FABRICATED NEGATIVES IN MULTI-LABEL CONTRASTIVE LEARNING: WHAT THE LOSS SEES, WHAT REMOVAL CHANGES, AND WHAT THE METRIC CAN TELL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-label contrastive training treats every label absent from a sample's record as a verified absence; under partial annotation this turns coordinates that nobody examined into negative supervision. We make the observation process explicit and report three findings, each of which reverses a conclusion the standard pipeline draws from the same runs. **(1) Fabricated negatives are common and must be counted where the loss receives them.** We define a fabrication rate from the label matrix and a recorded observation mask, split into an *undecidable* clause (coordinate not examined) and a *contradicted* clause (sample observed positive), and computed on CPU before training. On a renal-pathology corpus it is 0.26 over all pairs, 0.49 at the loss, and 0.52 as realized in training; on Open Images, 96% of train-tier candidate negatives cannot be certified, and the usual coverage approximation misses by 0.37. **(2) Which pairs a rule removes, not how many, decides what removal does.** Random removal of up to 66% of the candidates the loss receives leaves a profile probe unchanged; at the same volume, a knowledge-graph rule that our pipeline had called certification lowers it by 0.25, because 72% of what it removes are certifiably correct negatives, while removing only the contradicted pairs that partially overlap the target raises all three probes over eight seeds. Where hidden labels are known, the undecidable pairs are worth removing too: removal raises macro AUC on every seed of a pathology testbed, on COCO the gain from not treating unexamined labels as absent grows with the wrong-negative mass the diagnostic predicts before training (Spearman 1.0, declared in advance), and on the federated record of LVIS it raises macro AUC by 0.019 on every seed against complete labels. **(3) The standard metric cannot see any of this.** Scoring each model against its own training vocabulary ranks a control trained on *meaningless text* above every real-text arm and makes removal look beneficial and dose-dependent (Spearman $+0.90$); on shared held-out text no removal ranks first. Scoring recorded profiles in their closed world errs the other way: models that gain on complete labels lose on every seed, and a constant prediction outscores them all. We distill the findings into a six-item reporting protocol that makes a negative-mining claim auditable whichever arm ranks first.

1 INTRODUCTION

Contrastive objectives learn from a decision about which pairs should be pushed apart (van den Oord et al., 2018; Chen et al., 2020; He et al., 2020; Radford et al., 2021). In multi-label data that decision is usually made by the shortcut that negative sampling inherited from word embeddings and implicit feedback (Mikolov et al., 2013; Hu et al., 2008): if label ℓ is not in sample i's record, then i is a negative for every sample or text that carries ℓ. The shortcut is harmless when the label matrix is complete. It is not harmless when the matrix records only what annotators looked at, because "absent from the record" then means either "examined and absent" or "not examined", and the loss collapses the two.

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

Missing labels are a well-studied problem for classifiers (Sun et al., 2010; Yu et al., 2014; Durand et al., 2019; Cole et al., 2021), and false negatives are a well-studied problem for contrastive learning (Arora et al., 2019; Huynh et al., 2022; Chen et al., 2022; Chuang et al., 2020; Robinson et al., 2021). Existing corrections estimate which negatives to distrust from the representation being trained. We study the complementary case in which missingness is *instrumental and recorded*: the annotation protocol determines which coordinates could have been scored, so the relevant object exists before any model does. Examples are stain-dependent lesion scoring in pathology, verified-present and verified-absent rows in Open Images (Kuznetsova et al., 2020), exposure logs in recommendation (Liang et al., 2016), and radiology labels mined from reports, where a blank means "not mentioned" rather than "ruled out" (Wang et al., 2017; Irvin et al., 2019; Oakden-Rayner, 2020).

What the standard analysis would have concluded. The campaign behind this paper began as a method paper. We reconstructed the observation mask of a $47{,}909$-crop pathology corpus, removed the negatives flagged by our mining rule, a knowledge-graph relation rule that consults the mask and that we took to implement certification, and scored each model against the prompts it was trained on. Both mask-dependent removal arms improved mean AUC over the untouched baseline by 0.04 to 0.07 with tight seeds, and the score grew with the fraction of pairs removed (Spearman $+0.90$). The conclusion this invites, that the rule works and more removal works better, is wrong, and four controls that are cheap relative to the campaign show why. Counting fabricated negatives on realized batches shows that the loss receives twice the corpus rate. Replaying the rule on fixed training pairs shows that 72% of what it removes are certifiably correct negatives, so the campaign never tested the removal of fabricated negatives. Removing random pairs at the fractions the loss actually receives, up to two thirds of its candidates, leaves the probes where no removal puts them. A control trained on meaningless text ranks *first* under the metric behind the improvement, while on shared held-out text no removal ranks first.

Contributions.

1. **A pre-training diagnostic** (§3–4): the fabrication rate ρ_{fab}, its two clauses, and a pair-distance form that replaces the one-coordinate coverage approximation, computed on CPU at three denominators. It predicts the rate the loss realizes in training (0.495 predicted, 0.516 realized, 0.260 in the corpus), and on Open Images it exposes a train-tier rate of 0.96.
2. **A sampler result** (§4): group balancing moves the loss-side rate from 0.656 to 0.416 without reaching the corpus rate, 94.9% of the change through the contradicted clause, which is why removal must be measured at the loss.
3. **An evaluation control** (§6): a meaningless-text arm gives every metric an empirical zero. It is at or above every real arm under matched scoring and below all of them on held-out text.
4. **A removal result explained by an audit** (§5, §7.2): at matched loss-side volume, random removal is inert and the knowledge-graph rule costs 0.253 on the group probe, half its anchored range, while removing only the contradicted pairs that partially overlap the target raises all three probes over eight seeds, by 0.021, 0.071, and 0.034.
5. **Ground truth for the undecidable clause** (§7.3): removing it raises macro AUC on every seed of a testbed that hides known pathology labels, most when positives are hidden preferentially; on COCO the gain from not treating unexamined labels as absent grows with the wrong-negative mass predicted before training, as declared, and on LVIS it raises macro AUC by 0.019 on every seed against COCO's complete labels, while the record's own prevalence, 25 times the audited one, costs 0.14 AP as a target. Scored as profiles in the record's closed world, the same testbed models lose on every seed: the cost removal shows on the record belongs to the endpoint.

Our claim is about identification. The observation model, the loss-side rate, the removal fraction at the loss, the evaluation vocabulary and labels, and the anchored range are each needed to interpret a negative-mining result under partial labels, and on our own runs leaving any one of them implicit reversed the conclusion.

2 RELATED WORK

False negatives and partial labels. Contrastive objectives assume trustworthy negatives (Chen et al., 2020; Khosla et al., 2020; Radford et al., 2021; Zhai et al., 2023), and collisions between positives and negatives bound downstream performance (Arora et al., 2019). Corrections cancel, reweight, or filter suspected false negatives using the representation being trained or a pretrained matcher (Huynh et al., 2022; Chuang et al., 2020; Robinson et al., 2021; Chen et al., 2022; Byun et al., 2024; Bulat et al., 2024; Balmaseda et al., 2025), including medical image–text methods (Wang et al., 2022b; 2023). Single-positive and partial-label multi-label learning faces the same missingness and likewise takes its evidence from the model (Cole et al., 2021; Kim et al., 2022; Chen et al., 2023; Song et al., 2024). Our C1 clause is the positive–unlabelled ambiguity (Elkan and Noto, 2008; Bekker and Davis, 2020) restricted to recorded coordinates, and the observation mask is a missing-data indicator (Rubin, 1976; Little and Rubin, 2020) whose dependence on stain is structured missingness (Mitra et al., 2023). Because our diagnostic reads its evidence from the recorded observation process, it exists before training and does not move during it.

Sampling and evaluation validity. Resampling changes a loss's effective distribution (Cui et al., 2019; Kang et al., 2020); we measure how it changes the composition of mined negatives, as exposure models do for recommendation (Liang et al., 2016; Schnabel et al., 2016). The meaningless-text arm follows control tasks and randomized-label diagnostics (Zhang et al., 2017; Hewitt and Liang, 2019), moved into training. Prompt-based scores survive random descriptor words (Roth et al., 2023) and word-order scrambling (Yuksekgonul et al., 2023), and gains that vanish under matched protocols recur across subfields (Melis et al., 2018; Ferrari Dacrema et al., 2019; Musgrave et al., 2020). App. O extends this discussion.

3 FABRICATED NEGATIVES

Let $\mathcal{L}$ be the candidate label set with $|\mathcal{L}| = L$. For sample i we keep $Y_i \subseteq S_i \subseteq \mathcal{L}$: the labels observed to hold and the coordinates that were examined. A coordinate in $S_i \setminus Y_i$ is a verified absence, one outside S_i was never examined, and the closed-world convention silently sets $S_i = \mathcal{L}$. A mining rule produces an ordered multiset $\mathcal{C}$ of candidate negatives. Each candidate has a target i, a source j (a sample, a text, or a label profile), and a nonempty *required negative set* $R_{ij} \subseteq \mathcal{L}$ that must be absent from i for (i, j) to be a valid negative: $R_{ij} = \{\ell\}$ for the rule "j carries ℓ", the source's positive set for the group-based rule of our pipeline, and $R_{ij} = Y_j$ for the observation arms of §5 (App. A; Fig. 1a applies Definition 1 to one target row).

Definition 1 (Fabricated negative)**.** A candidate $(i, j) \in \mathcal{C}$ is assigned to clause C1 when $R_{ij} \not\subseteq S_i$: at least one required coordinate was not examined in the target, so the negative cannot be certified. It is assigned to clause C2 when C1 does not fire and $R_{ij} \cap Y_i \neq \emptyset$: the negative is contradicted by an observed positive. The short-circuit order makes C1 and C2 disjoint; a candidate is *certified* when neither holds. For a finite $\mathcal{C}$, ρ_{C1} and ρ_{C2} are the fractions of candidates in each clause and $\rho_{\mathrm{fab}} = \rho_{\mathrm{C1}} + \rho_{\mathrm{C2}}$.

A C2 pair is *wrong*: the record contradicts it. A C1 pair is *undecidable*: the record cannot certify it, and it is a correct negative if the unexamined coordinate is in fact absent, so ρ_{C1} measures supervision at risk rather than supervision known to be wrong. Rates are computed over ordered candidates at three denominators, all from (Y_i, S_i) and the candidate rule on CPU before training: all candidate pairs of the training split (*corpus*), the candidates reached by a frozen sequence of sampler indices (*schedule*), and the candidates delivered to the loss (*loss*). A removal intervention's *design fraction* is counted over the $\binom{n}{2}$ unordered pool pairs, a pair counting as removed if it is masked in either orientation, and its *fraction at the loss* over the directed candidate slots of realized batches.

Pair-distance form. Let κ be the probability that the target is positive on some required coordinate (the C2 rate when every coordinate is examined), and for the other candidates let $D = |R_{ij}| \geq 1$ and $q(d) = \Pr[D = d,\ \text{no positive overlap}]$.

Lemma 1 (Pair-distance decomposition)**.** *If each required coordinate is examined independently with probability c, then $\rho_{\mathrm{fab}}(c) = \kappa + \sum_{d \geq 1} q(d)\,(1 - c^d) = 1 - \sum_{d \geq 1} q(d)\,c^d$, which is non-increasing in c, strictly decreasing whenever $\kappa < 1$, and tends to κ as $c \to 1$. With coordinate-specific c_ℓ, c^d becomes $\prod_{\ell \in R_{ij}} c_\ell$.*

The proof is a two-case partition (App. B). The scalar approximation $1 - c(1 - \kappa)$ obtained from a coverage number alone is the special case $D = 1$; §4 shows that it misses by 0.37 on Open Images. Independence enters only when a rate is *predicted* from a coverage model; with S_i observed the rate is *measured*, and when examination status is known for only part of the stream the record still brackets it (App. B).

