot AUC rises then falls while training loss decreases monotonically; we call this **O1: rise-then-fall** (Figure 1(A)). At large batch ($B{=}1024$), AUC remains near random and loss stays close to $\log B$ throughout, with no learning; we call this **O2: large-batch collapse** (Figure 1(B)). Both patterns hold across all three text granularities.

[1] Project repository: `https://github.com/yhygao/CorLIP`

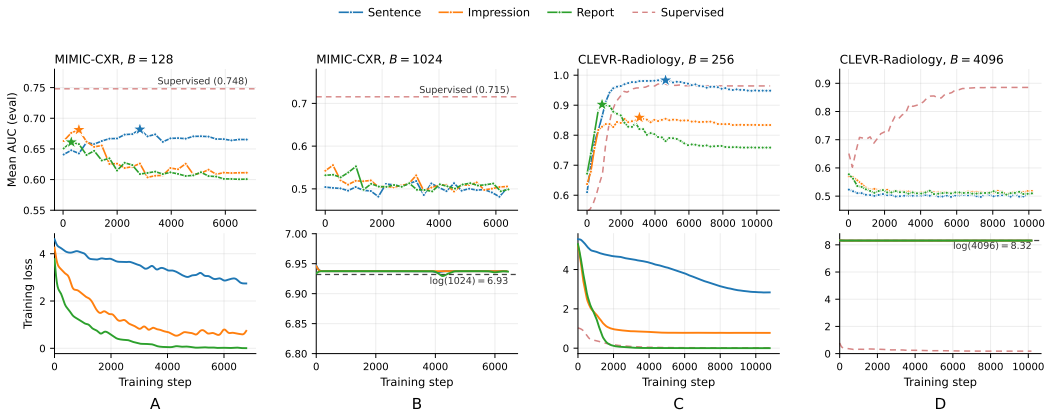


Figure 1: **Two dynamics of vanilla CLIP.** On MIMIC-CXR (A, B) and *CLEVR-Radiology* (C, D), moderate-batch training shows rise-then-fall evaluation while loss keeps decreasing; large-batch training collapses, with AUC near random and loss near $\log B$. Stars mark peak-AUC checkpoints. The supervised and CLIP losses are not directly comparable in absolute value; we contrast their training trajectories, not their levels.

Why these dynamics are puzzling. Neither failure is explained by ordinary causes. For O1, a supervised classifier on the same data and backbone shows no late-training degradation, and vanilla CLIP's peak stays well below this supervised baseline, arguing against generic overfitting as the sole explanation. For O2, large-batch training is routinely beneficial for general-domain CLIP [1, 17, 18]; here it prevents learning entirely. The contrastive target is therefore the natural suspect.

Structural hypothesis. What data structure would make the contrastive target fail this way? Unlike web image–caption data, where captions across images rarely overlap, radiology image–report pairs share a *shared-schema fact-bundle structure*: images depict a largely fixed anatomy (lungs, heart, mediastinum, pleura); each region carries finding states from a small clinical vocabulary; and reports bundle assertions about many such states. A short assertion like "no pleural effusion" is supported by many images in the batch; a longer report like "Mild cardiomegaly. No pleural effusion. Lungs are clear." makes three independent assertions about distinct anatomical regions, and an unpaired report typically matches the anchor image on most of them. Yet vanilla CLIP treats every off-diagonal text as a pure negative, pushing apart what is in fact compatible. We hypothesize this structural property, not any medical-specific artifact, drives both failures, and that the bias compounds with batch size.

***CLEVR-Radiology* isolates the structure.** To test this without confounders from anatomy variation, acquisition artifacts, clinical language, or incomplete annotation, we designed a controlled experiment: *CLEVR-Radiology* is a synthetic testbed where only the three structural ingredients are preserved by construction. Each image contains a fixed set of simple objects (two cubes, a sphere, a cylinder) playing the role of anatomy; each object carries attribute slots (color, size, material) drawn from small vocabularies with one normal and several abnormal values, playing the role of finding states; a template-based generator emits sentence-, impression-, and report-level texts bundling assertions about these slots. Exact fact metadata is recorded for every (image, text) pair (Appendix D). On CLEVR-Radiology, both failures recur (Figure 1(C, D)): at moderate batch, evaluation rises then falls while loss decreases; at large batch (B=4096), AUC stays near chance with loss near $\log(4096)$. The structure alone is sufficient, supporting the hypothesis. Section 3 attributes both failures to a single gradient bias that compounds with batch size.

3 Theory: A Single Misspecified Target Explains Both Failures

Section 2 located the cause of O1 and O2 in a structural property of the data. We now make this precise. Under that structure, vanilla CLIP's paired-identity target is misspecified in a way that is *deterministic* (present at every training step) and *batch-amplified* (worsens as B grows). A single decomposition of the vanilla CLIP gradient recovers O1 and O2 as its two regimes.

3.1 Setup: Fact Compatibility

Vanilla CLIP treats each image as having exactly one matching text. In radiology, many texts can be semantically true of the same image; we formalize this and use it to define an alternative reference target alongside vanilla's paired identity.

Image-text setup. A training batch contains B image–text pairs $\{(x_i, t_i)\}_{i=1}^{B}$. Let s_{ij} be the image–text similarity logit, $p_i^\theta(j) = \exp(s_{ij}) / \sum_\ell \exp(s_{i\ell})$ the row softmax under model parameters θ, and $\ell_i(q_i) = -\sum_j q_i(j) \log p_i^\theta(j)$ the row cross-entropy against any target distribution q_i. Vanilla CLIP uses the paired-identity target $e_i(j) = \mathbf{1}\{j = i\}$. We analyze image-to-text rows; text-to-image is symmetric.

Compatibility kernel. Each image x_i has a latent fact state $z_i = (z_{i,1}, \ldots, z_{i,F})$ over F discrete attribute slots, and each text t asserts a set $A(t)$ of (slot, value) pairs. We say x_i *entails* t_j, written $z_i \models A(t_j)$, if every $(a, v) \in A(t_j)$ satisfies $z_{i,a} = v$. In radiology, a text's meaning reduces to its asserted facts: two texts that assert the same facts are semantically equivalent regardless of surface form (e.g., "cardiomegaly" and "enlarged heart" mean the same for any image), and an image semantically supports a text exactly when the image's facts satisfy the text's assertions. The *compatibility kernel* $K_{ij} = \mathbf{1}\{z_i \models A(t_j)\}$ formalizes this as the simplest fact-level equivalence, collapsing all surface variation. With $m_i = \sum_{j=1}^{B} K_{ij}$ compatible texts in a batch of B, the row-normalized *compatibility reference* $q_i^K(j) = K_{ij}/m_i$ is the simplest semantic target consistent with this kernel: it asks the model to assign equal probability $1/m_i$ to every text in the batch that is semantically supported by the image, and zero to incompatible texts, treating semantically equivalent texts as interchangeable. Vanilla CLIP's e_i implicitly assumes $m_i = 1$; under the structure of Section 2, $m_i > 1$ is typical.

Bundled reports. For multi-finding texts, exact entailment is too strict: a report may share most asserted facts with the anchor while disagreeing on one. We therefore also define a soft reference q_i^R via the partial-agreement fraction $R_{ij} = |A(t_j)|^{-1} \sum_{(a,v)\in A(t_j)} \mathbf{1}\{z_{i,a} = v\}$, and write $q_i^\star \in \{q_i^K, q_i^R\}$ when arguments apply to either (Appendix B.5). The analysis below uses q_i^K for clarity.

Notation across domains. *In radiology*, slots index findings such as cardiomegaly and pleural effusion with values in {present, absent, uncertain}; "no pleural effusion" asserts a single (slot, value) pair, so any effusion-free image entails it. *In CLEVR-Radiology*, slots index object attributes (color, size, material) with small per-attribute vocabularies, and z_i is fully observed. *In general-domain CLIP*, captions are largely instance-specific, so $A(t_j)$ is essentially unique per image, $m_i \approx 1$.

3.2 Decomposing Vanilla CLIP's Gradient

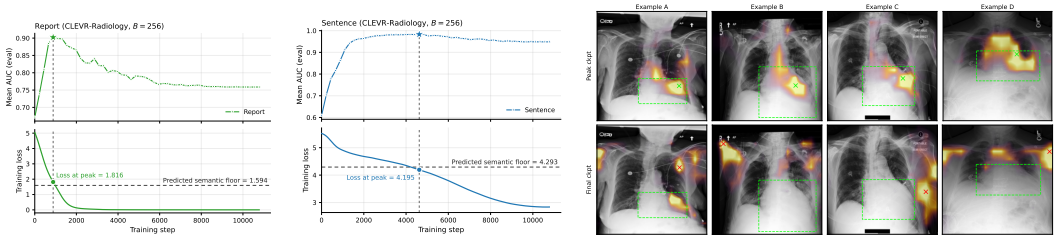
In logit space, vanilla CLIP's gradient can be decomposed into two parts: a *semantic* part that drives the model to score equally any text that is true of the image, and a fixed *identity* part that rewards singling out the paired text from those compatible alternatives.

Lemma 1 (Vanilla CLIP's gradient decomposition)**.** *For any compatibility reference $q_i^\star$, vanilla CLIP's logit-space gradient descent direction decomposes as*

$$-\nabla_{s_i} \ell_i(e_i) \;=\; \underbrace{e_i - p_i^\theta}_{d_i^{\mathrm{van}}(\theta)} \;=\; \underbrace{(q_i^\star - p_i^\theta)}_{d_i^{\mathrm{sem}}(\theta)} \;+\; \underbrace{(e_i - q_i^\star)}_{d_i^{\mathrm{id}}}.$$

The lemma follows from $\nabla_{s_i} \ell_i(e_i) = p_i^\theta - e_i$ by adding and subtracting $q_i^\star$ (Appendix B.2). The decomposition isolates a structural fact: d_i^{id} depends only on the batch, so vanilla CLIP carries a deterministic gradient bias relative to the compatibility reference, independent of training state.

The two components play different roles. $d_i^{\mathrm{sem}}(\theta)$ is the *semantic learning signal*: it is the descent direction of training directly against q_i^K, and vanishes during semantic learning, as the model approaches scoring all m_i compatible texts equally. d_i^{id} is what remains: it pushes up the paired logit s_{ii} and pushes down every equally compatible but unpaired s_{ij}. This is a wrong learning signal, asking the model to distinguish semantically equivalent texts via non-fact-level features. The next two subsections show how d_i^{sem} and d_i^{id} produce O1 and O2.



(A) Report semantic floor (B) Sentence semantic floor (C) Shortcut drift

Figure 2: **Evidence for the semantic-floor and shortcut-learning predictions.** (A, B) On *CLEVR-Radiology*, the eval-peak loss aligns with the generator-computed semantic floor. (C) On MIMIC-CXR, cardiomegaly grounding shifts from the clinical region at peak to irrelevant cues at final.

3.3 Semantic Floor and Shortcut Incentives

Once vanilla CLIP has saturated semantic learning, the only remaining gradient rewards non-semantic features that single out the paired text. This explains O1.

An ideal fact-level encoder would assign equal similarity to all compatible texts, since they are semantically equivalent for the image. We call logits with this property *purely semantic*: $K_{ij} = K_{ik} = 1 \Rightarrow s_{ij} = s_{ik}$. Under this ideal, $p_i^\theta(i) \leq 1/m_i$ (Appendix B.3), with equality when $p_i^\theta = q_i^K$ (incompatible logits $\to -\infty$). The vanilla loss is therefore bounded below by $\log m_i = H(q_i^K)$, the entropy of the compatibility reference; we call this the *semantic floor*. As the model approaches the floor, $p_i^\theta = q_i^K$, so by Lemma 1 the vanilla CLIP's negative gradient reduces to d_i^{id}: the semantic part vanishes and only the fixed identity residual remains.

Proposition 1 (Shortcut reward at the semantic floor)**.** *At $p_i^\theta = q_i^K$, for any feature direction u that raises only the paired logit s_{ii}, $\frac{\partial \ell_i(e_i)}{\partial u} = -\left(1 - \frac{1}{m_i}\right)$, $\frac{\partial \ell_i(q_i^K)}{\partial u} = 0$.*

Proof in Appendix B.3. In the encoder, such u correspond to text-side cues (specific wording, sentence ordering) or image-side cues (border markers, medical devices, in-image text), not the underlying clinical facts. Below the semantic floor, the symmetric vanilla loss forces both encoders to break fact-level clustering: x_i must rank t_i above other compatible t_j, so the text encoder scatters semantic-equivalent texts along wording axes; t_i must rank x_i above other images that entail it, so the image encoder scatters same-finding images along image-specific cues. This is not benign overfitting: zero-shot prompts carry none of these training-side cues, so encoder capacity that should encode *what* fact an image asserts is reallocated to encoding *which* text it was paired with, and clean class prompts no longer find their image neighbors. The longer training stays below the floor, the wider this misalignment grows, producing the rise-then-fall.

Predicting O1. Before the floor, d_i^{sem} dominates the gradient and improves both loss and zero-shot evaluation; the fixed d_i^{id} is present throughout but overshadowed. When approaching the semantic floor, the semantic component vanishes and vanilla training shifts to rewarding these shortcut cues, so zero-shot classification can decline while training loss keeps falling. The persistent d_i^{id} also caps the best vanilla checkpoint: optimization effort is constantly diverted to shortcuts even when d_i^{sem} dominates, leaving vanilla's peak below the supervised baseline. The supervised classifier has no analogue of d_i^{id} (its multi-label target carries no paired-instance direction), so its peak is not similarly capped, and it shows no late-training drift.

Evidence. Two complementary tests confirm the theory. On CLEVR-Radiology, the theory predicts the floor $\log m_i$ from generator metadata before any training, and predicts that loss below this floor coincides with evaluation decline. Real training dynamics confirms this (Figure 2(A, B)). On CXR data, the latent state is hidden, but the predicted shortcuts are visible directly in the model's similarity map: phrase-grounding heatmaps after the peak show attention drifting from the cardiac silhouette to markers and devices, with cardiomegaly pointing-game accuracy dropping from 0.394 at peak to 0.213 at final (a 46% relative loss) while training loss continues to decrease (Figure 2(C)).

3.4 Batch-amplified Identity Pressure

At large batch, the identity component dominates the gradient from the very first step; the model can fail to build semantic representations. This explains O2.

At initialization, $p_i^\theta \approx \pi_B$, the uniform distribution over the B batch texts, so Lemma 1 specializes to

$$e_i - \pi_B \;=\; \underbrace{(q_i^K - \pi_B)}_{d_i^{\mathrm{sem}}(0)} \;+\; \underbrace{(e_i - q_i^K)}_{d_i^{\mathrm{id}}},$$

and the two components are now exactly orthogonal (Appendix B.4). $d_i^{\mathrm{sem}}(0)$ raises every compatible text and lowers every incompatible one (the semantic direction that the model would learn); orthogonality lets us cleanly compare component magnitudes.

Proposition 2 (Batch-amplified identity pressure)**.** *Let ω be the probability that an unpaired text is compatible with an independent image, so $\mathbb{E}[m_i] \sim B\omega$. At initialization,*

$$\Gamma_i(0) \;=\; \frac{\|d_i^{\mathrm{id}}\|_2^2}{\|d_i^{\mathrm{sem}}(0)\|_2^2} \;=\; \frac{B(m_i-1)}{B-m_i} \;\sim\; \frac{\omega}{1-\omega}\,B.$$

The semantic share of the vanilla gradient energy, $1/(1+\Gamma_i(0))$, shrinks as $O(1/B)$.

Proof in Appendix B.4. The fixed d_i^{id} asks the model to pick one paired text from $m_i - 1$ equally compatible alternatives before any fact-level representation has formed, an ill-posed objective at this stage. The component that could build such a representation has only $O(1/B)$ energy share, so training stalls near the uniform-softmax regime with loss near $\log B$ and AUC near random. Radiology reports usually state each finding as present or absent (e.g., "cardiomegaly" vs. "no cardiomegaly"), so any image shares many compatible assertions with many unpaired reports, regardless of finding prevalence; this keeps ω often far from 0 and produces severe dilution at large batches. For general-domain CLIP, $\omega \approx 0$ makes d_i^{id} vanish and the gradient stays essentially semantic, which is why large-batch vanilla CLIP works on web data.

4 Mitigation: Approximating the Compatibility Reference

O1 and O2 share one root: the fixed identity residual d_i^{id}. CorLIP shrinks it by replacing the vanilla target e_i with a closer approximation of q_i^K, the only modification to the contrastive objective; architecture, optimizer, and batch construction are unchanged.

Estimating image states and text assertions. The references q_i^K and q_i^R require known fact states z_i, which is observed in *CLEVR-Radiology* but latent in real radiology data. The paired report, however, is itself the radiologist's reading of z_i from the image, so CorLIP recovers $\hat{z}_i = E_{\mathrm{state}}(r_i)$ and $\hat{A}(t_j) = E_{\mathrm{text}}(t_j)$ by parsing reports offline as a preprocessing step. The construction pipeline can be extended to other radiology modalities by swapping the extractor (Figure 6).

Choosing fact granularity. When are facts from two reports compatible? The answer is non-trivial and depends on the chosen granularity. At coarse granularity, "left pleural effusion" and "right pleural effusion" can both map to "pleural effusion" and be treated as compatible; this expands the compatible pool but also asserts equivalence between clinically distinct sub-findings, which may not match how the downstream task evaluates them. At fine granularity, these distinctions can be preserved, but the compatible pool shrinks and the shared structure across related sub-findings is no longer reflected in the target. No single choice is optimal across all evaluations.

For CXR we instantiate CorLIP with two practical schemas spanning this tradeoff. *CheXbert-13d* [31] is a coarse 13-label schema aligned with CheXpert-style evaluations; we drop the aggregation "No Finding" label since it is not an independent finding. *MedGemma-30d* is a finer schema we define using MedGemma-27B-Text-IT [32] as a structured text parser, covering more findings and distinguishing laterality. Both schemas operate on a flat list of present/absent finding facts; we deliberately leave richer compatibility kernels (severity grading, anatomy/finding hierarchies, inter-finding dependencies) to future work to isolate the contribution of target correction itself. Unless stated otherwise, corrected CXR results use MedGemma-30d as the less benchmark-coupled default.

Scoring compatibility. Given a schema, we score only facts explicitly asserted by the candidate text; unmentioned findings are not treated as contradictions. For sentence-level texts, compatibility is close

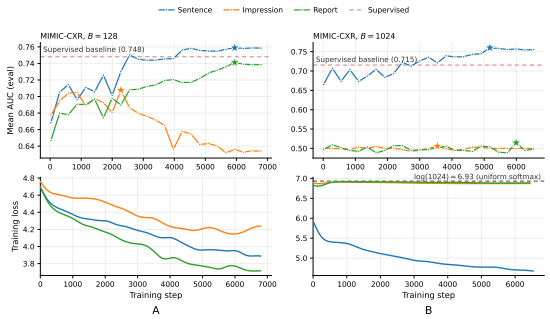


Figure 3: **Corrected CLIP training dynamics on MIMIC-CXR.** (A) CorLIP at B=128. Sentence- and report-level rise-then-fall are largely fixed. (B) CorLIP at B=1024. Sentence CorLIP trains successfully at large batch size. Corrected with CorLIP-MG.

		MIMIC AUC			MIMIC F1		
Method	Text	pk	fn	Δ	pk	fn	Δ
Vanilla CLIP	Rep.	66.1	60.1	−6.0	28.9	23.6	−5.4
Vanilla CLIP	Imp.	68.1	61.1	−7.0	28.7	23.9	−4.9
Vanilla CLIP	Sen.	68.2	66.5	−1.7	28.5	27.5	−0.9
SAT [26]	Rep.	67.9	59.3	−8.7	29.9	24.2	−5.7
MedCLIP [19]	Sen.	77.4	77.2	−0.2	44.7	45.0	+0.3
Ko&Park [20]	Rep.	78.9	77.9	−1.0	45.4	46.8	+1.3
CorLIP-CB	Rep.	77.3	75.4	−1.9	38.6	42.8	+4.2
CorLIP-MG	Rep.	74.1	73.9	−0.2	33.0	32.5	−0.5
CorLIP-CB	Imp.	79.0	72.5	−6.5	42.1	42.2	+0.0
CorLIP-MG	Imp.	70.8	63.4	−7.4	32.7	27.5	−5.2
CorLIP-CB	Sen.	**81.6**	**81.4**	−0.2	**48.5**	**48.1**	−0.4
CorLIP-MG	Sen.	75.8	75.8	0.0	35.5	35.5	0.0

Table 1: **In-domain MIMIC-CXR comparison.** Peak (pk), final (fn), and delta are shown. The pk and fn checkpoints are selected by AUC; F1 is reported at the same checkpoint.

to binary since a sentence asserts one or two facts: $\hat{C}_{ij}^{\text{hard}} = \mathbf{1}\{\hat{z}_i \models \hat{A}(t_j)\}$ approximates K, so multiple sentences supporting the same finding share target mass instead of competing as negatives. For impressions and reports, exact entailment is too strict because a bundled text may agree on most facts while disagreeing on one, so partial agreement $\hat{C}_{ij}^{\text{soft}} = |\hat{A}(t_j)|^{-1} \sum_{(a,v)\in\hat{A}(t_j)} \mathbf{1}\{\hat{z}_{i,a} = v\}$ approximates R; a report sharing four of five asserted findings then receives partial target mass rather than being a pure negative.

Contrastive targets. Given a nonnegative $\hat{C}$, we row- and column-normalize to form targets $\hat{q}_{ij}^{I\to T} = \hat{C}_{ij}/\sum_\ell \hat{C}_{i\ell}$ and $\hat{q}_{ji}^{T\to I} = \hat{C}_{ij}/\sum_\ell \hat{C}_{\ell j}$ that drop into the standard symmetric CLIP cross-entropy. We retain the paired diagonal; handling of empty or non-clinical texts, uncertainty, unmentioned findings, and extractor conventions is in Appendix C.

Why sentence-level correction is best aligned. The residual under target $\hat{q}_i$ is $\hat{q}_i - q_i^\star$ (Lemma 1); the closer $\hat{q}_i$ is to the reference, the smaller this residual. Sentence-level texts are closest to atomic assertions, so the compatibility decision is least ambiguous and hard entailment is closest to K. Bundled texts mix findings, omissions, uncertainty, and stylistic conventions, forcing many fact-level agreements into one scalar. By Proposition 2, this residual is further amplified by batch size, so sentence-level CorLIP is predicted to be the most stable configuration as B grows.

5 Experiments

Setup. *Data and training.* We train on 60k frontal MIMIC-CXR image–report pairs [27]. Report uses FINDINGS+IMPRESSION, impression uses IMPRESSION only, and sentence samples one report sentence per iteration. *Controlled comparison.* To isolate the contribution of the contrastive target, all methods share the same DINOv3-initialized ViT-B image encoder [33], BioClinicalBERT text encoder [34], training data, augmentations, optimizer, and schedule. Methods differ *only* in their contrastive target. Classical CXR VLP objectives are absorbed into this controlled pipeline rather than compared as separate checkpoints: CheXzero [4] and ConVIRT [2] training reduce to vanilla CLIP at report/impression and sentence granularity respectively, so we reinstantiate their paired-identity objective under our pipeline rather than reusing original checkpoints, which differ in training recipe. *Evaluation.* For zero-shot classification, we test in-domain on MIMIC-CXR 14-class and out-of-distribution on CheXpert 14-class [30] and VinDr 28-class [35]. For dense evaluation, we use phrase grounding on MS-CXR bounding boxes [5]. Full hyperparameters are in Appendix E.

The experiments are designed as theory checks, not architecture search. They ask whether correcting the target changes the training trajectory, whether data scale helps once the target is corrected, and whether large batches become useful when compatible texts are no longer repelled.

Correcting the target changes the trajectory, not only the score. At B=128, the key change is trajectory-level. Figure 3 shows that sentence-level CorLIP largely removes the post-peak drop seen in vanilla CLIP training (Figure 1). Table 1 quantifies the same effect: vanilla CLIP loses 1.7–7.0

AUC points by the final checkpoint, whereas CorLIP-CB sentence reaches the best peak and final scores on both MIMIC AUC and F1, with only 0.2 AUC and 0.4 F1 drop. Report-level correction also reduces vanilla report degradation; impression-level correction is less reliable, suggesting that compatibility quality is not monotonic in text length. Impressions are lossy summaries, neither atomic like sentences nor information-rich enough for stable soft report-level correction. Notably, corrected sentence training exceeds the supervised baseline, showing that the contrastive objective recovers its expected advantage once the target is corrected.

Data scaling helps more when it scales the right target. Appendix A.3 shows that doubling data from 60K to 120K raises vanilla CLIP's best peak AUC only from 68.2 to 70.6, and all granularities still degrade after the peak. With corrected targets, 60K CorLIP-CB sentence already reaches $81.6/81.4$ peak/final AUC, over 10 points above the best 120K vanilla checkpoint; at 120K, it further improves to $86.5/86.0$. Data scaling is powerful when it scales the right target, but inefficient when it scales a misspecified one.

The gains transfer to out-of-distribution data and dense tasks. Table 2 asks whether the stronger MIMIC representation transfers and how it works on dense tasks. CorLIP results in the best result on all evaluated settings: CorLIP-MG sentence on CheXpert, CorLIP-MG report/sentence on VinDr, and CorLIP-MG sentence on MS-CXR grounding. Compared with Table 1, CheXbert-based correction is strongest on in-domain MIMIC, whose evaluation labels derive from the same CheXpert-style labeling as its correction schema. MedGemma-based correction is stronger on OOD settings and dense tasks in our runs, making it the stronger fixed default of the two; since the schemas differ in both granularity and extractor, we do not attribute this gap to granularity alone.

Large batches become useful once the target is compatible. The right column of Figure 3 shows the large-batch regime. At $B{=}1024$, report- and impression-level correction still fail to train reliably, while sentence-level correction escapes the uniform-softmax floor and reaches a useful AUC score. Replacing paired identity is therefore not enough if the text unit still bundles many facts into a lossy partial-compatibility score.

Figure 4(A) broadens the comparison across methods and batch sizes. A few non-sentence configurations still train at $B{=}512$, including MedCLIP and impression-level runs, but they collapse by $B{=}1024$. Among the evaluated configurations, only CorLIP sentence remains stable through $B{=}2048$, where it still achieves competitive peak AUC values. We stop at $B{=}2048$ because larger batches exceed our compute budget.