4 THE RATE AT THE LOSS IS NOT THE RATE IN THE CORPUS

Renal pathology. On the $37{,}886$ training crops of our pathology corpus (§7), the corpus, schedule, and loss rates are 0.2602, 0.4954, and 0.4948: the corpus rate, taken over $1{,}260{,}054{,}450$ eligible ordered pairs, understates what the loss receives by $1.9\times$. Sweeping the balancing strength α of the group-balanced sampler over 11 settings and 4 seeds, with no model trained, lowers the loss-side rate monotonically from 0.656 (fully balanced) to 0.416 (balancing off), never reaching the corpus rate (App. C, Tables 7 and 8; Fig. 1b). C2 carries 94.9% of the swing, while ρ_{C1} moves by at most 0.013 anywhere in the sweep: balancing changes how often profiles that share a positive meet, and it cannot examine a coordinate. In training, the observation arm of §5, which removes exactly the fabricated candidates, removes 51.6% of the candidates the loss receives (C1 22.3%, C2 29.4%; four seeds, SD 0.07 points), against 49.5% predicted on CPU.

Open Images: a second annotation process. We test the diagnostic, with no training, on an annotation process we did not design. Open Images (Kuznetsova et al., 2020) records human-verified present *and* absent labels, which define S_i without reading pixels. On $98{,}065$ validation and test images and a 12-label candidate set fixed before any rate was computed, the rate is 0.47 on this native tier and 0.96 on the boxable train tier. The scalar approximation misses the native tier by 0.37, and its miss grows with the candidate-set size, from 0.23 at $L = 4$ to 0.37 at $L = 12$, while the pair-distance form stays within 0.011 (App. D). On CheXpert (Irvin et al., 2019), where a blank means "not mentioned", reading blanks as unexamined makes 73% of the ordered candidate pairs undecidable (C1) and 87% fabricated (C1 or C2; App. L). On VinDr-CXR (Nguyen et al., 2022), a latent-class model fitted to randomly thinned reads recovers the fabricated share within 0.015 down to coverage 0.20, and overstates it by 0.069 under label-dependent hiding (App. L).

5 REMOVAL RULES AND WHAT THEY ACTUALLY REMOVE

Definition 1 gives an operator that removes exactly the fabricated negatives. *Observation certification* sets the supervision mask $M_{ij} = \mathbf{1}\{\text{certified}\}$ with $R_{ij} = Y_j$ (OBS-C1C2), or removes only the C1 candidates (OBS-C1) or only the C2 candidates (OBS-C2). It relabels nothing, imputes nothing, and uses no representation-dependent threshold; the three arms differ from NOELIM, which removes nothing, only in the mask.

The pipeline this paper started from used a different rule, and auditing it is itself a finding. It removes a candidate when the two lesion profiles are *reachable* from one another in a knowledge graph of glomerular lesions, with relations built in an observation-aware profile-overlap space (REACH-33, and REACH-7 at a hash-subsampled fraction). The original configurations also count same-group samples as positives and up-weight knowledge-graph-exclusive negatives by 1.2 (REACHSG-33, and REACHSG-7 at a low fraction). Because the rule consults the mask we had described it as certification. Replaying its loss-side logic on 10^6 fixed ordered training pairs (Table 1) shows that 71.6% of what it removes are *certified* negatives, 21.3% are C2, and 7.1% are C1. It is a rule for removing related, mostly correct negatives, and we name its arms accordingly.

Arms. Besides NOELIM and the three observation arms, we train random removal of 7.26%, 33.28%, 51.65%, and 66.20% of pairs (RAND-7, RAND-33, RAND-52, RAND-66); the knowledge-graph rule (REACH-7, REACH-33); its same-group-positive variants (REACHSG-7, REACHSG-33) and CLOSEDWORLD-42, which is REACHSG-33 with $S_i := \mathcal{L}$ imposed before relations are built; three corrections that keep every pair, namely knowledge-graph-exclusive negatives reweighted by 1.2 (KGX-W1.2), false-negative cancellation (FNC; Huynh et al., 2022), and debiased contrastive learning (DEBIASED; Chuang et al., 2020), each adapted at one operating point; and the control NONSENSE of §6. The random arms and REACH-7 differ from REACH-33 only in which pairs are removed; configurations, codes, and per-seed removal fractions are in App. E.

Table 1: What each rule removes, on a fixed sample of 10^6 ordered, non-diagonal training pairs; each rule's loss-side logic is replayed on the same pairs. Parentheses give shares of the row total.

	C1	C2	certified	total
All sampled pairs	150,838 (15.1%)	168,684 (16.9%)	680,478 (68.0%)	1,000,000
Removed by REACH-33	23,050 (7.1%)	69,122 (21.3%)	232,433 (71.6%)	324,605
share of each clause removed	15.3%	41.0%	34.2%	32.5%
Removed by OBS-C1C2	150,838 (47.2%)	168,684 (52.8%)	0 (0.0%)	319,522

Removal fractions must be counted at the loss. Over uniform pairs the two rules remove almost the same share of the sample (31.95% and 32.46%). The loss receives a different mixture: the group-balanced sampler brings knowledge-graph-related profiles into one batch far more often than uniform sampling, so in training REACH-33 removes 66.2% of the loss-side candidate slots and OBS-C1C2 51.6%, while random removal, being uniform, removes its design fraction. We therefore compare rules at the loss: RAND-52 is matched to OBS-C1C2, and RAND-66 to REACH-33.

6 AN EMPIRICAL ZERO FOR EACH METRIC

A metric that orders arms is not yet a scale: two arms can be reliably ordered and still be separated by less than the metric's resolution, and a metric can rank a meaningless model highly. We therefore train one extra arm, NONSENSE, identical to NOELIM except that every training string is replaced by a meaningless string, and for a metric R and the set A of real-text arms define the anchored range

$$\Delta(R) = \max_{a \in A} R(a) - R(\text{NONSENSE}). \tag{1}$$

$\Delta \leq 0$ means the metric cannot tell semantic training from the control. Each metric can also be run against two vocabularies. In the *matched* setting every arm is scored against the prompts of its own training text, which measures compatibility with an internal codebook; in the *shared held-out* setting all arms are scored against one external paraphrase bank that no arm saw, which measures transfer. We always report the pair (metric, vocabulary).

What the matched setting measures. Two mechanism controls (App. F) show that a matched score reads agreement between the two towers' codebooks rather than content: crossing a clinical vocabulary with a bare label vocabulary between training and evaluation drops it from 0.904 and 0.899 to 0.665 and 0.664, and with an identity vocabulary, permuting the lesion-to-description assignment on one side alone gives 0.471 or 0.442 while permuting both sides together keeps 0.895 (baseline 0.904). The matched score certifies that the image tower agrees with *some* codebook, not what the codebook refers to; only the shared held-out setting tests transfer.

7 EXPERIMENTS

Setup. The corpus is a retrospective, de-identified collection of renal-biopsy whole-slide images in four routine stains, cropped into single glomeruli: a pool of $47{,}909$ crops over 11 lesion coordinates, of which the $37{,}886$ crops of the training split enter alignment. Training and evaluation are slide-disjoint; the clean evaluation split has $1{,}438$ crops from 346 slides, read once by a single annotator. The scoring protocol routes each lesion coordinate to the stains on which it is scored and the annotation queue scored each crop only for the coordinates routed to its stain, so S_i is the routed set of crop i's stain intersected with its queue. Examination is tiered by stain (Table 6), the mask differs from the closed-world default in 22.6% of the training label entries, and it is far from coordinate-independent, so every rate is measured with the mask rather than predicted from a coverage scalar. All arms train the same image and text projection heads on features of the same in-domain image backbone, a ViT-B/16 (Dosovitskiy et al., 2021) pretrained with DINO (Caron et al., 2021) on glomerulus crops, and of the same biomedical text encoder (Gu et al., 2021), with the same sampler, candidate generation, ten epochs, and evaluation (App. G). As a frozen feature extractor the backbone outperforms PLIP (Huang et al., 2023) and CLIP (Radford et al., 2021) (App. N).

Table 2: Five arms on six (metric, vocabulary) pairs; Y1/Y2 on the clean split, probes on the test split. Y1 and probes: four stride-100 seeds, subscripts sample SDs; Y2: the four consecutive seeds (3411–3414) evaluated on the frozen held-out bank, no SD (§7). Y1/Y2 are mean AUC, probes accuracies. NONSENSE is each column's empirical zero; Δ is the anchored range of Eq. 1 and ρ the Spearman correlation between score and removal fraction over the five real arms.

		mean AUC			profile probes		
	Removed	Y1 matched	Y2-A h/o	Y2-B h/o	group	hard	simple
NOELIM	0.00%	$0.8197_{\pm 0.0084}$	0.6544	0.7096	$0.540_{\pm 0.017}$	$0.575_{\pm 0.039}$	$0.566_{\pm 0.016}$
REACHSG-7	7.26%	$0.8293_{\pm 0.0136}$	0.6272	0.6413	$0.514_{\pm 0.022}$	$0.452_{\pm 0.046}$	$0.546_{\pm 0.023}$
REACHSG-33	33.02%	$0.8853_{\pm 0.0082}$	0.6203	0.6482	$0.386_{\pm 0.026}$	$0.269_{\pm 0.073}$	$0.435_{\pm 0.027}$
REACH-33	33.28%	$0.8608_{\pm 0.0198}$	0.6215	0.6737	$0.297_{\pm 0.018}$	$0.116_{\pm 0.012}$	$0.285_{\pm 0.021}$
CLOSEDWORLD-42	42.00%	$0.8995_{\pm 0.0109}$	0.6482	0.6704	$0.465_{\pm 0.005}$	$0.547_{\pm 0.064}$	$0.549_{\pm 0.006}$
NONSENSE	0.00%	$0.9089_{\pm 0.0034}$	0.5806	0.6194	$0.040_{\pm 0.022}$	$0.002_{\pm 0.002}$	$0.045_{\pm 0.027}$
Δ		-0.0094	$+0.0738$	$+0.0902$	$+0.5006$	$+0.5734$	$+0.5209$
Spearman ρ		$+0.90$	-0.30	-0.10	—	—	—

Metrics. *Y1* scores each crop by the dot product between its L2-normalized projected image feature and the embedding of the arm's own training prompt for each stain×lesion row, and averages the per-row AUC against the gold label. *Y2-A* and *Y2-B* apply the same protocol to 24 stain×lesion coordinates with two independently constructed shared held-out paraphrase banks. The three *profile probes* classify each crop's label group by similarity between its projected image feature and frozen profile text embeddings, over all groups (*group* top-1), over predefined hard single-profile groups (*hard*), and over groups of at most one lesion (*simple*); they test whether the embedding keeps the profile distinctions that fabricated negatives teach. Checkpoints are selected by validation group top-1; Y1 and Y2 are read on the clean split and the probes on the test split (App. G).

Seeds. Because our original schedule seeded each epoch with $\mathrm{seed}+\mathrm{epoch}$, runs with consecutive seeds share most of their sampling streams (App. H). Y1 and probe results therefore use four seeds spaced by 100 (3407–3707), eight (3407–4107) for the contradicted-clause and weighting arms; the held-out columns come from the only batch evaluated on the frozen held-out bank, four consecutive seeds (3411–3414), for which we print no SD (App. M). SDs are sample SDs.