Figure 4(B) isolates the stable regime. With corrected sentence targets, larger batches are helpful up to $B{=}512$: peak AUC rises from 71.1 at $B{=}16$ to 77.0 at $B{=}512$, then remains stable at 76.0 and 75.7 for $B{=}1024$ and $B{=}2048$. The usual lesson from medical CLIP would be to avoid large batches [2, 19–21], and commonly used CXR batch sizes are often below 128 [4, 7]. Our result changes the diagnosis: large batches are harmful under a misspecified target, but more negatives in large batches become useful again after target correction. The peak arrives much earlier than in web-scale CLIP systems [1, 17, 18], plausibly because CXR has a smaller effective semantic space, so a few hundred images already provide enough genuinely distinct negatives.

Beyond softmax CLIP. Appendix A.6 shows the same conclusion for SigLIP-style sigmoid contrastive training: vanilla SigLIP exhibits the same rise-then-fall and large-batch collapse, while our correction can largely mitigate the issue.

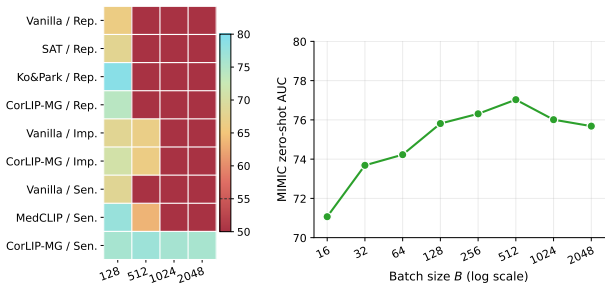
Correction transfers to 3D CT. The bias we identified is target-level, not modality-level. To check this, we apply CorLIP to Merlin 3D CT image–report pretraining with a CT-adapted MedGemma extractor; encoder, preprocessing, optimizer, and schedule are fixed at Merlin's defaults (Appendix E.3). Report-level vanilla CLIP shows the same rise-then-fall dynamic as on CXR: mean AUC peaks at 67.9 and then degrades by about 5 points while the training loss keeps decreasing. Sentence-level training raises the peak to 75.8, and sentence-level CorLIP reaches 79.3 (Table 3). The CXR pattern (sentence beats report, correction beats vanilla) holds on CT despite the shift in modality, encoder backbone, and report style.

Table 3: **3D CT transfer (Merlin).** Peak mean AUC over 30 findings, $\times 100$, with 95% CI.

Text	Method	AUC
Rep.	Vanilla	67.9 (66.4–69.5)
Sen.	Vanilla	75.8 (74.5–77.2)
Sen.	CorLIP	**79.3** (78.0–80.6)

Method	Text	CheXpert	VinDr	MS-CXR PG
Vanilla CLIP	Rep.	79.8	83.0	22.4
Vanilla CLIP	Imp.	78.7	79.6	30.8
Vanilla CLIP	Sen.	82.1	83.1	30.7
SAT [26]	Rep.	80.7	83.9	34.0
MedCLIP [19]	Sen.	80.9	82.8	32.3
Ko&Park [20]	Rep.	83.5	84.1	35.8
CorLIP-CB	Rep.	84.5	83.7	34.5
CorLIP-MG	Rep.	<u>86.2</u>	**87.3**	<u>36.9</u>
CorLIP-CB	Imp.	81.7	82.7	33.1
CorLIP-MG	Imp.	85.9	<u>85.2</u>	34.1
CorLIP-CB	Sen.	85.6	83.7	37.7
CorLIP-MG	Sen.	**86.6**	**87.3**	**41.6**

Table 2: **OOD classification and grounding.** Mean AUC for CheXpert and VinDr, mean pointing-game accuracy for MS-CXR phrase grounding.



(A) Batch trainability (B) Batch scaling

Figure 4: **Large-batch trainability and scaling.** (A) Which methods keep learning or collapse as B grows. (B) Batch scaling within CorLIP-MG sentence training, the only stable large-batch regime.

6 Related Work

Medical false-negative correction in VLP. We focus on methods that modify the CLIP targets or negatives. MedCLIP [19] expands paired data combinatorially and replaces InfoNCE with a medical-knowledge-guided semantic matching loss. SAT [26] triages pairs by inter-report similarity. Ko and Park [20] build dynamic soft labels from text, CheXbert, and RadGraph similarities, combined with negation-based hard negatives. These methods soften image-text pairs through similarity using labels, knowledge, or graphs. CorLIP instead corrects the CLIP targets to assertion-level compatibility: a text is positive when its asserted facts are entailed by the image state, regardless of global report similarity, while the rest of CLIP is left unchanged. Visual-entailment formulations for positive and negative mentions [36] are related; our focus is deriving the contrastive target distribution and using its misspecification to explain both training dynamics.

Contrastive-learning theory and supervised contrastive learning. General contrastive-learning theory explains when instance discrimination yields semantic representations through latent classes [37], alignment and uniformity [38], or identifiability of latent generative factors [39]. False-negative and soft-target methods relax one-hot contrastive targets by correcting sampled negatives [40] or using intra-modal similarity to build softened targets [41]. These analyses are informative, but radiology lies outside their usual regime: compatible off-diagonal texts are dense rather than rare, reports bundle multiple dependent findings, and many true statements are negative assertions. We therefore view the failure as deterministic target misspecification rather than stochastic class collision. Our mitigation objective is also related to supervised contrastive learning [42], but the relation is not symmetric class equivalence. Radiology compatibility is multi-label, compositional, and directional: an image can entail an assertion without being equivalent to the full report containing it. Unlike a SupCon label set, which defines the entire semantic space, the schema only corrects targets within its coverage; standard CLIP supervision applies elsewhere, preserving open-vocabulary zero-shot inference.

Structured reports and noisy labels. We use clinical NLP structure only to estimate the contrastive target distribution, drawing on tools for findings, negation, uncertainty, and relations [43, 30, 31, 44]. This differs from prior uses of report structure as classification labels, grounding supervision, or knowledge injection [3, 5, 25, 24]. The rise-then-fall dynamic also resembles noisy-label memorization [45–47], but the mechanism is different: the paired label is not randomly corrupted; the one-hot target is structurally incomplete and assigns negative mass to clinically compatible off-diagonal assertions.

7 Discussion and Limitations

The unit of contrast in radiology should be the assertion, not the report. Reports bundle reusable assertions, so supervision should ask whether an image supports a text assertion, not whether the text is the unique paired report. Sentences are the closest practical approximation to atomic assertions; for bundled reports, even the best soft correction can supply only a single scalar per (image, text)

pair, which is lossy when many facts are bundled, leaving the encoder to recover fact-level structure on its own. Sentence-level CorLIP delivers this: across iterations, the same image is queried by different supported assertions, while corrected targets stop repelling compatible off-diagonal texts, encouraging the encoder to encode the image fact state rather than paired-report identity.

Several limits of medical CLIP turn out to be target-specific rather than fundamental. On CXR, large batches once considered harmful now become stable through $B{=}2048$ and deliver useful negatives up to $B{=}512$. Scaling data alone delivers only marginal gains while scaling with corrected targets delivers substantial improvements. The same correction transfers from CXR to 3D CT, indicating the bias is target-level rather than modality-level. Together, these results point to the contrastive target as a key bottleneck of medical CLIP: many practical compromises (small batches, limited data efficiency, modest OOD transfer) stem largely from paired-identity misspecification rather than fundamental constraints. In radiology, defining which pairs count as positive is itself a central design decision, not a fixed convention inherited from general-domain CLIP.

As a design decision, the definition of positives carries a trade-off. CorLIP reads compatibility off a finite schema, so the schema determines which clinical distinctions the target preserves and which it treats as interchangeable. A coarse schema can over-correct: merging left- and right-sided findings, for example, assigns positive target mass to clinically distinct cases, potentially weakening representations of laterality. Removing the incentive to distinguish such cases does not force identical embeddings, but may reduce information useful for finer-grained tasks. Conversely, an increasingly fine flat schema can under-correct: compatible sets become sparse, while related findings receive no shared supervision through their higher-level concepts; in the extreme, compatibility approaches paired identity. Texts with no schema-covered fact retain only their paired positive, so content outside the schema keeps pair-specific supervision. Schema design therefore trades off preserving clinically meaningful distinctions against correcting recurring false-negative conflicts. Our experiments establish benefits for the two evaluated schemas, which differ in both granularity and extractor, rather than identifying an optimal schema.

Our correction is intentionally minimal, and richer compatibility kernels are a natural way to ease this trade-off. Fine-grained assertions over finding, anatomy, laterality, severity, uncertainty, and temporal change, combined through clinical ontologies into hierarchical compatibility, could preserve detailed distinctions while sharing supervision across related findings. Evaluating the corrected representations beyond the image–text similarity probes used here, on report generation, segmentation, and supervised fine-tuning, is an important next step. CorLIP is also orthogonal to other axes of medical CLIP improvement: stronger radiology foundation models and backbones for CXR and CT [6, 9, 8, 12], region-level or global-local objectives [3, 5], specialized clinical knowledge and language models [24, 32], and larger curated biomedical or CXR corpora [48, 7]; combining target correction with these advances is a promising direction.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Medical Imaging and Data Resource Center (MIDRC), which is funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under contract 75N92020C00021 and through the Advanced Research Projects Agency for Health (ARPA-H). The authors declare no competing interests.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[2] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR, 2022.

[3] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In

Proceedings of the IEEE/CVF international conference on computer vision, pages 3942–3951, 2021.

[4] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.

[5] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[6] Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

[7] Chong Wang, Yabin Zhang, Yunhe Gao, Maya Varma, Clemence Mottez, Faidra Patsatzi, Jiaming Liu, Jin Long, Jean-Benoit Delbrouck, Sergios Gatidis, et al. A data-and compute-efficient chest x-ray foundation model beyond aggressive scaling. *arXiv preprint arXiv:2602.22843*, 2026.

[8] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19, 2026.

[9] Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, pages 1–11, 2026.

[10] Qi Zhang, Jianpeng Zhang, Weiwei Cao, Zilin Lu, Wanxing Chang, Haonan Ding, Cao Chen, Zhi Li, Xing Xue, Sinuo Wang, et al. An expert-level generalist AI for abdominal CT diagnosis. *Science*, 393(6817):eaec6129, 2026.

[11] Xingxin He, Aurora Rofena, Ruimin Feng, Haozhe Liao, Zhaoye Zhou, Albert Jang, and Fang Liu. Omnimri: A unified vision–language foundation model for generalist mri interpretation. *arXiv preprint arXiv:2508.17524*, 2025.

[12] Chenhui Zhao, Yiwei Lyu, Asadur Chowdury, Edward Harake, Akhil Kondepudi, Akshay Rao, Xinhai Hou, Honglak Lee, and Todd Hollon. Towards scalable language-image pre-training for 3d medical imaging. *arXiv preprint arXiv:2505.21862*, 2025.

[13] Kumar Krishna Agrawal, Longchao Liu, Long Lian, Michael Nercessian, Natalia Harguindeguy, Yufu Wu, Peter Mikhael, Gigin Lin, Lecia V Sequist, Florian Fintelmann, et al. Pillar-0: A new frontier for radiology foundation models. *arXiv preprint arXiv:2511.17803*, 2025.

[14] Zhixiu Lu, Hailong Li, Nehal A Parikh, Jonathan R Dillman, and Lili He. Radclip: Enhancing radiologic image analysis through contrastive language–image pretraining. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[15] Yogesh Kumar and Pekka Marttinen. Improving medical multi-modal contrastive learning with expert annotations. In *European Conference on Computer Vision*, pages 468–486. Springer, 2024.

[16] Yaxiong Chen, Chuang Du, Chunlei Li, Jingliang Hu, Yilei Shi, Shengwu Xiong, Xiao Xiang Zhu, and Lichao Mou. Unicrossadapter: Multimodal adaptation of clip for radiology report generation. In *International workshop on foundation models for general medical AI*, pages 113–123. Springer, 2024.

[17] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[18] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

[19] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022.

[20] Hanbin Ko and Chang-Min Park. Bringing clip to the clinic: Dynamic soft labels and negation-aware learning for medical analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25897–25906, 2025.

[21] Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel CF Codella, et al. Comprehensive language-image pre-training for 3d medical image understanding. *arXiv preprint arXiv:2510.15042*, 2025.

[22] George Fisher. Vision-language foundation models do not transfer to medical imaging classification: A negative result on chest x-ray diagnosis. *medRxiv*, pages 2025–12, 2025.

[23] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[24] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21372–21383, October 2023.

[25] Che Liu, Sibo Cheng, Miaojing Shi, Anand Shah, Wenjia Bai, and Rossella Arcucci. Imitate: Clinical prior guided hierarchical vision-language pre-training. *IEEE Transactions on Medical Imaging*, 44(1):519–529, 2024.

[26] Bo Liu, Donghuan Lu, Dong Wei, Xian Wu, Yan Wang, Yu Zhang, and Yefeng Zheng. Improving medical vision-language contrastive pretraining with semantics-aware triage. *IEEE Transactions on Medical Imaging*, 42(12):3579–3589, 2023.

[27] 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.

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[30] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[31] 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)*, pages 1500–1519, 2020.

[32] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[33] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[34] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78, 2019.

[35] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

[36] WeiLong Wu, Jingzhi Yang, Xun Zhu, Xiao Zhang, ZiYu Liu, Miao Li, and Ji Wu. Medical contrastive learning of positive and negative mentions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 392–401. Springer, 2025.

[37] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning*, pages 5628–5637. PMLR, 2019.

[38] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[39] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International conference on machine learning*, pages 12979–12990. PMLR, 2021.

[40] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[41] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1860–1868, 2024.

[42] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[43] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018.

[44] Saahil Jain, Ashwin Agrawal, Adriel Saporta, SQ Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports (2021). *arXiv preprint arXiv:2106.14463*, 2021.

[45] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[46] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[47] Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*, pages 24851–24871. PMLR, 2022.

[48] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[49] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

[50] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

A Additional Evidence for Target Misspecification and Correction

A.1 When Paired-Identity Target Creates False Negatives

Figure 5 gives a schematic view of the target mismatch studied in the paper. Vanilla CLIP uses paired identity: in each image-to-text row, only the paired text receives positive mass, and every off-diagonal text is treated as a negative. For radiology image–report data, this row target is too fine-grained. A short assertion can be entailed by many different images, so exact-compatible off-diagonal sentences should not be pure negatives. A longer report may disagree with the anchor image on one finding while still sharing many other asserted facts, so whole-report negatives are often partial positives at the fact level.

The corrected target changes only this supervision matrix. For short assertions, hard correction assigns positive mass to texts exactly entailed by the image. For bundled reports, soft correction assigns mass according to partial fact agreement. This visualization is not an additional method; it is a compact guide to the notation used in Sections 3 and 4: paired identity e_i, exact compatibility K_{ij}, and partial agreement R_{ij}.

False negatives in real MIMIC-CXR batches. To quantify this on real data, we sample 1,000 training batches at $B{=}128$ and $B{=}1024$ and count, for each sentence anchor, the off-diagonal images compatible with it under CheXbert-13d. False negatives arise for both negative and positive assertions. "No lung opacity" has $84.9{\pm}1.2$ false-negative images per batch at $B{=}128$ and $686.3{\pm}4.6$ at $B{=}1024$ (66.8% and 67.1% of off-diagonal entries). "Lung opacity" has 13.7 ± 0.1 and 111.2 ± 0.4 (10.8% and 10.9%). Even the rare finding "fracture" averages 2.1 ± 0.1 and 17.2 ± 0.1 (1.6% and 1.7%), and because rare findings already receive weak learning signal, these false negatives carry disproportionate weight.

A.2 How Compatibility Targets Are Constructed

Figure 6 illustrates how CorLIP constructs the corrected target in the sentence-level hard-correction case. Since the true image fact state is unobserved, we approximate it by extracting a structured fact vector from the paired full report. The sampled sentence supplies the text-side assertion. The compatibility kernel then records whether each image-side fact state entails each sentence assertion.

Row-normalizing the compatibility kernel gives the image-to-text corrected target. Texts that express the same fact in different phrasings receive the same target support for any image that satisfies that fact, so the correction acts as implicit text augmentation inside the contrastive target. For report and impression correction, the same construction is used with a soft compatibility score: the binary entailment entry is replaced by the fraction of asserted facts satisfied by the image-side state estimate.

A.3 Target Correction Matters More Than Scaling Data

Table 4 separates two possibilities: vanilla CLIP may fail because the target is biased, or because 60K training studies are not enough. Doubling the training set to 120K helps vanilla CLIP only modestly.

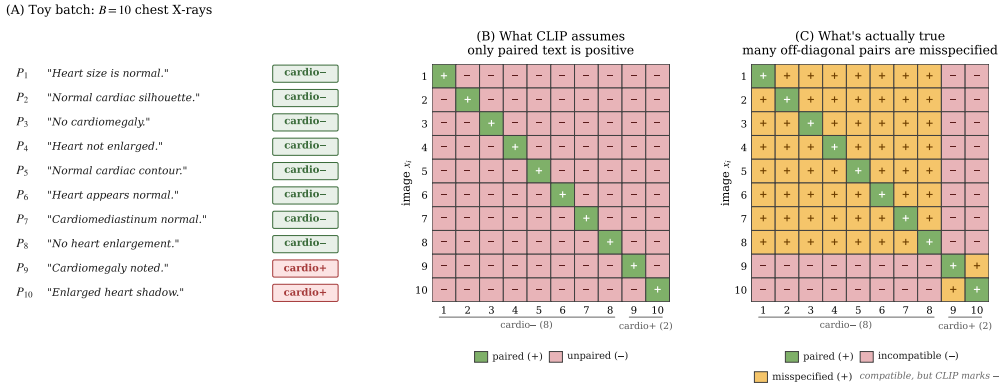


Figure 5: **Target misspecification in medical CLIP.** (A) A $B{=}10$ toy batch of CXR reports with 20% cardiomegaly prevalence; eight reports describe the same normal finding using different paraphrases while two reports describe cardiomegaly with different phrases. (B) CLIP's paired-identity target marks all 90 off-diagonal image–text pairs as negative. (C) A compatibility-aware target reveals that 58/90 off-diagonal cells (64.4%, yellow) are semantically compatible, while the remaining cells (pink) are genuinely incompatible.

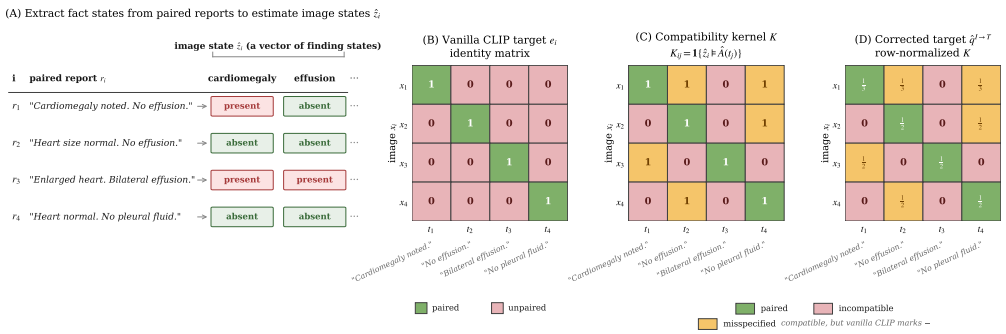


Figure 6: **Constructing corrected targets.** (A) An offline extractor maps each paired report to an estimated image-level fact state $\hat{z}_i$. (B) Vanilla CLIP uses the identity matrix as its target. (C) For sampled sentences, the hard compatibility kernel K marks whether each sentence assertion is entailed by each estimated image state; yellow cells are compatible pairs that vanilla CLIP would mark as negatives. (D) Row-normalizing K gives the image-to-text target $\hat{q}^{I\to T}$, distributing mass across compatible sentences. The text-to-image target $\hat{q}^{T\to I}$ is obtained by column-normalizing K.

The best vanilla checkpoint rises from 68.2 to 70.6 mean AUC, and all three text granularities still degrade after their peak at 120K. Thus the rise-then-fall dynamic is not explained by insufficient data. More data moves the biased trajectory upward, but it does not remove the bias that makes the final checkpoint worse than the peak.

The contrast with target correction is much larger. This diagnostic uses the CheXbert correction. CorLIP-CB trained on only 60K sentence pairs reaches 81.6 peak AUC and 81.4 final AUC, already more than 10 points above vanilla CLIP trained on 120K, even when vanilla is evaluated at its own best checkpoint. Target correction also makes additional data more useful: with 120K, corrected sentence-level training reaches 86.5 peak and 86.0 final AUC, while its peak-to-final drop remains below one point. The lesson is that scale and target quality are not interchangeable. Adding more data to a misspecified target gives limited gains; correcting the target changes the training trajectory and then allows data scaling to pay off.

A.4 Corrected Targets on *CLEVR-Radiology*

Figure 7 shows the controlled testbed after replacing paired identity with compatibility targets. At both moderate and large batch, corrected training produces stable evaluation trajectories. Sentence-level correction reaches near-perfect AUC fastest and exceeds the supervised classifier at both $B{=}256$

Table 4: **Data scaling on MIMIC-CXR at B=128.** Peak (pk), final (fn), and $\Delta = \text{fn} - \text{pk}$ MIMIC zero-shot mean AUC, multiplied by 100.

Method	Text	60K			120K		
		pk	fn	Δ	pk	fn	Δ
Vanilla CLIP	Rep.	66.1	60.1	−6.0	68.9	63.6	−5.3
Vanilla CLIP	Imp.	68.1	61.1	−7.0	69.0	61.5	−7.5
Vanilla CLIP	Sen.	68.2	66.5	−1.7	70.6	67.4	−3.2
CorLIP-CB	Rep.	77.3	75.4	−1.9	84.5	82.1	−2.4
CorLIP-CB	Imp.	79.0	72.5	−6.5	84.4	80.7	−3.6
CorLIP-CB	Sen.	81.6	81.4	−0.2	86.5	86.0	−0.6

and B=4096; report-level correction also approaches the supervised curve, while impression-level correction remains lower, consistent with impression text being a lossy bundle of multiple facts. This behavior is natural for a corrected contrastive objective: once positives are defined by compatibility, the objective uses many compatible texts per image, much like supervised contrastive learning uses many same-class positives. Prior work shows that supervised contrastive learning can outperform cross-entropy empirically [42], and that supervised contrastive and cross-entropy losses share the same class-collapse simplex geometry at their minima while exhibiting different optimization behavior [49]. Our result is the fact-level analogue: corrected contrastive learning does not change the underlying labels in *CLEVR-Radiology*, but it reaches high-quality semantic representations faster by using the compatibility structure during training.

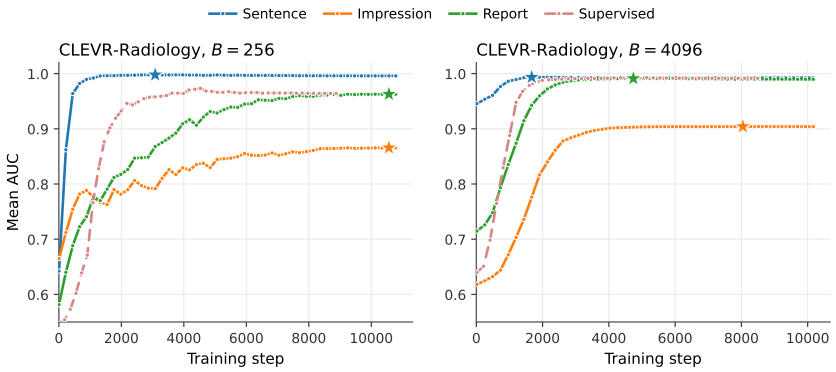


Figure 7: **Corrected targets on *CLEVR-Radiology*.** Evaluation mean AUC at B=256 and B=4096. Stars mark peak checkpoints; the supervised curve is shown for reference.

A.5 Extractor Choice Changes the Tradeoff, Not the Conclusion

Table 5 collects the vanilla and CorLIP rows from Tables 1 and 2. The goal is not to rank extractors in isolation, since the two schemas expose different fact sets. Instead, the table checks whether the main conclusion depends on one particular extractor. It does not: both CheXbert and MedGemma correction improve over vanilla CLIP, and both identify sentence-level correction as the most reliable training regime.

The tradeoff follows from the granularity of the extracted fact schema. CheXbert-13d is closely aligned with the CheXpert-style label space used for in-domain MIMIC evaluation, so it gives the strongest MIMIC AUC values. This alignment is also a limitation. The schema is coarse: it covers a small set of findings, does not distinguish laterality, and cannot correct false negatives for findings outside the 13-dimensional label space. For example, left- and right-lung findings can be mapped to the same extracted state, making them compatible under the correction even when localization should distinguish them. MedGemma-30d exposes a broader and finer fact schema, including more findings and laterality. It is therefore less tightly matched to the in-domain MIMIC labels, but transfers better to OOD and dense evaluations, especially VinDr's 28-label classification and MS-CXR grounding.

Thus the gains are not explained only by reusing the CheXbert label space: changing the extractor changes which evaluations benefit most, but compatibility correction remains beneficial.

Table 5: **Extractor comparison at B=128.** MIMIC reports peak zero-shot mean AUC; CheXpert, VinDr, and MS-CXR report best-checkpoint values. All values are multiplied by 100.

Text	Method	Extractor	MIMIC AUC	CheXpert AUC	VinDr AUC	MS-CXR PG
Rep.	Vanilla	–	66.1	79.8	83.0	22.4
Rep.	CorLIP	CheXbert-13d	77.3	84.5	83.7	34.5
Rep.	CorLIP	MedGemma-30d	74.1	86.2	**87.3**	36.9
Imp.	Vanilla	–	68.1	78.7	79.6	30.8
Imp.	CorLIP	CheXbert-13d	79.0	81.7	82.7	33.1
Imp.	CorLIP	MedGemma-30d	70.8	85.9	85.2	34.1
Sen.	Vanilla	–	68.2	82.1	83.1	30.7
Sen.	CorLIP	CheXbert-13d	**81.6**	85.6	83.7	37.7
Sen.	CorLIP	MedGemma-30d	75.8	**86.6**	**87.3**	**41.6**

A.6 SigLIP Dynamics Under Vanilla and Corrected Targets

Figure 8 repeats the MIMIC-CXR dynamics with a SigLIP-style sigmoid contrastive loss [17]. SigLIP uses the same MIMIC-CXR data, encoders, training schedule, and evaluation protocol as the CLIP experiments. This is a robustness check rather than a full parallel benchmark, but the pattern is informative. Vanilla SigLIP exhibits the same two dynamics as vanilla CLIP: at B=128, evaluation rises early and then falls for all three text granularities, even though training continues; at B=1024, it collapses to near-random AUC, matching the large-batch failure in Figure 1. After replacing paired identity with MedGemma-30d compatibility targets, sentence-level correction becomes stable at both batch sizes and exceeds the supervised baseline in both settings. Thus the correction is not tied to the softmax form of CLIP; it addresses the target structure shared by CLIP and SigLIP-style objectives.