7.1 THE EVALUATION VOCABULARY DETERMINES THE ORDERING

The matched metric cannot tell semantics from structure. Under matched Y1, NONSENSE scores 0.9089, above the best real arm (CLOSEDWORLD-42, 0.8995) on all four seeds (per-seed margin $+0.002$ to $+0.017$) and with the smallest seed spread of any arm. Scored against the real arms' vocabulary instead of its own, the same checkpoint falls to 0.596, last of six (App. F). One checkpoint thus spans 0.002 to 0.909 across the six keys of Table 2, and a score reported without its key is not comparable to another.

Under held-out text, no removal ranks first. Under matched Y1 the order is CLOSEDWORLD-42, REACHSG-33, REACH-33, REACHSG-7, NOELIM, and the score rises with removal fraction ($\rho = +0.90$). On both held-out banks NOELIM is first and the correlation disappears (-0.30 and -0.10). The direction holds run by run: NOELIM exceeds REACHSG-33 and REACH-33 on each of the four seeds of both banks (Table 15). Dropping each of the 11 lesions in turn leaves NOELIM first in 21 of the 22 bank–lesion variants (Table 14). A difference of 0.03 is a third of the Y2-B anchored range (0.0902) but only 6% of the group-probe range (0.5006), which is why every gap is read against Δ.

7.2 WHICH PAIRS, NOT HOW MANY

The arms below were trained after the pre-declared held-out budget was spent and are read on the profile probes, where removal is far from inert and the pairs removed decide how.

Equal volume at the loss, different pairs. Random removal leaves all three probes at the no-removal level across the whole range the loss sees, from 7.26% to 66.16% of its candidates (group 0.540 to 0.553 against 0.540; RAND-52−NOELIM $+0.012$, 95% paired t-interval

Table 3: What the removed pairs carry. Profile probes (mean$_{\pm\text{sample SD}}$, four stride-100 seeds) with each arm's removal fraction at the loss. Top: random removal. Middle: the knowledge-graph rule and observation certification. Bottom: interventions that keep every pair. Per-seed values, and OBS-C2 over eight seeds, are in App. I.

Arm	Rule	At loss	group	hard	simple
NOELIM	none	0.00%	$0.540_{\pm 0.017}$	$0.575_{\pm 0.039}$	$0.566_{\pm 0.016}$
RAND-7	random	7.26%	$0.547_{\pm 0.016}$	$0.543_{\pm 0.050}$	$0.571_{\pm 0.019}$
RAND-33	random	33.30%	$0.540_{\pm 0.013}$	$0.595_{\pm 0.040}$	$0.573_{\pm 0.018}$
RAND-52	random	51.67%	$0.553_{\pm 0.006}$	$0.564_{\pm 0.037}$	$0.580_{\pm 0.004}$
RAND-66	random	66.16%	$0.550_{\pm 0.009}$	$0.566_{\pm 0.047}$	$0.579_{\pm 0.012}$
REACH-7	knowledge graph	14.43%	$0.540_{\pm 0.022}$	$0.532_{\pm 0.068}$	$0.558_{\pm 0.023}$
REACH-33	knowledge graph	66.20%	$0.297_{\pm 0.018}$	$0.116_{\pm 0.012}$	$0.285_{\pm 0.021}$
OBS-C1	observation, C1	22.28%	$0.532_{\pm 0.019}$	$0.558_{\pm 0.048}$	$0.553_{\pm 0.023}$
OBS-C2	observation, C2	29.37%	$0.562_{\pm 0.020}$	$0.616_{\pm 0.060}$	$0.598_{\pm 0.031}$
OBS-C1C2	observation, C1∪C2	51.65%	$0.483_{\pm 0.011}$	$0.574_{\pm 0.049}$	$0.510_{\pm 0.021}$
KGX-W1.2	exclusive $\times 1.2$	0.00%	$0.534_{\pm 0.016}$	$0.574_{\pm 0.038}$	$0.566_{\pm 0.016}$
FNC	top-2 visual neighbours	—	$0.527_{\pm 0.019}$	$0.537_{\pm 0.092}$	$0.559_{\pm 0.019}$
DEBIASED	denominator, $\pi = 0.01$	0.00%	$0.533_{\pm 0.017}$	$0.574_{\pm 0.038}$	$0.564_{\pm 0.017}$

$[-0.012, +0.036]$; Table 3). At the same loss-side share, the knowledge-graph rule costs 0.253 on the group probe relative to RAND-66, half of the probe's anchored range, and 0.45 and 0.29 on the hard and simple probes. Removing exactly the fabricated candidates costs 0.070 relative to RAND-52, and 0.057 relative to NOELIM with the same sign on all four seeds ($[-0.094, -0.020]$). Down-sampling the knowledge-graph rule's removed set to 14.4% of the candidates at the loss (REACH-7, identical switches) leaves the group probe at the no-removal level (0.540): within one rule the fraction matters, and across rules at equal volume the pairs do.

Decomposing observation certification. The two clauses behave differently (Table 16). Removing only the undecidable pairs (OBS-C1, 22.3% at the loss) lowers the group and simple probes on every seed, by 0.008 ($[-0.011, -0.006]$) and 0.014 ($[-0.026, -0.001]$), a cost that §7.3 traces to the closed-world endpoint. Removing only the contradicted pairs (OBS-C2, 29.4%) gives positive mean differences over eight seeds ($+0.012$, $+0.055$, $+0.023$; positive on 5, 7, and 6), and every interval contains zero (Table 17). Removing both costs more than the two clauses separately (interaction -0.071 on the group probe, $[-0.092, -0.049]$, negative on every seed), at a volume where random removal is inert. At the loss 85% of the contradicted candidates are partial overlaps; the other 15% share the target's full profile, the same-class negatives that supervised contrastive learning would treat as positives (Khosla et al., 2020). Splitting the clause over eight seeds locates the gain (Table 18). Removing only the full-profile pairs (4.5% at the loss) moves no probe (-0.006, $+0.002$, -0.007). Removing only the partial overlaps (25%) raises all three, by 0.021 ($[+0.005, +0.036]$), 0.071 ($[+0.030, +0.112]$), and 0.034 ($[+0.016, +0.053]$), positive on 7, 8, and 8 seeds, meeting the success criterion declared before the runs, whose hypothesis had placed the gain in the full-profile pairs.

Remark 1 (Prevalence-weighted certification). If each unexamined required coordinate $\ell \in U_{ij} = R_{ij} \setminus S_i$ of a C1 candidate is absent with probability $1-\pi_\ell$ independently of whether it was examined, the pair is a correct negative with probability $w_{ij} = \prod_{\ell \in U_{ij}} (1 - \pi_\ell)$, computable from the record alone. Weighting C1 pairs by w_{ij} instead of dropping them uses no threshold, and the diagnostic bounds its effect: with $\mathbb{E}[w_{ij} \mid \text{C1}] = 0.847$ the weighted arm removes in expectation 15.3% of the C1 mass that OBS-C1 removes, which moves the group probe by about 0.001 if the cost scales with the removed mass, under half the half-width of the interval that resolves OBS-C1.

Corrections that keep every pair. KGX-W1.2, FNC, and DEBIASED stay within seed noise of NOELIM on every probe (largest deviation 0.75 pooled SD). So does WEIGHT-C1, which weights each C1 pair by w_{ij} (Remark 1) and was trained after the remark predicted it invisible: over eight seeds it moves the group probe by -0.0015 ($[-0.0045, +0.0015]$), inside the null band $|\Delta| < 0.005$ pre-declared for it (Table 18). Across record-, graph-, and representation-driven interventions, only the removal of partial overlaps has a paired interval above zero.

Table 4: Ground truth for the undecidable clause. *True / record*: the wrong share of undecidable supervision against hidden labels, and the record's own estimate of it (testbed: C1 pairs vs. Remark 1; COCO at $c = 0.2$: the wrong-negative mass; LVIS: hidden-positive rate vs. examined prevalence). Gains are arm minus the closed world on complete labels, with 95% paired t-intervals over four seeds (five on LVIS); the last column rescores the same contrast on recorded labels in the closed world.

Setting	True / record	Removing or masking	Prevalence as target	On the record
Testbed, hiding ignores the label	0.260 / 0.256	OBS-C1: AUC $+0.009$ $[+0.005, +0.013]$	WEIGHT-C1: AUC $+0.001$ $[-0.004, +0.006]$	top-1 $-0.002 \rightarrow -0.044$
Testbed, positives hidden preferentially	0.415 / 0.067	OBS-C1: AUC $+0.024$ $[+0.022, +0.026]$, top-1 $+0.030$ $[+0.008, +0.053]$	WEIGHT-C1: AUC $+0.0003$ $[+0.0002, +0.0004]$	top-1 $+0.030 \rightarrow -0.064$
COCO, $c = 0.2$	0.1125 / 0.1127	MASK: AUC $+0.012$ $[+0.011, +0.014]$	ET-PRIOR: AUC $+0.013$ $[+0.009, +0.017]$	AUC $+0.012 \rightarrow +0.012$
LVIS record	0.025 / 0.62	MASK: AP $+0.011$ $[+0.005, +0.017]$, AUC $+0.019$ $[+0.018, +0.020]$	ET-PRIOR: AP -0.140 $[-0.180, -0.100]$; audited prior: AP $+0.002$ $[-0.001, +0.005]$	—

7.3 GROUND TRUTH FOR THE UNDECIDABLE CLAUSE

The record cannot say how many undecidable pairs are wrong; three datasets can, and Table 4 collects what they show.

Pathology testbed. Four coordinates are examined in every row (Table 6). We hide about half of their training entries with one fixed draw, independently of the label or preferentially for positives, train NOELIM and OBS-C1 on what remains observed, and score complete test labels (four paired seeds; App. J). OBS-C1 raises macro AUC on every seed and every coordinate under both mechanisms, and under preferential hiding also profile top-1 on every seed. It gains more where it removes less (7.3% of the slots at the loss against 21.3%), where 41% of those pairs are wrong against 26%. The record predicts that share only when hiding ignores the label (Remark 1); when positives are hidden preferentially, weighting C1 pairs by the record's prevalence gains nothing where removal gains most.

Scored against the record, the gain reverses. The pathology probes score recorded profiles in a closed world. Rescored that way, the same checkpoints turn OBS-C1's top-1 gain into a loss on every seed, and predicting the most common profile for every crop scores 0.73 and 0.94, above every trained arm (0.48 on complete labels, below all of them). On the pathology record itself, restricting each crop's candidates to the profiles its own record could have produced ($P \subseteq S_i$) removes the cost of OBS-C1 on an all-profile probe, from -0.013 on every seed to $+0.001$ (App. J).

COCO. We train projection heads on frozen CLIP ViT-B/16 features (Radford et al., 2021) of COCO (Lin et al., 2014) with a sigmoid loss over the 80 category prompts, examine each training image on a random $\mathrm{round}(80c)$ of the categories at five coverages c, and read macro AUC on val2017 once, after all runs, with prompts no arm was trained on (four seeds; App. K). Replacing unexamined labels by their examined prevalence, the expected target of Remark 1, raises macro AUC at every coverage below one, and the gain grows with the predicted mass as declared before training, from 0.005 at $c = 0.75$ to 0.013 at $c = 0.2$ (Spearman 1.0); dropping them from the loss, the analogue of OBS-C1, recovers 71% of the 0.017 the closed world loses at $c = 0.2$.