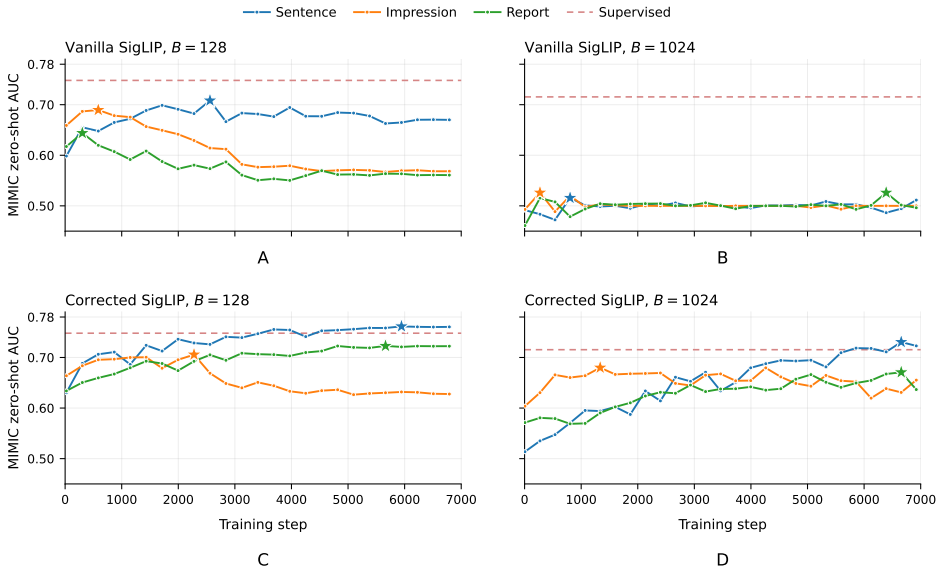


Figure 8: **SigLIP dynamics on MIMIC-CXR.** (A, B) Vanilla SigLIP at B=128 and B=1024. (C, D) Corrected SigLIP using MedGemma-30d compatibility targets at B=128 and B=1024. Stars mark peak checkpoints; dashed lines are supervised baselines.

B Complete Proofs

B.1 Fact-State Notation and Compatibility

This appendix expands the notation used in Section 3. Let

$$z_i = (z_{i,1}, \ldots, z_{i,F})$$

be the latent fact state of image x_i, where $a \in \{1, \ldots, F\}$ indexes an atomic finding slot, such as CARDIOMEGALY or PLEURAL EFFUSION, and $z_{i,a}$ is the value of that slot, such as present or absent. For a text t, let $\mathcal{S}(t)$ be the set of fact slots asserted by the text. For each asserted slot $a \in \mathcal{S}(t)$, let $v_a(t)$ be the value asserted by the text. The asserted fact set is

$$A(t) = \{(a, v_a(t)) : a \in \mathcal{S}(t)\}.$$

We say that image x_i entails text t_j, written $z_i \models A(t_j)$, if every asserted fact in t_j agrees with the image fact state:

$$z_{i,a} = v_a(t_j) \quad \text{for all } a \in \mathcal{S}(t_j).$$

The exact compatibility kernel is $K_{ij} = \mathbf{1}\{z_i \models A(t_j)\}$. The paired text is compatible with its image by construction, so $K_{ii} = 1$. The row-normalized exact-compatibility reference is $q_i^K(j) = K_{ij} / \sum_{\ell=1}^{B} K_{i\ell}$.

B.2 Proof of Lemma 1

Proof. The proof starts from the standard softmax cross-entropy gradient identity. For any target distribution q_i, $\nabla_{s_i} \ell_i(q_i) = p_i - q_i$. Indeed, $\partial \log p_i(j) / \partial s_{ik} = \delta_{jk} - p_i(k)$, so

$$\frac{\partial \ell_i(q_i)}{\partial s_{ik}} = -\sum_j q_i(j)\big[\delta_{jk} - p_i(k)\big] = p_i(k) - q_i(k),$$

using $\sum_j q_i(j) = 1$. Subtracting,

$$\nabla_{s_i} \ell_i(q_i) - \nabla_{s_i} \ell_i(\tilde{q}_i) = (p_i - q_i) - (p_i - \tilde{q}_i) = \tilde{q}_i - q_i.$$

The model distribution p_i cancels: the gradient gap depends only on the target mismatch, not on the current model state.

For vanilla CLIP, $q_i = e_i$, so the logit-space gradient-descent direction is

$$-\nabla_{s_i} \ell_i(e_i) = e_i - p_i.$$

Adding and subtracting any compatibility reference $q_i^\star$ gives

$$-\nabla_{s_i} \ell_i(e_i) = e_i - p_i = (q_i^\star - p_i) + (e_i - q_i^\star) = d_i^{\text{sem}} + d_i^{\text{id}}.$$

This is Lemma 1. The first term is the descent direction that would move the row toward the compatibility reference. The second term is the fixed identity residual. In the exact-reference case $q_i^\star = q_i^K$, this residual has value $1 - 1/m_i$ on the paired text, $-1/m_i$ on each of the other $m_i - 1$ compatible texts, and 0 on incompatible texts. □

Immediate consequence: approximate correction sufficiency. Let q_i^{ref} denote the compatibility reference used for a row: q_i^K for short exact assertions, or q_i^R for bundled reports. If $\hat{q}_i$ is the practical target used for training, the same softmax identity used in Lemma 1 gives

$$\nabla_{s_i} \ell_i(\hat{q}_i) - \nabla_{s_i} \ell_i(q_i^{\text{ref}}) = q_i^{\text{ref}} - \hat{q}_i.$$

Equivalently, the difference between the corresponding descent directions is $\hat{q}_i - q_i^{\text{ref}}$. Thus, in logit space, approximate correction improves over vanilla one-hot whenever

$$\|\hat{q}_i - q_i^{\text{ref}}\|_1 < \|e_i - q_i^{\text{ref}}\|_1. \tag{1}$$

Parameter-gradient deviations are the corresponding score-Jacobian-weighted versions of these logit-gradient deviations. This is the appendix form of the design principle used in Section 4: the closer the constructed target is to the compatibility reference, the smaller the structural gradient bias.

B.3 Proof of Proposition 1

Proof. **Part (i): Semantic floor.**

Let $P_i = \{j : K_{ij} = 1\}$ be the compatible set, with $|P_i| = m_i$. A purely semantic predictor assigns the same logit c_i to every $j \in P_i$. Therefore

$$p_i(i) = \frac{\exp(c_i)}{m_i \exp(c_i) + \sum_{j \notin P_i} \exp(s_{ij})} \leq \frac{1}{m_i}. \tag{2}$$

The vanilla row loss is $\ell_i(e_i) = -\log p_i(i)$, hence

$$\ell_i(e_i) \geq \log m_i. \tag{3}$$

Because K is binary, q_i^K is uniform over P_i, so $H(q_i^K) = \log m_i$. The lower bound is approached when all incompatible logits go to $-\infty$. If the observed vanilla loss drops below $\log m_i$, the model must be distinguishing among compatible texts using non-semantic features.

Part (ii): What happens at $p_i = q_i^K$.

The proposition states the behavior at the semantic floor. At $p_i = q_i^K$, the paired text has probability $p_i(i) = q_i^K(i) = 1/m_i$, so

$$\ell_i(e_i) = -\log p_i(i) = \log m_i.$$

The vanilla descent direction is

$$-\nabla_{s_i} \ell_i(e_i) = e_i - p_i = e_i - q_i^K = d_i^{\mathrm{id}}.$$

Thus the semantic component has vanished, and the only remaining vanilla descent direction is the fixed identity residual.

Part (iii): Shortcut incentive.

Assume $\partial s_{ii}/\partial u = 1$ and $\partial s_{ij}/\partial u = 0$ for $j \neq i$. By the chain rule:

$$\frac{\partial \ell_i(q_i)}{\partial u} = \frac{\partial \ell_i}{\partial s_{ii}} \cdot \frac{\partial s_{ii}}{\partial u} = p_i(i) - q_i(i), \tag{4}$$

using $\partial \ell_i / \partial s_{ii} = p_i(i) - q_i(i)$ from the proof of Lemma 1.

For vanilla ($q_i(i) = 1$): $\partial \ell_i(e_i)/\partial u = p_i(i) - 1$.

For exact-compatibility-aware training ($q_i(i) = q_i^K(i)$): $\partial \ell_i(q_i^K)/\partial u = p_i(i) - q_i^K(i)$.

At the semantic symmetric point with $p_i(i) = q_i^K(i) = 1/m_i$:

$$\frac{\partial \ell_i(e_i)}{\partial u} = \frac{1}{m_i} - 1 = -\left(1 - \frac{1}{m_i}\right) < 0, \qquad \frac{\partial \ell_i(q_i^K)}{\partial u} = 0. \tag{5}$$

Since gradient descent updates $u \leftarrow u - \eta \cdot \partial \ell / \partial u$, a negative derivative means u increases; the model is *rewarded* for encoding nuisance features under vanilla training. □

B.4 Proof of Proposition 2

At initialization $p_i \approx \pi_B$, where $\pi_B(j) = 1/B$. It is convenient to analyze the vanilla descent direction $e_i - \pi_B$, since gradient descent on logits moves in the negative gradient direction $e_i - p_i$. The proof decomposes this direction into a semantic part and an identity part, then compares their squared norms.

Orthogonal decomposition. The decomposition

$$e_i - \pi_B \;=\; \underbrace{(q_i^K - \pi_B)}_{d_i^{\mathrm{sem}}} \;+\; \underbrace{(e_i - q_i^K)}_{d_i^{\mathrm{id}}} \tag{6}$$

satisfies $\langle d_i^{\mathrm{sem}}, d_i^{\mathrm{id}} \rangle = 0$.

Proof. Recall $q_i^K(j) = (1/m_i)\,\mathbf{1}\{K_{ij} = 1\}$ and $\pi_B(j) = 1/B$. Expand:

$$\langle q_i^K - \pi_B,\ e_i - q_i^K \rangle \;=\; \langle q_i^K, e_i \rangle - \langle q_i^K, q_i^K \rangle - \langle \pi_B, e_i \rangle + \langle \pi_B, q_i^K \rangle.$$

The four inner products evaluate as

$$\langle q_i^K, e_i\rangle = q_i^K(i) = 1/m_i, \qquad \langle q_i^K, q_i^K\rangle = m_i \cdot (1/m_i)^2 = 1/m_i,$$
$$\langle \pi_B, e_i\rangle = \pi_B(i) = 1/B, \qquad \langle \pi_B, q_i^K\rangle = (1/B)\textstyle\sum_j q_i^K(j) = 1/B,$$

giving $\langle d_i^{\mathrm{sem}}, d_i^{\mathrm{id}}\rangle = 1/m_i - 1/m_i - 1/B + 1/B = 0$. □

Component magnitudes. Assume $m_i < B$. If $m_i = B$, every batch text is compatible, $q_i^K = \pi_B$, and the semantic component is exactly zero, so the identity-pressure ratio below is infinite.

$$\|d_i^{\mathrm{sem}}\|_2^2 \;=\; \frac{B - m_i}{m_i\, B}, \qquad \|d_i^{\mathrm{id}}\|_2^2 \;=\; 1 - \frac{1}{m_i}. \tag{7}$$

Derivation. For d_i^{sem}, partition the B coordinates into m_i compatible texts and $B - m_i$ incompatible texts:

$$\|q_i^K - \pi_B\|_2^2 = m_i\left(\frac{1}{m_i} - \frac{1}{B}\right)^2 + (B - m_i)\cdot\frac{1}{B^2}$$
$$= \frac{(B-m_i)^2}{m_i B^2} + \frac{B - m_i}{B^2} \;=\; \frac{(B-m_i)\big[(B-m_i) + m_i\big]}{m_i B^2} \;=\; \frac{B - m_i}{m_i B}.$$

For d_i^{id}, the difference $e_i(j) - q_i^K(j)$ is $1 - 1/m_i$ at $j = i$, $-1/m_i$ at the other $m_i - 1$ compatible j, and 0 at incompatibles, so

$$\|e_i - q_i^K\|_2^2 = \left(1 - \frac{1}{m_i}\right)^2 + (m_i - 1)\cdot\frac{1}{m_i^2} = \frac{(m_i-1)^2 + (m_i - 1)}{m_i^2} = 1 - \frac{1}{m_i}.$$

Taking the ratio gives the exact expression in Proposition 2:

$$\Gamma_i \;=\; \frac{\|d_i^{\mathrm{id}}\|_2^2}{\|d_i^{\mathrm{sem}}\|_2^2} \;=\; \frac{(m_i - 1)/m_i}{(B - m_i)/(m_i B)} \;=\; \frac{B(m_i - 1)}{B - m_i}. \tag{8}$$

Batch scaling. Let ω be the probability that an independently sampled off-diagonal text is compatible with the anchor image. Then K_{ij} for $j \neq i$ has mean ω, and $K_{ii} = 1$, so

$$\mathbb{E}[m_i] = 1 + (B - 1)\omega \sim B\omega.$$

Substituting the typical value $m_i \approx B\omega$ into the exact ratio gives

$$\Gamma_i \;\sim\; \frac{\omega}{1 - \omega}\, B \quad \text{for large } B. \tag{9}$$

By orthogonality, $\|d_i^{\mathrm{van}}\|_2^2 = \|d_i^{\mathrm{sem}}\|_2^2 + \|d_i^{\mathrm{id}}\|_2^2$. Therefore the semantic share of the vanilla descent-direction energy is

$$\frac{\|d_i^{\mathrm{sem}}\|_2^2}{\|d_i^{\mathrm{van}}\|_2^2} \;=\; \frac{1}{1 + \Gamma_i} \;\sim\; \frac{1 - \omega}{\omega}\cdot\frac{1}{B} \quad \text{for large } B, \tag{10}$$

which decays as $1/B$. Intuitively, larger batches add more compatible alternatives. The semantic part still asks the model to bring all compatible texts together, but the identity part asks it to choose the paired text among them; the latter increasingly dominates the vanilla direction as B grows. □

B.5 Partial positives in bundled reports

For bundled reports, exact compatibility K is conservative: a report can fail $K_{ij} = 1$ on a single contradicted finding while still sharing most asserted facts with the anchor image. The partial-agreement fraction is

$$R_{ij} \;=\; \frac{1}{|A(t_j)|}\sum_{(a,v)\in A(t_j)} \mathbf{1}\{z_{i,a} = v\} \;\in\; [0, 1]. \tag{11}$$

For example, if a candidate report asserts five facts and four agree with the anchor image while one disagrees, then exact compatibility gives $K_{ij} = 0$, but partial agreement gives $R_{ij} = 4/5$. Treating

this report as a pure negative would therefore suppress mostly shared facts, not mostly distinguishing facts.