Natural records. LVIS (Gupta et al., 2019) records, per COCO image, verified-present and verified-absent categories; every other category is unexamined. On its $100{,}170$ shared images and the 63 categories that map one-to-one, the record examines 2.4% of the entries, present categories 39 times more often than absent ones, so the examined prevalence averages 0.62 while a random audit of 1% of the unexamined entries puts theirs at 0.025 (App. L). Training the COCO heads on this record and scoring val2017's complete labels over five seeds, masking unexamined entries beats the closed world on every seed; the examined prevalence as a target costs 0.14 AP, the audited prevalence as a target removes that cost, and the hidden labels themselves, as a reference, would add 0.024 ($[+0.021, +0.027]$), half of which masking alone recovers. Masking also beats the closed

world with a classifier trained from scratch (0.67 against 0.53, two seeds) and in a two-seed Open Images pilot on $2{,}048$ images (0.896 macro AP against 0.750; App. L).

8 DISCUSSION

What removal does here, and why. Three findings replace the account a reader would take from the rule's name. First, the large probe losses come from removing negatives between knowledge-graph-related profiles, 72% of them certifiably correct hard negatives between neighbouring profiles, which the balanced sampler delivers at twice their uniform rate. Second, the fabricated pairs that can be checked are worth removing. Among the contradicted pairs, the partial overlaps, which push an image away from a text naming some but not all of its lesions, carry the gain. The undecidable pairs help wherever their hidden labels are known, and the cost they show on the record is the endpoint's: a closed-world probe rewards the convention NOELIM is trained to reproduce and OBS-C1 is not. Third, the diagnostic locates where the loss is *uncertain*, not where it is wrong: how many uncertifiable pairs are wrong depends on why coordinates went unexamined. When examination ignores the label the record predicts it, on the testbed and at every COCO coverage; when positives are hidden preferentially it predicts 0.067 against 0.415, and on LVIS its 0.62 overstates the audited 0.025, so the reason for missingness belongs in the report, or an audit replaces it.

A reporting protocol. The diagnostic needs only a candidate rule and an account of which coordinates were examined, such as masks, verified rows, exposure logs, or report fields. The evidence supports six reporting items, none of which depends on a removal rule beating a baseline: (1) report (Y_i, S_i) and why coordinates went unexamined, or audit a random sample of them; if S_i cannot be reconstructed, say why and give the bound; (2) report ρ_{C1} and ρ_{C2} on realized batches next to the corpus rate; (3) characterize each mining intervention by its rule and its fraction at the loss, with random removal at the same fraction and the same rule at another; (4) replay each rule on fixed pairs and report what it removes over C1, C2, and certified pairs; (5) anchor each metric with a semantics-free arm and report Δ; (6) name the evaluation vocabulary, use a shared held-out one for any transfer claim, and score each sample only against candidates its own record could have produced. Each item is cheap, and each failure it prevents occurred in our audit.

Limitations. Training interventions under natural missingness are evaluated on the pathology corpus and on LVIS; the testbed and COCO hide labels by rule, and Open Images and CheXpert enter through the diagnostic and a small pilot.

9 CONCLUSION

Partial annotation manufactures a large, measurable amount of negative supervision, and it must be measured where the loss receives it: the loss sees twice the corpus rate, as the CPU diagnostic predicts. The rule our pipeline had called certification removed mostly correct negatives between related profiles and cost half of the probed signal; random removal of up to two thirds of what the loss receives cost nothing. The fabricated negatives that can be checked are worth removing: dropping the contradicted pairs that partially overlap the target raises all three probes, and treating unexamined labels as unknown rather than absent raises macro AUC wherever the hidden labels are known: on the testbed, on COCO in step with the wrong-negative mass the diagnostic predicts before training, and on the federated record of LVIS. A mining intervention is therefore identified by its rule, its fraction at the loss, what it removes, and why coordinates went unexamined, which the record alone does not reveal. Evaluation needs an empirical zero, a named vocabulary, and labels outside the training's closed world: a meaningless-text control matches or exceeds every real model under matched-vocabulary scoring, and models that gain on complete labels lose once their profiles are scored in the record's closed world.

ETHICS STATEMENT

The pathology study uses de-identified archival material under institutional ethics approval with a waiver of consent; institution and cohort identifiers are withheld during review, and slides cannot be redistributed. All other data are public. No result here is a clinical performance claim.

REPRODUCIBILITY STATEMENT

Every number in the paper is generated from frozen manifests and run artifacts by a verification script (App. M). Definitions and proofs are in App. A–B, configurations and per-seed values of every arm in App. E–I, and the testbed, COCO, and natural-record experiments in App. J–L. The public-data experiments on Open Images, COCO, LVIS, CheXpert, and VinDr-CXR, described in App. D, K, and L, do not require the pathology corpus.

AI USE STATEMENT

In this work, AI tools were used to aid and polish the writing of this paper. Specifically, ChatGPT was used for grammar checking, language polishing, and suggesting improvements to paragraph structure. All AI-assisted outputs were reviewed, verified, and edited by the authors, who take full responsibility for the content of this work. AI tools were not used to fabricate research data, experimental results, or citations.

REFERENCES

Anish Acharya, Sujay Sanghavi, Li Jing, Bhargav Bhushanam, Dhruv Choudhary, Michael Rabbat, and Inderjit S. Dhillon. Understanding contrastive representation learning from positive unlabeled (PU) data. *arXiv preprint arXiv:2402.06038*, 2024.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5628–5637, 2019.

Vicente Balmaseda, Bokun Wang, Ching-Long Lin, and Tianbao Yang. Discovering global false negatives on the fly for self-supervised contrastive learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label classification with partial annotations using class-aware selective loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4764–4772, 2022.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing. In *Computer Vision – ECCV 2022*, pages 1–21. Springer, 2022.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4843–4855, 2021.

Serhat Selcuk Bucak, Rong Jin, and Anil K. Jain. Multi-label learning with incomplete class assignments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2801–2808, 2011.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. FFF: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14172–14182, 2024.

Jonathon Byrd and Zachary C. Lipton. What is the effect of importance weighting in deep learning? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 872–881, 2019.

Jaeseok Byun, Dohoon Kim, and Taesup Moon. MAFA: Managing false negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27314–27324, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Cheng Chen, Yifan Zhao, and Jia Li. Semantic contrastive bootstrapping for single-positive multi-label recognition. *International Journal of Computer Vision*, 131(12):3289–3306, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive representation learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 933–942, 2021.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9268–9277, 2019.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619, 2021.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 703–711, 2014.

Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 647–657, 2019.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9588–9597, 2021.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 213–220, 2008.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys)*, pages 101–109, 2019.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021.

Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5356–5364, 2019.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 107–112, 2018.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2733–2743, 2019.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, pages 263–272, 2008.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual-language foundation model for pathology image analysis using medical Twitter. *Nature Medicine*, 29:2307–2316, 2023.

Dat Huynh and Ehsan Elhamifar. Interactive multi-label CNN learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9423–9432, 2020.

Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2785–2795, 2022.

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*, pages 590–597, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 4904–4916, 2021.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Myunsoo Kim, Seong-Woong Shim, and Byung-Jun Lee. FALCON: False-negative aware learning of contrastive negatives in vision-language pretraining. *arXiv preprint arXiv:2505.11192*, 2025.

Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14156–14165, 2022.

Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1675–1685, 2017.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096, 2019.

Dawen Liang, Laurent Charlin, James McInerney, and David M. Blei. Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 951–961, 2016.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *ACM Queue*, 17(1):45–77, 2019.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 3rd edition, 2020.

Biao Liu, Jie Wang, Ning Xu, and Xin Geng. Can class-priors help single-positive multi-label learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W. Tsang. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7955–7974, 2022.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

Benjamin M. Marlin and Richard S. Zemel. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys)*, pages 5–12, 2009.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations (ICLR)*, 2018.

Aditya Krishna Menon, Brendan van Rooyen, Cheng Soon Ong, and Robert C. Williamson. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 125–134, 2015.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3111–3119, 2013.

Robin Mitra, Sarah F. McGough, Tapabrata Chakraborti, Chris Holmes, Ryan Copping, Niels Hagenbuch, Stefanie Biedermann, Jack Noonan, Brieuc Lehmann, Aditi Shenvi, Xuan Vinh Doan, David Leslie, Ginestra Bianconi, Ruben Sanchez-Garcia, Alisha Davies, Maxine Mackintosh, Eleni-Rosalina Andrinopoulou, Anahid Basiri, Chris Harbron, and Ben D. MacArthur. Learning from data with structured missingness. *Nature Machine Intelligence*, 5(1):13–23, 2023.

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision – ECCV 2020*, pages 681–699. Springer, 2020.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1196–1204, 2013.

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

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021a.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021b.

Luke Oakden-Rayner. Exploring large-scale public medical image datasets. *Academic Radiology*, 27(1):106–112, 2020.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4609–4622, 2020.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 5835–5847, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5389–5400, 2019.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, 2017.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 452–461, 2009.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91, 2021.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*, 2021.

Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15746–15757, 2023.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1670–1679, 2016.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2018.

Gonglei Shi, Li Xiao, Yang Chen, and S. Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70:101979, 2021.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 761–769, 2016.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2888–2913, 2021.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

GyeongRyeol Song, Noo-Ri Kim, Jin-Seop Lee, and Jee-Hyong Lee. IGNORE: Information gap-based false negative loss rejection for single positive multi-label learning. In *Computer Vision – ECCV 2024*, pages 472–488. Springer, 2024.

Harald Steck. Training and testing of recommender systems on data missing not at random. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 713–722, 2010.

Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 593–598, 2010.

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*, 6:1399–1409, 2022.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528, 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, 2020.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, 2021.

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. PiCO: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations (ICLR)*, 2022a.

Peiqi Wang, Yingcheng Liu, Ching-Yun Ko, William M. Wells, Seth Berkowitz, Steven Horng, and Polina Golland. Sample-specific debiasing for better image-text models. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, PMLR, 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 9929–9939, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.

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 (EMNLP)*, pages 3876–3887, 2022b.

Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2840–2848, 2017.

Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4302–4309, 2018.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations (ICLR)*, 2021.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *Computer Vision – ECCV 2022*, pages 668–684. Springer, 2022.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S. Dhillon. Large-scale multi-label learning with missing labels. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 593–601, 2014.

Mert Yuksekgonul, Federico Bianchi, Pradyumna Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2023.

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 (ICCV)*, pages 11975–11986, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023.

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 *Proceedings of the 7th Machine Learning for Healthcare Conference (MLHC)*, pages 2–25, 2022.

Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10042–10051, 2021.

Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Acknowledging the unknown for multi-label learning with single positive labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

SUPPLEMENTARY MATERIAL

The appendices that follow are the supplementary text of this submission. Appendices A–B give the observation model, an illustration of Definition 1 and of the sampler sweep (Figure 1), the rate ledger, and the proof of Lemma 1; Appendices C–D the full sampler sweep and the Open Images tables; Appendices E–G the arm configurations, rule replays, vocabulary controls, and training configuration; Appendices H–I the seed audit, held-out robustness, and per-seed values behind every paired interval in the main text; Appendices J–K the ground-truth testbed, its rescorings, and the COCO experiment of §7.3; Appendix L the natural records of LVIS, Open Images, CheXpert, and VinDr-CXR; and Appendices M–O the provenance ledger, the frozen-feature probe, and extended related work.