Soft compatibility reference. Set $C^R_{ii} = 1$, $C^R_{ij} = R_{ij}$ for $j \neq i$, and row-normalize:

$$q^R_i(j) \;=\; \frac{C^R_{ij}}{\sum_{\ell=1}^{B} C^R_{i\ell}}. \tag{12}$$

Let $S^R_i = \sum_{j\neq i} R_{ij}$. Then $q^R_i(i) = 1/(1+S^R_i)$ and $\sum_{j\neq i} q^R_i(j) = S^R_i/(1+S^R_i)$. Unlike q^K_i, q^R_i is not an exact-entailment target; it is a soft reference measuring how much shared fact content the one-hot objective assigns the negative sign.

Fact-level sign decomposition. The soft score R_{ij} is useful because it separates shared facts from distinguishing facts. To see this with the simplest algebra, suppose a bundled report score decomposes additively over its asserted facts:

$$s_{ij} = \frac{1}{|A(t_j)|} \sum_{(a,v)\in A(t_j)} h_i(a,v),$$

where $h_i(a,v)$ is the model's score that image x_i supports fact (a,v). For an off-diagonal report $j \neq i$, vanilla one-hot has target $e_i(j) = 0$, so

$$\frac{\partial \ell_i(e_i)}{\partial s_{ij}} = p_i(j).$$

Therefore every asserted fact in t_j receives

$$\frac{\partial \ell_i(e_i)}{\partial h_i(a,v)} = \frac{p_i(j)}{|A(t_j)|}, \qquad (a,v) \in A(t_j).$$

Gradient descent lowers all these fact scores. This is right-signed for facts contradicted by x_i, but wrong-signed for facts that are true of x_i. The fraction of asserted facts in t_j that are true for x_i is exactly R_{ij}, while the distinguishing fraction is $1 - R_{ij}$. Thus the off-diagonal update carries wrong-signed shared-fact mass $p_i(j)R_{ij}$ and right-signed distinguishing mass $p_i(j)(1-R_{ij})$. This is why a report can be a poor negative even when it is not exactly compatible.

Partial-report bias is a deterministic batch-composition function.

Proposition 3 (Partial-report bias)**.** *For bundled reports, the deviation between paired identity and the soft compatibility reference is*

$$\|q^R_i - e_i\|_1 \;=\; 2\,\frac{S^R_i}{1+S^R_i}, \tag{13}$$

which is a deterministic function of the partial-agreement fractions $\{R_{ij}\}_{j\neq i}$ in the batch. Under independently sampled off-diagonal texts at granularity g,

$$\mathbb{E}[S^R_i] \;=\; (B-1)\rho_g, \qquad \rho_g \;=\; \mathbb{E}[R(Z, T_g^-)]. \tag{14}$$

The bias is therefore present in every batch, scales with B, and does not average out with more data or longer training.

Proof. Since $e_i(i) = 1$ and $e_i(j) = 0$ for $j \neq i$,

$$\|q^R_i - e_i\|_1 \;=\; \big|q^R_i(i) - 1\big| + \sum_{j\neq i} q^R_i(j) \;=\; \left(1 - \frac{1}{1+S^R_i}\right) + \frac{S^R_i}{1+S^R_i} \;=\; \frac{2S^R_i}{1+S^R_i}.$$

Equation (14) follows from linearity of expectation applied to $S^R_i = \sum_{j\neq i} R_{ij}$ using $\mathbb{E}[R_{ij}] = \rho_g$ for off-diagonal j. □

Relation to the exact-compatibility proof. The decomposition of Section 3.4 still holds algebraically for the soft reference:

$$e_i - \pi_B \;=\; \underbrace{(q^R_i - \pi_B)}_{d_i^{\mathrm{sem},R}} \;+\; \underbrace{(e_i - q^R_i)}_{d_i^{\mathrm{id},R}}. \tag{15}$$

When R is binary, $q_i^R = q_i^K$, the two components are orthogonal, and Proposition 2 is recovered exactly. For general bundled reports, R_{ij} is fractional and q_i^R is not uniform over a compatible set, so the squared-norm ratio depends on the full distribution of partial-agreement scores, not only on their sum S_i^R. We therefore use S_i^R and $\|q_i^R - e_i\|_1$ as the clean report-level pressure statistics. They are sufficient for the claim made in Section 3.4: increasing B adds more partial-positive mass, and by the softmax target-gap identity that mass is exactly the logit-space bias introduced by paired identity.

C Extractor Schemas and Compatibility Construction

This appendix summarizes how compatibility matrices are instantiated in practice for CXR and CT settings. The main design choice is the fact schema: a coarse schema removes more false negatives by merging related descriptions, while a finer schema preserves clinical distinctions such as laterality, organ system, and device or disease subtype. The schemas below are intentionally simple fixed instantiations of this tradeoff.

C.1 CXR Fact Schemas

Table 6: **Category coverage of the two CXR fact schemas.**

Clinical group	**CheXbert-13d**	**MedGemma-30d**
Lung parenchyma	Lung opacity, lung lesion, edema, consolidation, pneumonia, atelectasis, pneumothorax.	Right/left opacity, consolidation, atelectasis, edema, pneumothorax, mass or nodule; global vascular congestion, hyperinflation, interstitial markings.
Pleura	Pleural effusion, pleural other.	Right/left effusion, right/left pleural thickening.
Cardiac / mediastinal	Enlarged cardiomediastinum, cardiomegaly.	Cardiac enlargement, mediastinal widening, hilar enlargement.
Devices	One support-devices slot.	Airway tube, vascular line, enteric tube, chest tube, cardiac device.
Skeletal / other	Fracture.	Fracture, degenerative spine change, diaphragmatic abnormality.

Table 6 lists only the category coverage. CheXbert [31] is a fixed labeler over the CheXpert-style label space; "No Finding" is retained only as auxiliary metadata. MedGemma-27B-Text-IT [32] is a text-only LLM used as a structured report parser: it extracts anatomy, finding, laterality, and status, which are then resolved into the 30-slot schema above. Sentence-level facts are aggregated over the paired full report to approximate the image-side state $\hat{z}_i$. The empirical pattern in Appendix A.5 matches the intended tradeoff: CheXbert-13d is strongest when the evaluation label space matches the coarse schema, while MedGemma-30d is stronger when transfer or localization requires finer clinical structure.

MedGemma-30d prompt. The prompt defines the finding and anatomy vocabularies of the 30-slot schema and asks for a JSON list of facts per sentence. Its rules extract negated findings as absent, expand broad negations (e.g., "lungs are clear") only to findings within their anatomical scope, and keep stable or unchanged findings present. The full prompt is released with our code.

Extraction cost. Both extractors are run offline before contrastive training. On a single H100 80GB GPU, CheXbert processes the 120K-report pool in under half an hour. MedGemma-27B-Text-IT is a larger model, but with vLLM inference acceleration [50] it parses the same pool in about 1.5 hours. This preprocessing adds no cost during contrastive training or test-time evaluation, and is small relative to the cost of pretraining the vision-language model.

C.2 CT Fact Schemas

For the 3D CT transfer experiment, we use a CT-specific MedGemma parser rather than the CXR schemas above. CT reports are commonly organized by organ-system sections, so the parser first segments each report into sections, uses section headers as metadata for implicit anatomy, extracts

sentence-level triples $(\text{anatomy}, \text{finding}, \text{laterality}, \text{status})$, and resolves them into the Merlin-30d finding schema used for zero-shot evaluation.

Table 7: **Category coverage of the CT Merlin-30d fact schema.**

Clinical group	**Merlin-30d categories**
Lower thorax / cardiac	Atelectasis, pleural effusion, cardiomegaly, coronary calcification, aortic-valve calcification.
Hepatobiliary / pancreas / spleen	Hepatomegaly, hepatic steatosis, biliary ductal dilation, gallstones, surgically absent gallbladder, pancreatic atrophy, splenomegaly.
Renal / adrenal / pelvic organs	Renal hypodensities, renal cyst, hydronephrosis, prostatomegaly.
Vascular / lymph / oncology	Thrombosis, abdominal aortic aneurysm, atherosclerosis, lymphadenopathy, metastatic disease.
GI / peritoneum / abdominal wall / MSK	Submucosal edema, bowel obstruction, appendicitis, hiatal hernia, ascites, anasarca, free air, fracture, osteopenia.

The CT extractor uses MedGemma-27B-Text-IT [32] as a text-only structured parser. Its section-aware design is important for CT: a sentence such as "normal size and appearance" is interpreted differently under liver, kidneys, adrenal glands, or lymph nodes. We therefore use section headers to supply default anatomy for implicit statements and to expand organ-level normal statements into absent findings for the relevant Merlin-30d categories. Present findings override absent findings when sentence-level configs are aggregated into a report-level state. Uncertain triples are not treated as positive assertions in the CT parser. As in CXR, parsing is an offline preprocessing step and MedGemma is not used during contrastive training or evaluation.

Compatibility scoring. Across CXR and CT, compatibility is evaluated only on facts asserted by the candidate text t_j. Unmentioned findings in t_j contribute neither agreement nor disagreement. Sentence-level correction uses hard entailment, $\hat{C}_{ij} = 1$ if all asserted facts in t_j are satisfied by $\hat{z}_i$, and 0 otherwise. Impression- and report-level correction uses soft agreement, the fraction of asserted facts in t_j satisfied by $\hat{z}_i$. In CXR experiments, uncertain findings are mapped to absent/negative; in the CT parser, uncertain triples are ignored. If no clinical fact is extracted from a text unit, we keep only its paired entry positive rather than treating the empty assertion set as compatible with every image. We also force $\hat{C}_{ii} = 1$ before normalization to guard against extractor inconsistency.

Symmetric targets. The same nonnegative compatibility matrix $\hat{C}$ is used in both directions. Image-to-text normalizes each row,

$$\hat{q}_{ij}^{I\to T} = \frac{\hat{C}_{ij}}{\sum_{\ell=1}^{B} \hat{C}_{i\ell}},$$

and the text-to-image target normalizes each column,

$$\hat{q}_{ji}^{T\to I} = \frac{\hat{C}_{ij}}{\sum_{\ell=1}^{B} \hat{C}_{\ell j}}.$$

This is the practical version of the row-wise references q_i^K and q_i^R used in the theory; the text-to-image side is the symmetric counterpart.

Scope. These schemas are intentionally simple instantiations of the compatibility idea. They use binary or fractional agreement over a fixed set of extracted facts, rather than a learned clinical ontology, severity-aware entailment, uncertainty-aware matching, or a kernel that weights facts by hierarchy and diagnostic importance. The CXR and CT schemas are modality-specific, but the construction is the same: parse reports into asserted facts, estimate the image-side fact state from the paired report, and replace paired identity with compatibility over asserted facts. Designing richer schemas and compatibility kernels is a natural direction for future work, as discussed in Section 7.

D *CLEVR-Radiology*: Construction and Statistics

This appendix documents the synthetic testbed introduced in Section 2: the generation pipeline, the slot and template inventories used by the semantic-floor computation in Section 3.3, qualitative samples, dataset statistics, and the floor derivation. The released dataset contains $60{,}000$ image-report pairs, split into $40{,}000/10{,}000/10{,}000$ train/validation/test pairs. Images are rendered at 224×224 and generated by a deterministic, seedable pipeline: scene-config sampling, template-based report rendering, Blender image rendering, and image-noise injection.

D.1 Generation pipeline

The pipeline runs in four stages.

Stage 1 (scene config). A scene contains 4 objects with fixed identities at fixed ground-plane positions: `left_cube`, `right_cube`, `sphere`, `cylinder`. Each object carries 3 attribute slots (`color`, `size`, `material`), giving 12 independent slots per scene. Each slot is sampled from an independent Bernoulli: with probability $1-p_{\mathrm{abn}}$ the slot takes its normal value (`gray`, `medium`, `rubber` for color, size, material respectively); otherwise an abnormal value is drawn uniformly from that attribute's abnormal vocabulary. Slots are split into 6 *common* (p_{abn}=0.20) and 6 *rare* (p_{abn}=0.03); the assignment is in Table 8. There are no co-occurrence constraints, so the joint config distribution factorizes across slots.

Stage 2 (report). A template grammar emits a FINDINGS section followed by an IMPRESSION section. FINDINGS iterates over objects and slots: every abnormal slot is mentioned via one of 6 paraphrase templates; normal slots are mentioned with probability 0.2 via either an explicit-normal template (3 concrete, e.g., "The sphere is medium in size."), a positive-abstract template (3, e.g., "The sphere is normal in size."), or a negative-abstract template (3, e.g., "No abnormal size of the sphere."). With probability 0.5 all slots of a fully-normal object are collapsed into a single object-level all-clear sentence (5 templates, e.g., "The sphere is unremarkable."). The two cubes are merged bilaterally with probability 0.5 when they share an abnormal value (4 bilateral-abnormal templates) or are jointly normal (5 bilateral-all-clear templates). IMPRESSION lists abnormal slots joined with commas and "and", wrapped in one of 4 frames (e.g., "In summary, ⟨findings⟩."); fully-normal scenes use one of 6 all-clear impression templates ("Unremarkable examination.", "Normal study.", etc.). Every emitted sentence keeps its slot tuple metadata $\{(\text{obj}, \text{attr}, \text{value}, \text{is_abnormal})\}$, so for any (image, sentence) pair the entailment $z_i \models A(t_j)$ is known exactly.

Stage 3 (image). Blender (Cycles, 128 samples) renders each scene at 224×224. Object meshes are the standard CLEVR primitives [23] (smoothed cube, sphere, smoothed cylinder); colors come from the standard CLEVR palette; size is applied as a scale factor; material is either rubber (default Lambertian) or metal (specular shader). Camera position, three-light rig, and per-object position receive small Gaussian jitter (std 0.15–1.0) seeded by image index for reproducibility. Per-image metadata includes 2D bounding boxes, 3D world coordinates, and pixel-projected centers for each object, supporting downstream grounding evaluation.

Stage 4 (image noise). The rendered CLEVR scenes are visually simple and otherwise contain few image artifacts that could support shortcut learning. To make the visual side less pristine, we add random colored patches after rendering. The number, size, position, and color of these patches are sampled independently for each image and are not tied to the scene config or report text. This perturbation introduces nuisance visual variation while preserving the ground-truth object attributes and fact metadata.

The design mirrors the data-generation structure of radiology without preserving radiology-specific artifacts. The object layout is largely fixed, with small geometric variation, playing the role of shared anatomy. Each object has multiple attribute slots, analogous to finding states over an anatomical region; each slot has only a small value vocabulary, so the semantic space is compositional but not open-ended. The report generator follows radiology-style practice by bundling assertions, always mentioning abnormal states and selectively mentioning normal states. Finally, the random patches introduce image-specific non-semantic variation. Thus *CLEVR-Radiology* retains the shared-schema fact-bundle structure while removing clinical appearance, acquisition artifacts, and radiologist-specific language.

D.2 Slot table

Table 8 lists all 12 slots with their prior, abnormal vocabulary, and CXR analog. The CXR analog is purely a label for clinical readability; it does not enter any code path or training procedure.

Table 8: The 12 attribute slots of *CLEVR-Radiology*, with two-tier prevalence and CXR analogs.

Object	Attribute	Tier	Prior p_{abn}	Abnormal values	CXR analog
`left_cube`	`color`	common	0.20	red, blue, green	left lung opacity
`left_cube`	`size`	common	0.20	small, large	left atelectasis
`right_cube`	`color`	common	0.20	red, blue, green	right lung opacity
`right_cube`	`size`	common	0.20	small, large	right atelectasis
`sphere`	`size`	common	0.20	small, large	cardiomegaly
`cylinder`	`color`	common	0.20	red, blue, green	support devices
`left_cube`	`material`	rare	0.03	metal	left pleural effusion
`right_cube`	`material`	rare	0.03	metal	right pleural effusion
`sphere`	`color`	rare	0.03	red, blue, green	pulmonary edema
`sphere`	`material`	rare	0.03	metal	pleural other
`cylinder`	`size`	rare	0.03	small, large	fracture
`cylinder`	`material`	rare	0.03	metal	rare finding

D.3 Template inventory

Table 9 enumerates the template families and their cardinalities. "Mention probabilities" below refers to the per-slot Bernoulli that decides whether a normal slot is mentioned at all, and to the per-object Bernoulli that decides whether all-clear collapsing is used.

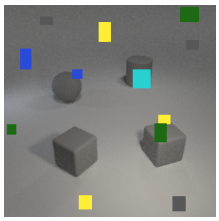
Table 9: Template families used by the *CLEVR-Radiology* report generator.

Family	Section	# templates	Example
Abnormal mention	FINDINGS	6	"The {shape} has {value} {attr}."
Normal individual, concrete	FINDINGS	3	"The {shape} is {normal_value} in {attr}."
Normal individual, positive	FINDINGS	3	"The {shape} is normal in {attr}."
Normal individual, negative	FINDINGS	3	"No abnormal {attr} of the {shape}."
Object all-clear	FINDINGS	5	"The {shape} is unremarkable."
Bilateral abnormal (cubes)	FINDINGS	4	"Bilateral {value} {attr} of the cubes."
Bilateral all-clear (cubes)	FINDINGS	5	"Both cubes are unremarkable."
Impression abnormal	IMPRESSION	4	"In summary, {findings}."
Impression all-normal	IMPRESSION	6	"Unremarkable examination."

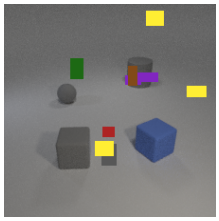
Mention probabilities are: normal-mention 0.20, object-all-clear 0.50, bilateral-merge 0.50 (when applicable), bilateral-all-clear 0.40, concrete-vs-abstract for normals 0.40, negation-vs-positive for abstract normals 0.30. These probabilities and template counts are configurable hyperparameters of the report generator, and they are the only inputs (alongside the slot priors) that the floor derivation in Appendix D.7 consumes.

D.4 Qualitative samples

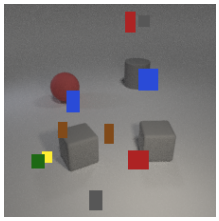
Figure 9 shows five image-report pairs that span the dominant report styles. The five samples were chosen to cover: an all-normal scene (using object-level and bilateral all-clear templates), a typical $n_{\text{findings}}=2$ scene with two common findings, a rare-only finding (a red sphere, where "red" is rare for the sphere's color slot), a bilateral-merged cube finding, and a multi-finding scene mixing common and rare slots. Additional dataset samples are in Figure 10.



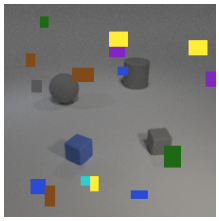
(A) **All-normal.** `n_findings=0`. *FINDINGS:* Both cubes are unremarkable. The sphere shows no abnormal color. The cylinder appears normal. *IMPRESSION:* Normal study.



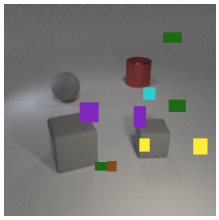
(B) **Two common findings.** `n_findings=2`: `right_cube.color=blue`, `sphere.size=small`. *FINDINGS:* The right cube demonstrates blue color. The sphere color is gray. The sphere has small size. The cylinder is normal in color. *IMPRESSION:* blue color of the right cube and small size of the sphere. No other acute abnormality.



(C) **Rare-only finding.** `n_findings=1`: `sphere.color=red` (a rare slot, p_{abn}=0.03). *FINDINGS:* Normal material of the left cube. Normal size of the left cube. No abnormality of the right cube. There is red color involving the sphere. The cylinder appears normal. *IMPRESSION:* red color of the sphere. No other acute abnormality.



(D) **Bilateral merge.** `n_findings=3`: both cubes share `size=small`, plus `left_cube.color=blue`. *FINDINGS:* Both cubes demonstrate small size. The left cube color is blue. No abnormality of the sphere. The cylinder is normal in material. *IMPRESSION:* blue color of the left cube, small size of the left cube and small size of the right cube.



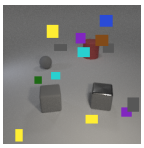
(E) **Multi-finding (common + rare).** `n_findings=3`: `left_cube.size=large` (common), `cylinder.color=red` (common), `cylinder.material=metal` (rare). *FINDINGS:* Large size of the left cube. The left cube material is rubber. The right cube shows no abnormal material. The sphere is medium in size. Metal material of the cylinder. Red color is seen in the cylinder. *IMPRESSION:* large size of the left cube, red color of the cylinder and metal material of the cylinder. No other acute abnormality.

Figure 9: Five *CLEVR-Radiology* samples illustrating the dominant report styles. Each row pairs the rendered image with the full FINDINGS and IMPRESSION text.

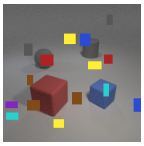
D.5 Additional random samples

D.6 Dataset statistics

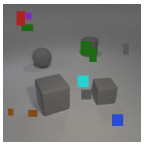
Table 10 summarizes the released split. Three numbers are worth highlighting for the theory: (i) report texts are nearly all unique (94.6%), so identity-target training has near-zero exact text duplication, yet (ii) the underlying *config* space collapses to 3,027 distinct configs and the *binary abnormality pattern* space to only 707 patterns, meaning many "different" reports describe the same finding set. This gap between text-level distinctness and fact-level distinctness is exactly the regime where the semantic floor of Section 3.3 becomes load-bearing. (iii) The all-normal rate observed at 21.9% matches the closed-form prediction $(1-0.20)^6 \cdot (1-0.03)^6 \approx 21.8\%$ to within sampling noise, a sanity check that the generator and the prior tabulation agree.



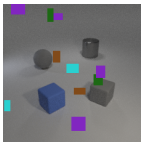
FINDINGS: The left cube is unremarkable. Metal material of the right cube. Gray color of the right cube. There is small size involving the sphere. The sphere is gray in color. Normal material of the cylinder. The cylinder has red color. *IMPRESSION:* metal material of the right cube, small size of the sphere and red color of the cylinder. No other acute abnormality.



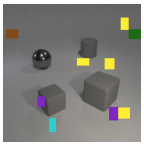
FINDINGS: Large size of the left cube. The left cube has red color. There is blue color involving the right cube. The sphere is clear. No abnormality of the cylinder. *IMPRESSION:* red color of the left cube, large size of the left cube and blue color of the right cube.



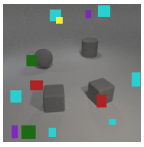
FINDINGS: The left cube color is unremarkable. The left cube has large size. The right cube appears normal. The sphere material is unremarkable. The cylinder is within normal limits. *IMPRESSION:* In summary, large size of the left cube.



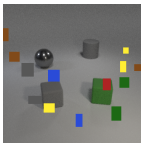
FINDINGS: Medium size of the left cube. The left cube has blue color. The right cube size is unremarkable. No abnormal size of the sphere. Metal material is seen in the cylinder. *IMPRESSION:* blue color of the left cube and metal material of the cylinder.



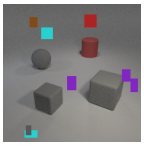
FINDINGS: Rubber material of the right cube. There is large size involving the right cube. The sphere has metal material. The sphere color is gray. No abnormality of the cylinder. *IMPRESSION:* Overall, large size of the right cube and metal material of the sphere.



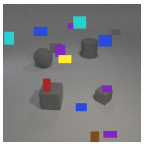
FINDINGS: The left cube is unremarkable. The right cube is medium in size. The sphere is unremarkable. Rubber material of the cylinder. *IMPRESSION:* Clear.



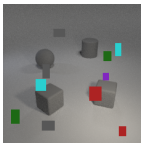
FINDINGS: The left cube is unremarkable. The right cube demonstrates green color. Metal material is seen in the sphere. The cylinder is gray in color. *IMPRESSION:* Overall, green color of the right cube and metal material of the sphere.



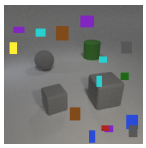
FINDINGS: Normal color of both cubes. There is large size involving the right cube. The sphere appears normal. The cylinder demonstrates red color. *IMPRESSION:* In summary, large size of the right cube and red color of the cylinder.



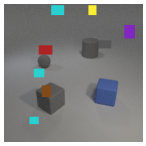
FINDINGS: Small size is seen in the right cube. The sphere is within normal limits. The cylinder appears normal. *IMPRESSION:* Overall, small size of the right cube.



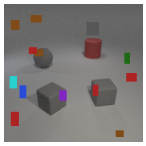
FINDINGS: The left cube is unremarkable. No abnormal material of the sphere. *IMPRESSION:* No acute findings identified.



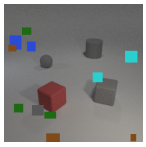
FINDINGS: The left cube is gray in color. There is large size involving the right cube. The right cube is gray in color. The sphere material is rubber. There is green color involving the cylinder. The cylinder material is unremarkable. *IMPRESSION:* In summary, large size of the right cube and green color of the cylinder.



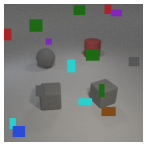
FINDINGS: The left cube shows no abnormal size. Blue color is seen in the right cube. No abnormal material of the right cube. There is small size involving the sphere. The cylinder appears normal. *IMPRESSION:* Overall, blue color of the right cube and small size of the sphere.