A OBSERVATION MODEL, RATES, AND REMOVAL FRACTIONS

Required negative sets. For the rule "j carries ℓ", $R_{ij} = \{\ell\}$. For the group-based rule of the pipeline and for the observation arms, R_{ij} is the source's observed positive set Y_j. The knowledge-graph rule removes a candidate when the source's and the target's lesion profiles are reachable from one another in a knowledge graph of glomerular lesions; REACH-33 and REACHSG-33 build the relations in the observation-aware profile-overlap space, and REACHSG-7 in the group-level relation space of the original pipeline. App. D describes the Open Images candidate set.

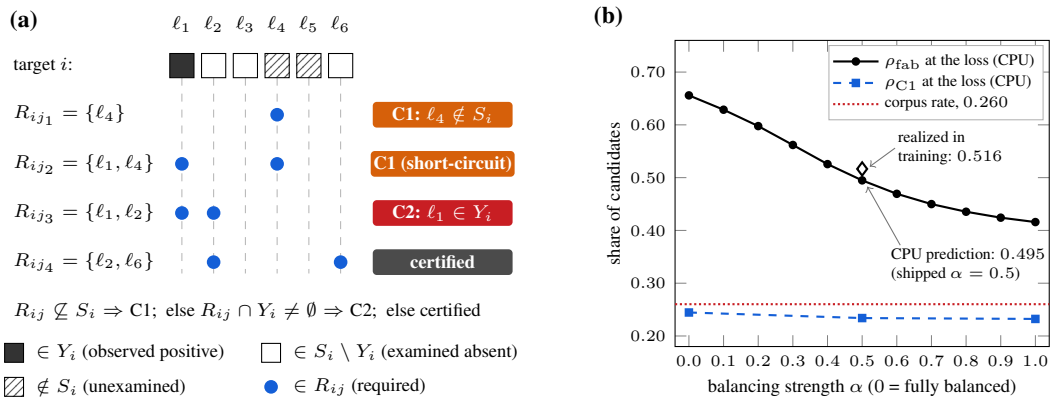


Figure 1: Definition 1 on one target row (a) and the fabrication rate at the loss across the sampler sweep (b); the values in (b) are those of Tables 5, 7, and 8. **(a)** Each of the $L = 6$ coordinates of the target is observed positive (filled, $\in Y_i$), examined absent (open, $\in S_i \setminus Y_i$), or unexamined (hatched, $\notin S_i$); four candidates are given by their required sets R_{ij} (dots). A required coordinate outside S_i puts a candidate in C1 even when another required coordinate is an observed positive (R_{ij_2}, the short-circuit that keeps the clauses disjoint); otherwise an observed positive among the required coordinates gives C2 (R_{ij_3}); otherwise the negative is certified (R_{ij_4}). **(b)** ρ_{fab} of the candidates delivered to the loss, computed on CPU at the 11 balancing strengths of Table 8 (four seeds each; the largest seed SD, 0.0013, is smaller than the marks), and ρ_{C1} at the three settings where it is tabulated (Tables 5 and 7). The swing is carried by C2 while ρ_{C1} stays nearly flat, and no setting reaches the corpus rate. At the shipped $\alpha = 0.5$ the CPU prediction, 0.495, is close to the rate realized in training by the removal counters of OBS-C1C2, 0.516 (Table 5).

Removal fractions. Hash-based removal indexes each unordered pair by $(\min(i,j), \max(i,j))$ before hashing, so both orientations share one decision. The design fraction counts, over the $\binom{47{,}909}{2} = 1{,}147{,}612{,}186$ unordered pool pairs, those masked in either orientation. The fraction at the loss is logged in every training batch as the number of removed directed candidate slots (knowledge-graph, observation, and hash removals together) divided by the number of positive pairs times 63, less the slots the loss skips, and is accumulated over ten epochs and averaged over four seeds. With one positive per anchor and batches of 64 this is the fraction of directed off-diagonal slots removed. Arms that also count same-group samples as positives have about four positive pairs

per anchor, which inflates the logged denominator about fourfold; for them we recount removed slots directly over the 23,869,440 directed off-diagonal slots of the realized batches (592 batches of 64 per epoch, ten epochs).

Rate ledger. Rates over different denominators count different objects, so Table 5 records the population behind each. The corpus row counts only the pairs that the candidate rule proposes, 175,256,660 (12.2%) fewer than the training-row pairs, and the pairs it leaves out are not a random subset. None of them is C1: the 216,285,283 C1 pairs of the first row, divided by the corpus count, give its $\rho_{\mathrm{C1}} = 0.1716$. About 130.5 million are C2, as many as the pairs whose source shares the target's nonempty profile ($Y_j = Y_i \neq \emptyset$; 9.1% of training-row pairs by the uniform sample, App. C), every one of which is C2 because $R_{ij} = Y_j = Y_i \subseteq S_i$; the remaining 44.8 million are certified. Hence ρ_{C2} falls from 0.169 over uniform pairs to 0.089 in the corpus, while ρ_{C1} rises from 0.151 to 0.172 only because the denominator shrinks.

Table 5: Rate ledger for the pathology training split. The first row uses all ordered pairs of training rows with R_{ij} the source's positive set (the population of Remark 1); the corpus row uses the eligible ordered candidate pairs; the schedule row the candidates reached by the frozen sampler sequence; the CPU loss row the candidates delivered to the loss at the shipped setting $\alpha = 0.5$; the realized row the removal counters of OBS-C1C2 during training (four seeds, ten epochs). [a]Estimated on the uniform sample of 10^6 training-row pairs of Table 1, whose C1 share (15.08%) matches the full count. [b]$\rho_{\mathrm{fab}} - \rho_{\mathrm{C1}}$ (Definition 1).

Denominator	Candidates	ρ_{C1}	ρ_{C2}	ρ_{fab}
Training-row pairs	1,435,311,110	0.1507	0.1687[a]	0.3195[a]
Corpus	1,260,054,450	0.1716	0.0886	0.2602
Schedule	22,732,800	0.191	0.304[b]	0.4954
Loss, predicted on CPU	—	0.234	0.261	0.4948
Loss, realized in training	—	0.2228	0.2937	0.5165

Table 6: Observation and prevalence per lesion coordinate on the 37,886 training rows that enter the alignment stage. "Examined" is the number of rows whose observation mask covers the coordinate; π_ℓ is the positive fraction among examined rows. Examination is tiered by stain: four coordinates are examined in every row, four in 29,533 (78.0%), one in 22,965 (60.6%), and two in 14,921 (39.4%). With R_{ij} the source's positive set, 216,285,283 of 1,435,311,110 ordered pairs are C1 ($\rho_{\mathrm{C1}} = 0.1507$), $\mathbb{E}[w_{ij} \mid \mathrm{C1}] = 0.8467$ and $\mathbb{E}[1 - w_{ij} \mid \mathrm{C1}] = 0.1533$.

Coordinate	Examined	Examined (%)	Positive	π_ℓ
mesangial lesion	37,886	100.0	13,033	0.344
mesangial, PAS-positive	14,921	39.4	2,439	0.163
mesangial, PAS-negative	29,533	78.0	1,750	0.059
ischemic wrinkling	14,921	39.4	1,771	0.119
segmental sclerosis	37,886	100.0	1,128	0.030
loop necrosis	22,965	60.6	2,358	0.103
membranous nephropathy	29,533	78.0	6,803	0.230
membranoproliferative	29,533	78.0	1,880	0.064
endocapillary	29,533	78.0	1,738	0.059
crescent	37,886	100.0	1,406	0.037
global sclerosis	37,886	100.0	4,951	0.131

B PROOF OF LEMMA 1

Partition the candidates by whether the target is positive on some required coordinate. (i) If it is, the candidate is fabricated whatever the mask: if that coordinate is unexamined the candidate is C1, and if it is examined it is recorded in Y_i, so the candidate is C2 when every required coordinate is examined and C1 otherwise. This case has probability κ, and at $c = 1$ it is exactly C2. (ii) Otherwise, with $D = d$ required coordinates, the candidate is certified exactly when all d are examined, which under independent examination has probability c^d, and it is C1 otherwise. Summing, $\rho_{\mathrm{fab}}(c) =$

$\kappa + \sum_d q(d)(1 - c^d)$, and since $\kappa + \sum_d q(d) = 1$, $\rho_{\mathrm{fab}}(c) = 1 - \sum_d q(d)c^d$. The derivative $-\sum_d d\, q(d)\, c^{d-1}$ is non-positive, and negative for $c > 0$ whenever some $q(d) > 0$, that is, whenever $\kappa < 1$; at $c = 1$, $\rho_{\mathrm{fab}} = 1 - \sum_d q(d) = \kappa$. With coordinate-specific probabilities the certification probability in case (ii) is $\prod_{\ell \in R_{ij}} c_\ell$, and the scalar approximation is the special case in which every case-(ii) candidate has $D = 1$.

Partial records. When examination status is known for only part of the stream, each candidate is fabricated on the record alone (an observed positive or a recorded unexamined coordinate among its required coordinates), a fraction ρ_{rec} of all candidates; certified on the record (every required coordinate recorded as examined and none observed positive); or undecided, a fraction u. Only the undecided candidates are unknown, and each is at most fabricated; hence $\rho_{\mathrm{rec}} \le \rho_{\mathrm{fab}} \le \rho_{\mathrm{rec}} + u$, a gate rather than a point estimate.

C SAMPLER SWEEP

Table 7: Clause decomposition at the endpoints of the balancing sweep (4 seeds per point; the largest seed SD of $\rho_{\mathrm{fab}}/0.2602$ anywhere in the sweep is 0.0051). The shipped setting $\alpha = 0.5$ gives the loss-side rate 0.4948.

	$\alpha = 0$ (balanced)	$\alpha = 1$ (off)	relative change
ρ_{C1} (undecidable)	0.2445	0.2323	5.0%
ρ_{C2} (contradicted)	0.4113	0.1835	55.4%
ρ_{fab}	0.6559	0.4158	—

Table 8: Complete balancing sweep (four stride-100 seeds per setting; no model is trained). The curve is monotone across all 11 settings and never reaches the corpus rate 0.2602: even with balancing disabled the loss sees $1.60\times$ the corpus rate. The shipped setting is $\alpha = 0.5$. A second sampler family gives 0.2602 ($1.00\times$) with sample-uniform batches (see below), so the amplification is a property of group balancing rather than of batching as such.

α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ρ_{fab} at the loss	0.656	0.629	0.598	0.562	0.526	0.495	0.469	0.450	0.435	0.424	0.416
$\times$ corpus rate	2.520	2.416	2.297	2.159	2.020	1.902	1.804	1.729	1.673	1.630	1.598
seed SD ($\times$ rate)	0.004	0.002	0.003	0.005	0.005	0.003	0.001	0.005	0.004	0.003	0.003

A second sampler family confirms that the amplification is a property of group balancing: sample-uniform batches return the loss-side rate to 0.2602, the corpus rate, while both group samplers amplify it, and across families the span of C2 is 4.4 times that of C1. Reconstructing every batch of the four stride-100 seeds over ten epochs gives $28{,}038{,}284$ loss-side C2 candidates, of which $4{,}317{,}168$ (15.40%) share the target's full profile and $23{,}721{,}116$ (84.60%) overlap partially; on a uniform sample of 10^6 pairs the shares are 53.97% and 46.03%. Because NOELIM treats only the diagonal as positive, same-group pairs enter its negatives, and OBS-C2 removes the part of them that is contradicted.

D OPEN IMAGES

The diagnostic uses $98{,}065$ validation and test images ($23{,}024 + 75{,}041$) and a 12-label candidate set fixed as the frequency-ordered prefix of the corpus before any rate was computed. Coverage is $c = \sum_i(|S_i| - |Y_i|)/\sum_i(L - |Y_i|)$. Table 9 gives the estimators on both tiers and Table 10 the candidate-set-size sweep.

Table 9: Open Images diagnostic under the fixed 12-label candidate set. "Own histogram" evaluates Lemma 1 with the tier's own pair-distance and profile distribution; "cross-tier" carries the validation/test distribution to the train tier and predicts from coverage alone. The fully observed row is not zero because C2 survives complete examination.

Tier	Estimator	c	ρ_{fab}	residual
Val+test	measured	0.189856	0.474243	—
Val+test	scalar $D = 1$	0.189856	0.846882	-0.372639
Val+test	own histogram	0.189856	0.484726	-0.010482
Train	measured	0.071729	0.960094	—
Train	own histogram	0.071729	0.961523	-0.001429
Train	cross-tier, coverage only	0.071729	0.511088	$+0.449006$
Val+test	fully observed limit	1.000000	0.193503	—

Table 10: Candidate-set-size sweep on the Open Images validation/test tier. The five label sets are nested frequency-ordered prefixes of one ordering fixed before any rate was computed. Residuals are measured minus predicted. As L grows the required negative set lengthens, which is exactly the information the one-coordinate approximation discards: its residual grows from 0.23 to 0.37, while the pair-distance form stays within 0.011. The two clauses move in opposite directions, so the total follows neither.

L	c	measured	own hist.	resid.	scalar D=1	resid.	$(\rho_{\mathrm{C1}}, \rho_{\mathrm{C2}})$
4	0.498803	0.360900	0.362036	-0.001	0.586204	-0.225	(0.266, 0.095)
6	0.325505	0.408021	0.409307	-0.001	0.723051	-0.315	(0.354, 0.054)
8	0.256209	0.467142	0.472280	-0.005	0.792272	-0.325	(0.414, 0.053)
10	0.214380	0.474732	0.483294	-0.009	0.827468	-0.353	(0.427, 0.047)
12	0.189856	0.474243	0.484726	-0.010	0.846882	-0.373	(0.431, 0.044)

E ARMS, RULE REPLAYS, AND REMOVAL FRACTIONS

Configuration diffs. The observation arms have the effective configuration of NOELIM: no same-group positives, no reachable or exclusive removal, and an exclusive weight that is present in the file but disabled. The random arms differ from REACH-33 only in turning reachable removal off and applying an all-pairs hash, and REACH-7 only in applying a hash to the reachable pairs (83,268,854 of 381,971,720). REACHSG-7 is the configuration of the original pipeline: same-group positives, reachable removal in the group-level relation space, and exclusive weighting.

Replays and nesting. The replay of run configurations also caught a second mislabel: an arm our ledger listed as random removal was the original pipeline at its low fraction. It appears here as REACHSG-7, and clean random arms at four fractions replace it. On the 10^6-pair sample of Table 1, REACHSG-33 removes 322,022 pairs: those REACH-33 removes, except 2,583 reachable C2 pairs that it keeps. The pipeline also removes candidate pairs whose records overlap. Ablating the observation model (CLOSEDWORLD-42) does not merely add negatives: relative to REACHSG-33 it restores no removed pair and removes 103,086,449 *additional* pairs, so REACHSG-33 is strictly contained in CLOSEDWORLD-42 (378,888,235 + 103,086,449 = 481,974,684). Declaring every coordinate examined creates artificial agreement on coordinates neither sample was assessed for, and that agreement propagates into overlap removal: applying the observation model only at negative mining fixes one error while leaving its mirror image in place.

Table 11: Arms. Codes in parentheses are the identifiers of the run ledger. All arms share architecture, optimization, data, candidate generation, sampler, and evaluation. *Design*: fraction of the unordered universe of 1,147,612,186 pool pairs, or [a]of the 10^6-pair audit sample (Table 1). *At loss*: fraction of directed candidate slots removed in training, mean of four seeds (Table 12). [b]Recounted over the directed slots of the realized batches, because the logged counter divides by positive pairs, which same-group positives multiply (App. A). [c]Multiplies the negative weight of knowledge-graph-exclusive pairs by 1.2 and removes nothing. [d]Cancels the two most visually similar in-batch candidates of each anchor. [e]Corrects the denominator with positive prior 0.01 and removes nothing. [f]Recounted from the reconstructed batches of App. C, where they are 15.40% and 84.60% of the loss-side C2 candidates; the two parts sum to OBS-C2.

Arm	What changes	Design	At loss
NOELIM (A1)	nothing	0.00%	0.00%
RAND-7	random pairs	7.26%	7.26%
RAND-33	random pairs	33.28%	33.30%
RAND-52	random pairs	51.65%	51.67%
RAND-66	random pairs	66.20%	66.16%
REACH-7 (G2-low)	hash subsample of the reachable pairs	7.26%	14.43%
REACH-33 (G2)	knowledge-graph-reachable pairs	33.28%	66.20%
REACHSG-7 (A4)	as REACHSG-33, group-level relations	7.26%	12.67%[b]
REACHSG-33 (B2)	reachable; same-group positives; exclusive $\times 1.2$	33.02%	63.69%[b]
CLOSEDWORLD-42 (TM-off)	REACHSG-33 with $S_i := \mathcal{L}$ before relations	42.00%	75.37%[b]
NONSENSE (R1b)	meaningless training text; mask unchanged	0.00%	0.00%
OBS-C1C2	C1 and C2 candidates (Def. 1)	31.95%[a]	51.65%
OBS-C1	C1 candidates only	15.08%[a]	22.28%
OBS-C2	C2 candidates only	16.87%[a]	29.37%
full profile	C2 candidates with $Y_j = Y_i$	—	4.52%[f]
full profile, text→image	the same, one loss direction only	—	—
partial overlap	the other C2 candidates	—	24.85%[f]
KGX-W1.2	exclusive negatives reweighted[c]	0.00%	0.00%
FNC	false-negative cancellation (Huynh et al., 2022)[d]	—	—
DEBIASED	debiased contrastive loss (Chuang et al., 2020)[e]	0.00%	0.00%
WEIGHT-C1	C1 candidates weighted by w_{ij} (Remark 1)	0.00%	0.00%
constant weight	C1 candidates weighted by 0.847	0.00%	0.00%

Table 12: Removal fractions at the loss (percent of directed candidate slots), by seed. Arms marked * count same-group samples as positives and are recounted over the 23,869,440 directed slots of their realized batches (App. A).

Arm	3407	3507	3607	3707	mean$_{\pm SD}$
RAND-7	7.2566	7.2608	7.2686	7.2596	$7.2614_{\pm 0.0051}$
RAND-33	33.3128	33.3161	33.2938	33.2968	$33.3049_{\pm 0.0112}$
RAND-52	51.6622	51.6801	51.6589	51.6593	$51.6651_{\pm 0.0101}$
RAND-66	66.2117	66.1421	66.1436	66.1613	$66.1647_{\pm 0.0325}$
REACH-7	14.4506	14.4241	14.4410	14.4024	$14.4295_{\pm 0.0212}$
REACH-33	66.3717	66.1495	66.2802	66.0164	$66.2044_{\pm 0.1550}$
REACHSG-7*	12.6373	12.6932	12.7071	12.6266	$12.6660_{\pm 0.0400}$
REACHSG-33*	63.8500	63.6392	63.7524	63.5074	$63.6872_{\pm 0.1477}$
CLOSEDWORLD-42*	75.4998	75.4006	75.4122	75.1834	$75.3740_{\pm 0.1346}$
OBS-C1	22.2731	22.3497	22.2534	22.2488	$22.2813_{\pm 0.0468}$
OBS-C2	29.4337	29.3485	29.3817	29.3012	$29.3663_{\pm 0.0558}$
OBS-C1C2	51.7068	51.6982	51.6351	51.5501	$51.6476_{\pm 0.0724}$

F VOCABULARY CONTROLS

A same-checkpoint control makes the matched-setting mechanism explicit: scoring the NONSENSE checkpoint against the real arms' vocabulary instead of its own moves it from 0.9089 to 0.5964 and

Table 13: Mechanism controls for the matched setting, on the clean evaluation split. *Left*: crossing a richly worded clinical vocabulary with a bare label vocabulary on the training and evaluation sides; the main effect of which vocabulary is used is 0.002–0.003, the matched-versus-crossed contrast is 0.237. *Right*: identity-vocabulary controls, three seeds. A score of 0.5 is chance. Permuting the lesion-to-description assignment on either side alone destroys the score; permuting both sides together restores it, so what the matched setting reads is the assignment shared by the two towers, not the wording.

Vocabulary factorial			Key permutation (identity vocabulary)	
	rich eval	bare eval	condition	matched AUC
rich train	0.9041	0.6645	baseline	0.9041
bare train	0.6650	0.8994	eval word order scrambled	0.8647
			both keys permuted	0.8951
			eval key permuted	0.4714
			train key permuted	0.4418
			one tower rotated	0.5012

from rank 1 to rank 6 of 6, while every real arm is unchanged by construction, and Δ becomes $+0.3030$.

G TRAINING CONFIGURATION

All arms use batches of 16 label groups with 4 crops each ($B = 64$), drawn by the group-balanced sampler at $\alpha = 0.5$; ten epochs; learning rate 10^{-3} and weight decay 10^{-4}; temperature 0.07; image and text projection heads with hidden width 1024 and output dimension 512; and the BiomedBERT text encoder (Gu et al., 2021) with inputs truncated to 64 tokens. The checkpoint of each run is the epoch with the best validation group top-1. Y1 and Y2 are read on the clean split. The profile probes are read on the test split against the 58 label-group profiles frozen at training time, over the 5,206 test crops whose profile is among them (*simple*: the 4,276 of these whose profile has at most one lesion). FNC cancels the two most visually similar in-batch candidates of each anchor; DEBIASED applies the correction of Chuang et al. (2020) with positive prior 0.01; KGX-W1.2 multiplies the weight of knowledge-graph-exclusive negatives by 1.2, which touches 344,032 pairs in the first epoch, 0.56% of them C2.

H SEEDS AND HELD-OUT ROBUSTNESS

With per-epoch seeds seed+epoch, a run with seed s and E epochs uses the streams $s, \ldots, s+E-1$, so runs with consecutive seeds share $E - 1$ of their E streams; eight nominal runs realized 17 distinct schedules instead of 80. Seeds spaced by 100 share none. The held-out batch (3411–3414) predates the audit, and its runs share streams pairwise, so we print no SD for it; an older consecutive-seed batch is retired (App. M). Each held-out bank is scored under an ambiguity-preserving policy (shown) and an ambiguity-dropping policy, which gives the same ordering. Table 14 audits the held-out columns coordinate by coordinate, and Table 15 seed by seed: NOELIM exceeds REACHSG-33 and REACH-33 on all four seeds of both banks, although the four runs share streams and are not independent replicates. NOELIM's lead over the two knowledge-graph arms at a third of the pairs covers 16 to 19 of the 24 coordinates; its lead over REACHSG-7 and CLOSEDWORLD-42 comes mostly from global sclerosis, which the three arms with same-group positives invert. The held-out readout is sensitive to the arm: NONSENSE's per-coordinate profile is uncorrelated with that of every real arm (r between -0.19 and $+0.16$, against at least 0.84 among the real arms), so the control fails on different coordinates rather than uniformly.