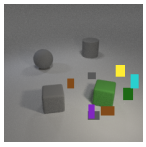
FINDINGS: The left cube is unremarkable. The sphere shows no abnormal size. No color abnormality of the sphere. Red color is seen in the cylinder. No abnormal material of the cylinder. *IMPRESSION:* Overall, red color of the cylinder.



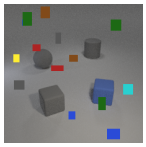
FINDINGS: Red color is seen in the left cube. Gray color of the right cube. The right cube is normal in size. The sphere has small size. The cylinder is unremarkable. *IMPRESSION:* Overall, red color of the left cube and small size of the sphere.



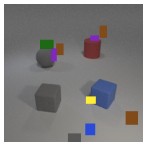
FINDINGS: The cubes are clear bilaterally. The sphere is clear. The cylinder shows no abnormal size. Red color of the cylinder. *IMPRESSION:* In summary, red color of the cylinder.



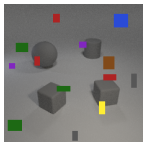
FINDINGS: The left cube appears normal. There is green color involving the right cube. The sphere is unremarkable. The cylinder is clear. *IMPRESSION:* green color of the right cube.



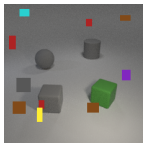
FINDINGS: The left cube is clear. The right cube demonstrates blue color. No abnormality of the sphere. The cylinder appears normal. *IMPRESSION:* In summary, blue color of the right cube.



FINDINGS: There is blue color involving the right cube. The sphere is within normal limits. The cylinder shows no abnormal size. Red color is seen in the cylinder. *IMPRESSION:* blue color of the right cube and red color of the cylinder.



FINDINGS: The left cube is unremarkable. No abnormality of the right cube. The sphere demonstrates large size. *IMPRESSION:* large size of the sphere.



FINDINGS: The left cube is clear. Green color is seen in the right cube. *IMPRESSION:* Overall, green color of the right cube.

Figure 10: Twenty uniformly sampled *CLEVR-Radiology* image-report pairs from the released split (fixed random seed), illustrating the diversity of the generated data.

Table 10: *CLEVR-Radiology* dataset statistics over the 60,000 released image-report pairs.

Scale	
Image-report pairs	60,000
Image format / resolution	PNG, 224×224
Diversity	
Unique report texts	56,751
Unique configs (exact attribute values)	3,027
Unique binary abnormality patterns	707
Unique sentence texts	6,202
Report length (words)	
Mean / median / range	35 / 34 / 7–99
Sentences per report (mean / median / range)	5.0 / 5 / 2–11
Finding count distribution	
$n_{\text{findings}}=0$ (all-normal)	21.9%
$n_{\text{findings}}=1$	37.0%
$n_{\text{findings}}=2$	26.8%
$n_{\text{findings}}=3$	11.0%
$n_{\text{findings}}\geq 4$	3.4%

D.7 Generator-based semantic floor

By Proposition 1, the row-i semantic floor of vanilla CLIP is $\log m_i = H(q_i^K)$, where q_i^K is the exact-compatibility reference over the B batch texts. For *CLEVR-Radiology* the generator records the full image config and the asserted fact set of every emitted text, so the compatibility probability can be computed directly from generator metadata. We summarize the report-level and sentence-level calculations used for the dashed lines in Figure 2(A, B).

Computation. For a fixed anchor, the floor depends only on how many other batch texts are compatible with it. We therefore estimate, under the generator distribution, the number of compatible off-diagonal texts in a B-sized batch and take the expected $\log$ positive count after adding the paired text itself. This is the same generator-based calculation used for the dashed lines in Figure 2(A, B).

Report level. For report-level training, compatibility is computed from the complete image config and the asserted fact set of a generated report. The config prior is known exactly from the slot sampling probabilities in Table 8, so the report floor is computed by enumerating the config prior and evaluating how often independently generated reports are entailed by each config. This gives the report-level dashed floor without using any model outputs.

Sentence level. For sentence-level training, the relevant object is a sentence pattern: the set of slot-value facts asserted by one generated sentence. We estimate the distribution of these patterns using the same generator defaults as the released data, including normal-mention probability, object-level all-clear sentences, bilateral merges, and the rule that training samples one sentence from a report. The sentence floor then uses the probability that a random sentence pattern is entailed by a random image config, averaged symmetrically over image-to-text and text-to-image directions.

Use in Figure 2. The dashed floors are precomputed from the slot priors, template probabilities, emitted assertion metadata, and batch size $B{=}256$. They are not fit to the training loss. The training curves are used only afterward to check whether the observed loss at the eval-peak step lies near the generator-predicted semantic floor.

E Experimental Setup Details

E.1 CXR Experiments

Data. MIMIC-CXR contains many image views and report formats. To keep the study focused on the contrastive target, we restrict pretraining to frontal images and reports containing both FINDINGS and IMPRESSION sections. This filtering yields a training pool slightly above 120K image–report pairs. Because MS-CXR is annotated on top of MIMIC-CXR, we remove all MS-CXR images from

the pretraining pool to avoid overlap with the grounding evaluation. We use two training scales, $60\mathrm{K}$ and $120\mathrm{K}$, and keep the official validation and test splits for validation and in-domain evaluation.

Text units. Report uses FINDINGS+IMPRESSION, impression uses IMPRESSION alone, and sentence samples one report sentence per training iteration.

Evaluation data. CheXpert and VinDr are used only for zero-shot out-of-distribution classification, never for training. We evaluate on their official labeled test splits: CheXpert uses the same 14-label disease vocabulary as MIMIC-CXR, while VinDr provides a finer 28-label CXR vocabulary. For dense evaluation, we use MS-CXR phrase-grounding annotations, which pair radiology finding phrases with radiologist-verified bounding boxes. Since MS-CXR is small, we use all annotated examples for testing.

Evaluation protocol. For zero-shot classification, all methods, including false-negative mitigation baselines, use the same class-name prompts and scoring rule. For phrase grounding, we compute a phrase heatmap by taking the similarity between the text embedding of the queried phrase and the dense image embeddings, then resizing the resulting similarity map to the input image size. We compute pointing-game accuracy on MS-CXR: a prediction is counted as correct if the maximum value of the phrase heatmap falls inside any ground-truth bounding box for that phrase.

For in-domain MIMIC-CXR training-dynamics analyses, "peak" denotes the logged checkpoint with the highest test-set metric and "final" denotes the last checkpoint. We use this retrospective peak only as a diagnostic, not as a deployable model-selection rule: the goal is to measure how far the final checkpoint falls from the best representation reached along the trajectory, and the full trajectory is shown whenever this comparison is central. For OOD classification and grounding tables, checkpoints are selected by validation performance and the corresponding test-set results are reported.

CXR training. All CXR methods are trained in the same framework. Vanilla CLIP, mitigation baselines, and CorLIP variants share the same DINOv3-initialized ViT-B/16 image encoder [33, 28], BioClinicalBERT text encoder [29, 34], 224×224 image resolution, augmentations, optimizer, schedule, and checkpointing. They differ only in the contrastive target used for each batch. We use AdamW [51] with default momentum parameters and weight decay 10^{-4}. The base learning rate is 10^{-4} at $B{=}128$; other batch sizes use batch-size-adapted learning rates capped at 4×10^{-4}. The schedule uses linear warmup over the first 5% of steps followed by cosine decay. Augmentations include random crop, brightness jitter, and Gaussian noise.

Baseline rationale. The vanilla CLIP rows at report, impression, and sentence granularity already instantiate the paired-identity objectives underlying several classic CXR VLP systems. Report- or impression-level paired training corresponds to CheXzero-style image–report CLIP [4], while sentence-level paired training corresponds to ConVIRT-style sentence sampling [2]. We do not reuse public checkpoints because they differ in data filtering, encoder initialization, optimization, image processing, and training scale; instead, we reinstantiate the paired-identity objective inside our controlled pipeline so that the target is the only intended difference. For false-negative mitigation baselines, we reimplement SAT [26], MedCLIP [19], and Ko&Park [20] in the same codebase. Method-specific target construction and hyperparameters follow the original papers, while all shared training choices follow the CXR setup above. These baselines are not incompatible with our framework; they can be viewed as alternative choices of text granularity and target kernel, using report similarity, semantic matching, graph similarity, or dynamic soft labels instead of explicit fact compatibility. CorLIP isolates the compatibility kernel predicted by our theory while keeping the rest of the training pipeline fixed.

We do not compare against every radiology VLP system that improves performance through an orthogonal axis. Methods such as GLoRIA [3] and BioViL [5] add global-local or dense alignment objectives; MedKLIP [24] and IMITATE [25] inject medical knowledge or clinical priors; larger foundation models change data scale, backbones, or language modeling capacity. These choices are valuable, but they test a different axis from the one studied here. CorLIP is meant to be a target-level correction and can in principle be combined with stronger architectures, local objectives, or larger pretraining corpora.

Supervised baseline. The supervised control uses the same CXR training data, image backbone, augmentations, optimizer, schedule, and checkpointing as the contrastive runs. The only architectural change is replacing the contrastive projection with a multi-label classification head. We train this head with supervised cross-entropy using the CheXpert 14-class labels provided with MIMIC-

CXR [27, 30]. This baseline serves only as a non-contrastive control for the training dynamics in Observation 1.

E.2 *CLEVR-Radiology* Experiments

For *CLEVR-Radiology*, we use an OpenCLIP ViT-B/32 image encoder [18] with random initialization, paired with the same BERT-style text encoder family unless otherwise stated. The image resolution, optimizer, schedule, and augmentation recipe follow the CXR setup. Dataset construction details are in Appendix D.

E.3 3D CT Transfer Experiment

To test whether target correction transfers beyond 2D CXR, we run a supporting experiment on the Merlin 3D CT image–report pretraining setup [9]. The training set contains $25{,}528$ CT scans paired with radiology reports. Following Merlin, we use the FINDINGS section as the text source, since it contains detailed organ-system observations. We evaluate three configurations while keeping image inputs, preprocessing, encoders, optimizer, and schedule fixed: full-report vanilla contrastive training, random sentence-level vanilla contrastive training, and sentence-level CorLIP, for which MedGemma-27B-Text-IT [32] is used offline as a structured text parser to extract CT finding assertions and construct sentence-level compatibility targets; it is not used during model training or evaluation.

The CT image encoder is an I3D ResNet-152 and the text encoder is a Clinical Longformer, with both encoders producing 512-dimensional embeddings. We train for 52 epochs using the same CT image preprocessing pipeline as Merlin. Optimization uses AdamW with $\beta = (0.9, 0.999)$, initial learning rate 10^{-5}, and a cosine learning-rate schedule parameterized over 300 epochs. Gradient checkpointing is applied to both encoders, and training uses FP16 mixed precision. These settings allow a maximum batch size of 18 on a single NVIDIA L40S 48GB GPU; evaluation reports zero-shot classification over 30 Merlin findings as mean AUC with 95% confidence intervals.

E.4 Compute Resources

CXR and *CLEVR-Radiology* experiments were run with BF16 precision on NVIDIA H100 80GB GPUs on nodes with approximately 500GB system memory. For report-level CXR training, each GPU supports a batch size of 128. For sentence- and impression-level training, the shorter text units allow a batch size of 256 per GPU, so the number of GPUs used for each global-batch experiment can be inferred from the reported batch size and the corresponding per-GPU capacity. The Merlin CT experiment was run separately with FP16 precision on a single NVIDIA L40S 48GB GPU because of the data storage requirements.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Main claims are stated in the abstract and introduction sections.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations are discussed in Section 7

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Theoretical assumptions are stated in Section 3, with complete proofs and notation in Appendix B.

Guidelines:

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

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Datasets, text granularities, training setup, baselines, evaluation protocols, and hyperparameters are described in Section 5 and Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data used in this paper are all publicly available, and the code to reproduce the main experimental results is publicly available. Instructions for reproducing the main experimental results are provided in Appendix E.

Guidelines:

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

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Training data, preprocessing, model initialization, optimizer, schedule, augmentations, baselines, and evaluation details are provided in Appendix E.

Guidelines:

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

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report multi-seed error bars or formal significance tests. The study reimplements all methods in a shared training framework for controlled comparison, and sweeps over many text granularities, batch sizes, data scales, extractors, and evaluation datasets; repeating the full grid across multiple seeds would exceed our compute budget. To make the conclusions inspectable, we report complete training trajectories, peak/final checkpoints, and consistent trends across granularities, datasets, and batch sizes.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

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

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments compute is described in Appendix E

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This work follows the NeurIPS Code of Ethics. We have reviewed the Code of Ethics and confirm that our research conforms to it in every respect.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Section 7 discusses potential benefits for medical representation learning.

Guidelines:

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

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model or scraped dataset.

    Guidelines:

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

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets, models, and tools are cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The synthetic dataset construction, statistics, templates, and examples are documented in Appendix D; the dataset is released together with the code.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not conduct new crowdsourcing or new human-subject experiments; it uses existing de-identified datasets and annotations.

    Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study uses existing de-identified public datasets and does not collect new human-subject data.

    Guidelines:

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

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: MedGemma-27B-Text-IT is used as a structured report parser for one extractor schema, as described in Appendix C; LLM use for writing/editing is not part of the scientific method.

    Guidelines:

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