Table 14: Per-coordinate audit of the held-out columns of Table 2: 24 stain×lesion coordinates, seeds 3411–3414, ambiguity-preserving policy; a re-analysis of stored scores with no new held-out read. < 0.5: coordinate–seed readings below chance. Worse: coordinates (of 24) whose seed-averaged AUC is below NOELIM's. Gap: difference to NOELIM in mean AUC. GS: share of the gap contributed by the three global-sclerosis coordinates. Global sclerosis, lesion-mean AUC (Y2-A/Y2-B): NOELIM 0.50/0.58, REACH-33 0.42/0.50, REACHSG-33 0.33/0.32, CLOSEDWORLD-42 0.33/0.31, REACHSG-7 0.29/0.16, NONSENSE 0.70/0.82. Leave-one-lesion-out over the 11 lesions: NOELIM is first in 10/11 (Y2-A; the exception drops global sclerosis and puts CLOSEDWORLD-42 first) and 11/11 (Y2-B) variants; Spearman ρ between score and removal fraction stays in $[-0.40, 0.00]$ (Y2-A) and $[-0.30, 0.00]$ (Y2-B); NONSENSE is below every real arm in 11/11 (Y2-A) and 7/11 (Y2-B) variants, the four exceptions dropping lesions on which it is weakest; Δ stays between 0.041 and 0.136.

	Y2-A				Y2-B			
Arm	< 0.5	worse	gap	GS	< 0.5	worse	gap	GS
NOELIM	12/96	—	—	—	13/96	—	—	—
REACHSG-7	20/96	12	−0.027	98%	20/96	15	−0.068	78%
REACHSG-33	20/96	18	−0.034	62%	19/96	19	−0.061	53%
REACH-33	20/96	16	−0.033	31%	17/96	16	−0.036	29%
CLOSEDWORLD-42	20/96	13	−0.006	>100%	19/96	18	−0.039	86%
NONSENSE	44/96	13	−0.074	—	38/96	14	−0.090	—

Table 15: Per-seed held-out differences under the readout of Table 2: for each seed, the mean AUC over the 24 coordinates; the four-seed means equal the main-table differences. Positive values favour NoElim.

Bank	Contrast	3411	3412	3413	3414	mean
Y2-A	NoElim−ReachSG-33	+0.022	+0.032	+0.029	+0.055	+0.034
Y2-A	NoElim−Reach-33	+0.047	+0.026	+0.022	+0.037	+0.033
Y2-B	NoElim−ReachSG-33	+0.048	+0.055	+0.049	+0.094	+0.061
Y2-B	NoElim−Reach-33	+0.066	+0.036	+0.029	+0.013	+0.036

I Per-seed values

Table 16: Observation arms minus NoElim at the same seed, with the mean and the 95% paired t-interval ($df = 3$). The interaction is Obs-C1C2−Obs-C1−Obs-C2+NoElim. For the matched-volume random control, Rand-52−NoElim is $+0.012$ $[-0.012, +0.036]$ (group), -0.011 $[-0.052, +0.030]$ (hard), and $+0.014$ $[-0.017, +0.045]$ (simple). Obs-C2 over eight seeds is in Table 17, and its two parts in Table 18.

Arm	Probe	3407	3507	3607	3707	Mean	95% interval
Obs-C1C2	group	−0.050	−0.085	−0.030	−0.064	−0.057	[−0.094, −0.020]
	hard	−0.023	−0.071	−0.013	+0.102	−0.001	[−0.118, +0.116]
	simple	−0.040	−0.096	−0.030	−0.059	−0.056	[−0.103, −0.009]
Obs-C1	group	−0.009	−0.008	−0.011	−0.006	−0.008	[−0.011, −0.006]
	hard	−0.047	+0.002	−0.021	−0.005	−0.018	[−0.052, +0.017]
	simple	−0.012	−0.008	−0.025	−0.009	−0.014	[−0.026, −0.001]
Obs-C2	group	+0.044	−0.016	+0.060	−0.000	+0.022	[−0.035, +0.079]
	hard	−0.023	+0.036	+0.003	+0.144	+0.040	[−0.077, +0.157]
	simple	+0.068	−0.017	+0.072	+0.004	+0.032	[−0.040, +0.104]
Interaction	group	−0.085	−0.061	−0.079	−0.058	−0.071	[−0.092, −0.049]
	hard	+0.047	−0.109	+0.005	−0.037	−0.024	[−0.129, +0.082]
	simple	−0.096	−0.071	−0.077	−0.054	−0.074	[−0.102, −0.047]

Table 17: Obs-C2−NoElim over eight stride-100 seeds: the four seeds of Table 16 and four new paired runs, with the mean, the 95% paired t-interval ($df = 7$), and the number of positive seeds. Over the first four seeds the means are $+0.022$, $+0.040$, and $+0.032$; over the new four, $+0.003$, $+0.070$, and $+0.013$. The criterion pre-declared for this contrast, a group interval above zero, is not met.

Seed	group	hard	simple
3407	+0.0438	−0.0227	+0.0683
3507	−0.0159	+0.0357	−0.0171
3607	+0.0601	+0.0032	+0.0725
3707	−0.0002	+0.1442	+0.0037
3807	−0.0355	+0.1426	−0.0323
3907	+0.0075	+0.0016	+0.0143
4007	+0.0061	+0.1248	+0.0295
4107	+0.0321	+0.0113	+0.0412
Mean	+0.0122	+0.0551	+0.0225
95% interval	[−0.0141, +0.0386]	[−0.0035, +0.1137]	[−0.0091, +0.0541]
Positive seeds	5/8	7/8	6/8

Table 18: Parts of the contradicted clause and weighted versions of the undecidable one, eight stride-100 seeds (3407–4107): arm minus NOELIM at the same seed, with the mean, the 95% paired t-interval ($df = 7$), and the number of positive seeds. *Full profile* removes the C2 candidates whose source and target have the same observed positive profile, and *partial overlap* the other C2 candidates; they are 15.40% and 84.60% of the loss-side C2 candidates (App. C), and their removal counts sum to those of OBS-C2 at every seed. *Text→image* applies the full-profile removal in that direction of the loss only. WEIGHT-C1 weights each C1 candidate by w_{ij} (Remark 1); the constant variant weights every C1 candidate by $\mathbb{E}[w_{ij} \mid \mathrm{C1}] = 0.847$; neither removes a pair. Pre-declared before the runs: the success criterion (interval above zero and at least seven positive seeds), the hypothesis that the full-profile pairs carry the gain of OBS-C2, and, for the weighted arms, a null band $|\Delta| < 0.005$ on the group probe, the primary endpoint.

Arm	Probe	Mean	95% interval	Positive
C2, full profile	group	-0.0060	$[-0.0216, +0.0097]$	3/8
	hard	$+0.0020$	$[-0.0425, +0.0466]$	4/8
	simple	-0.0070	$[-0.0281, +0.0141]$	3/8
C2, partial overlap	group	$+0.0207$	$[+0.0049, +0.0364]$	7/8
	hard	$+0.0707$	$[+0.0296, +0.1118]$	8/8
	simple	$+0.0343$	$[+0.0158, +0.0529]$	8/8
C2, full profile, text→image	group	$+0.0124$	$[-0.0030, +0.0278]$	6/8
	hard	$+0.0089$	$[-0.0246, +0.0425]$	5/8
	simple	$+0.0149$	$[-0.0056, +0.0353]$	5/8
WEIGHT-C1	group	-0.0015	$[-0.0045, +0.0015]$	4/8
	hard	-0.0071	$[-0.0154, +0.0012]$	2/8
	simple	-0.0016	$[-0.0046, +0.0013]$	4/8
WEIGHT-C1, constant weight	group	-0.0005	$[-0.0040, +0.0030]$	4/8
	hard	-0.0047	$[-0.0133, +0.0040]$	2/8
	simple	-0.0006	$[-0.0038, +0.0026]$	4/8

Table 19: Per-seed values for RAND-66 and the hard and simple probes of REACH-7; the group probe of REACH-7, $0.540_{\pm 0.022}$, is summarized in Table 3. The loss-side fraction is in percent of directed candidate slots.

Arm	Seed	At loss	group	hard	simple
RAND-66	3407	66.2117	0.5642	0.5138	0.5875
	3507	66.1421	0.5440	0.5883	0.5695
	3607	66.1436	0.5469	0.6191	0.5910
	3707	66.1613	0.5469	0.5430	0.5692
REACH-7	3407	14.4506	—	0.5348	0.5479
	3507	14.4241	—	0.5316	0.5592
	3607	14.4410	—	0.4489	0.5348
	3707	14.4024	—	0.6143	0.5893

J GROUND-TRUTH TESTBED

Construction. Four of the eleven coordinates, mesangial lesion, segmental sclerosis, crescent, and global sclerosis, are examined in every row (Table 6), so in the label space they span every row has complete labels. The testbed is built on the unfiltered pool of $48{,}068$ rows ($38{,}004$ training, $4{,}797$ validation, $5{,}267$ test) rather than on the $47{,}909$-crop pool of §7, and every test row is scored. We hide entries of the training rows only, with one draw of uniform random numbers (seed 20260926) shared by the two mechanisms, so that the conditions differ only in the hiding rule. Label-independent hiding examines every entry with probability 0.5 (realized 0.4963 for positive entries, 0.4990 for negative ones, 0.4986 overall); positive-preferential hiding examines positive entries with probability 0.10 and negative ones with 0.5629, which keeps the overall rate near one half (realized 0.1026, 0.5619, 0.4995). Training profiles, texts, and supervision masks are built from the entries that remain observed, which leaves 14 and 7 distinct observed training profiles, against 15 in the complete test labels; validation and test rows keep complete labels. NOELIM, OBS-C1, and WEIGHT-C1 are trained under each mechanism on seeds 3407–3707 with the configuration of

App. G, and WEIGHT-C1 uses the examined prevalence of its own condition. The sampler draws at most 16 label groups of 4 crops per batch and does not pad, so a batch holds at most 56 crops under independent and 28 under preferential hiding, and each run keeps the epoch with the best validation group top-1, the earlier one on ties.

Error rate of the undecidable clause. On 10^6 ordered, non-self training pairs drawn uniformly, with R_{ij} the source's observed positive set, 133,083 pairs are C1 under label-independent hiding and 26.04% of them are wrong against the hidden labels; under positive-preferential hiding 30,967 are C1 and 41.48% are wrong. Hiding positives thus removes C1 pairs of both kinds, about $34{,}700 \rightarrow 12{,}800$ wrong and $98{,}400 \rightarrow 18{,}100$ correct, so a higher wrong share is not a larger number of wrong pairs. Unlike the rates of §4, these are over uniform pairs; at the loss, over ten epochs and four seeds, OBS-C1 removes 21.30% of the negative slots under independent hiding and 7.33% under preferential hiding. The examined prevalences of the four coordinates, in the order above, are $(0.3425, 0.0301, 0.0384, 0.1289)$ under label-independent hiding and $(0.0889, 0.0059, 0.0067, 0.0254)$ under positive-preferential hiding. Plugged into Remark 1, they predict the wrong share from the observed labels alone: $\mathbb{E}[1 - w_{ij} \mid \text{C1}]$ is 0.2561 against the measured 0.2604 under independent hiding, and 0.0675 against 0.4148 under preferential hiding, where the examined prevalence understates the hidden one.

Endpoints and results. Profile top-1 assigns each test crop to the most similar of the 16 four-coordinate profile texts and scores it against the crop's complete profile; macro AUC averages over the four coordinates the AUROC of a per-coordinate score aggregated from the scores of all candidate profiles. Table 20 gives every run and Table 21 the paired differences. Hiding positives preferentially costs NOELIM 0.035 in macro AUC ($[+0.028, +0.041]$, every seed), and OBS-C1 recovers two thirds of it; its macro-AUC gain is 0.015 larger under preferential than under independent hiding ($[+0.012, +0.017]$, every seed). OBS-C1 raises the AUROC of every coordinate under both mechanisms, and each gain is larger under preferential hiding, where more of the C1 pairs on that coordinate are wrong (Table 22). A constant prediction of the most common profile scores 0.4817 on the complete labels (2,537 of 5,267 rows), below every run.

Table 20: Ground-truth testbed: every run, scored on complete test labels. *Independent*: examination independent of the label; *preferential*: positives examined with probability 0.10 and negatives with 0.5629.

		profile top-1				macro AUC			
Hiding	Arm	3407	3507	3607	3707	3407	3507	3607	3707
independent	NOELIM	0.7750	0.7572	0.7574	0.7674	0.9516	0.9486	0.9485	0.9503
	OBS-C1	0.7463	0.7602	0.7788	0.7653	0.9596	0.9589	0.9605	0.9562
	WEIGHT-C1	0.7760	0.7691	0.7613	0.7718	0.9524	0.9541	0.9491	0.9483
preferential	NOELIM	0.7501	0.7606	0.7756	0.7488	0.9125	0.9152	0.9185	0.9130
	OBS-C1	0.7906	0.7946	0.7849	0.7858	0.9348	0.9388	0.9439	0.9362
	WEIGHT-C1	0.7537	0.7631	0.7773	0.7513	0.9127	0.9157	0.9188	0.9133

Scoring against the record. The profile probes of §7.2 score each crop against the profiles of the record, in its closed world. Table 23 rescores the testbed checkpoints that way, without retraining: the test rows are hidden with the same draw as the training rows, and each crop is assigned to the most similar of the profiles observed in the hidden training split (14 and 7) and scored against its recorded profile, which is among them for all but one of the 5,267 rows. Every run loses, OBS-C1 the most, and its difference from NOELIM is negative on every seed under both mechanisms, while predicting the most common recorded profile for every crop scores above every run.

Stain-consistent candidates on the pathology record. The same reading applies to OBS-C1 on the corpus. We rescore the paired NOELIM and OBS-C1 checkpoints of Table 3 (seeds 3407–3707), without retraining, on all 5,267 test rows against all 164 recorded profiles, each profile text passing through the frozen text encoder and the checkpoint's own text head. Without restriction, this all-profile probe shows the cost of OBS-C1 that the frozen probes record, negative on every seed. Restricting each crop's candidates to the profiles whose positive coordinates it had examined

Table 21: Ground-truth testbed: arm minus NOELIM at the same seed and hiding draw, with the mean and the 95% paired t-interval ($df = 3$). Means and intervals are computed from unrounded scores and can differ in the last digit from values recomputed from printed ones. The interaction is the OBS-C1 effect under preferential hiding minus that under independent hiding. OBS-C1−WEIGHT-C1 is $+0.008$ $[+0.004, +0.012]$ (independent) and $+0.023$ $[+0.021, +0.025]$ (preferential) in macro AUC and $+0.028$ $[+0.006, +0.049]$ in profile top-1 under preferential hiding, positive on every seed.

Contrast	Endpoint	3407	3507	3607	3707	Mean	95% interval
Independent hiding							
OBS-C1	top-1	-0.0287	$+0.0030$	$+0.0215$	-0.0021	-0.0016	$[-0.0345, +0.0314]$
	AUC	$+0.0081$	$+0.0102$	$+0.0119$	$+0.0060$	$+0.0091$	$[+0.0049, +0.0132]$
WEIGHT-C1	top-1	$+0.0009$	$+0.0120$	$+0.0040$	$+0.0044$	$+0.0053$	$[-0.0021, +0.0128]$
	AUC	$+0.0008$	$+0.0054$	$+0.0006$	-0.0020	$+0.0012$	$[-0.0037, +0.0061]$
Preferential hiding							
OBS-C1	top-1	$+0.0404$	$+0.0340$	$+0.0093$	$+0.0370$	$+0.0302$	$[+0.0076, +0.0527]$
	AUC	$+0.0223$	$+0.0236$	$+0.0254$	$+0.0231$	$+0.0236$	$[+0.0215, +0.0257]$
WEIGHT-C1	top-1	$+0.0036$	$+0.0025$	$+0.0017$	$+0.0025$	$+0.0026$	$[+0.0013, +0.0038]$
	AUC	$+0.00024$	$+0.00043$	$+0.00026$	$+0.00023$	$+0.00029$	$[+0.00015, +0.00044]$
Interaction (preferential − independent)							
OBS-C1	top-1	$+0.0691$	$+0.0310$	-0.0122	$+0.0391$	$+0.0318$	$[-0.0216, +0.0851]$
	AUC	$+0.0142$	$+0.0134$	$+0.0135$	$+0.0171$	$+0.0146$	$[+0.0117, +0.0174]$

Table 22: Ground-truth testbed, per coordinate. *C1 wrong*: among the C1 pairs that require the coordinate and leave it unexamined, the share in which it is present. ΔAUC: OBS-C1−NOELIM in the coordinate's AUROC, mean over four seeds.

	Independent		Preferential	
Coordinate	C1 wrong	ΔAUC	C1 wrong	ΔAUC
mesangial lesion	34.6%	$+0.016$	51.9%	$+0.028$
segmental sclerosis	3.1%	$+0.007$	6.0%	$+0.034$
crescent	3.8%	$+0.012$	7.5%	$+0.026$
global sclerosis	13.3%	$+0.003$	23.7%	$+0.006$

Table 23: Testbed checkpoints rescored against the recorded labels: profile top-1 over the profiles observed in the hidden training split (14 under independent, 7 under preferential hiding), with test rows hidden by the same draw. *Majority*: the most common recorded profile predicted for every crop. Δ: OBS-C1−NOELIM at the same seed, with the mean and the 95% paired t-interval ($df = 3$), against the recorded labels and, repeating Table 21, against the complete ones.

Hiding	Row	3407	3507	3607	3707	Mean	95% interval
independent	NOELIM	0.6197	0.6271	0.6062	0.6226	0.6189	
	OBS-C1	0.5489	0.5667	0.5899	0.5943	0.5749	
	WEIGHT-C1	0.6138	0.6336	0.5990	0.5971	0.6109	
	majority		constant			0.7296	
	Δ, recorded	-0.0708	-0.0604	-0.0163	-0.0283	-0.0440	$[-0.0850, -0.0029]$
	Δ, complete	-0.0287	$+0.0030$	$+0.0215$	-0.0021	-0.0016	$[-0.0345, +0.0314]$
preferential	NOELIM	0.5738	0.5293	0.5202	0.5441	0.5419	
	OBS-C1	0.5033	0.5005	0.4492	0.4593	0.4781	
	WEIGHT-C1	0.5681	0.5269	0.5176	0.5407	0.5383	
	majority		constant			0.9421	
	Δ, recorded	-0.0704	-0.0289	-0.0710	-0.0849	-0.0638	$[-0.1023, -0.0252]$
	Δ, complete	$+0.0404$	$+0.0340$	$+0.0093$	$+0.0370$	$+0.0302$	$[+0.0076, +0.0527]$

($P \subseteq S_i$; the empty profile always qualifies, and every crop's own profile does) removes it: NOELIM scores 0.530 and OBS-C1 0.531 (Table 24). The frozen probes use the 58 profiles frozen at training

time and the 5,206 test crops whose profile is among them (App. G), so the restriction is compared within the all-profile probe, on the same rows, candidates, and checkpoints.

Table 24: OBS-C1−NOELIM on the pathology record at the same seed, with the mean and the 95% paired t-interval over the four seeds ($df = 3$). Frozen probes: Table 3. All-profile probe: 164 recorded profiles and 5,267 test rows, without and with the stain-consistent restriction $P \subseteq S_i$.

Probe	3407	3507	3607	3707	Mean	95% interval
frozen, group	-0.0088	-0.0083	-0.0106	-0.0061	-0.0085	$[-0.0114, -0.0056]$
frozen, simple	-0.0117	-0.0082	-0.0253	-0.0094	-0.0136	$[-0.0262, -0.0011]$
all-profile	-0.0144	-0.0122	-0.0161	-0.0080	-0.0127	$[-0.0183, -0.0071]$
all-profile, $P \subseteq S_i$	-0.0015	$+0.0021$	-0.0021	$+0.0046$	$+0.0008$	$[-0.0042, +0.0057]$

K COCO AT CONTROLLED COVERAGE

Setup. Training uses the 118,287 images of COCO train2017 and evaluation the 5,000 of val2017, with image-level labels for the 80 instance categories (Lin et al., 2014). Image and text features come from a frozen CLIP ViT-B/16 (Radford et al., 2021). Every arm trains the same projection head ($512 \to 1024 \to 512$, L2-normalized outputs) with a sigmoid image–text loss over all 80 category texts, in the manner of Zhai et al. (2023): logits are 10 times the cosine between projected image and text features, and the loss is the weighted binary cross-entropy summed over image–category pairs and divided by the batch's total weight. Training texts read "a photo containing a {category}"; every arm is evaluated with "a photograph of the {category}", which no arm saw. Each run trains three epochs with AdamW (learning rate 0.002, weight decay 10^{-4}, batch 512), and its final checkpoint is evaluated, so nothing is selected on validation data. At coverage $c \in \{1, 0.75, 0.5, 0.35, 0.2\}$ each training image is examined on $\mathrm{round}(80c)$ categories (80, 60, 40, 28, 16) drawn independently per image, and one mask per seed and coverage is shared by all arms. Examined labels keep their $0/1$ targets in every arm. AN gives unexamined categories target 0, the closed world; MASK gives them weight 0, the analogue of OBS-C1; ET-PRIOR gives them the soft target $\hat{\pi}_\ell$, the category's prevalence among the training images that examined it, which is the expected target under the independence assumption of Remark 1; NONSENSE keeps the targets of AN but trains on meaningless category strings, the control of §6. At $c = 1$, AN and ET-PRIOR have identical targets, per-epoch losses, and per-category AUCs on every seed. Macro AUC averages the per-category AUROC over the 80 categories on val2017, read once, after all runs had finished; intervals are 95% paired t-intervals over seeds 3407–3707 ($df = 3$).

Declared prediction. The dose is the share of the closed-world negative mass that falls on present categories. With $y_{i\ell}$ the complete label, $s_{i\ell}$ the examination indicator, and fixed category weights α_ℓ proportional to the number of training images containing ℓ, the negative mass of AN is $D = N^{-1} \sum_{i,\ell} \alpha_\ell (1 - y_{i\ell} s_{i\ell})$, its wrong part is $W = N^{-1} \sum_{i,\ell} \alpha_\ell \, y_{i\ell} (1 - s_{i\ell})$, and $M_{\mathrm{wrong}} = W/D$. The record estimates W by $\widehat{W} = N^{-1} \sum_{i,\ell} \alpha_\ell (1 - s_{i\ell}) \hat{\pi}_\ell$, with no hidden label, and $\widehat{M}_{\mathrm{wrong}} = \widehat{W}/D$ is within 2×10^{-4} of M_{wrong} at every coverage (Table 25). Before training we computed $\widehat{M}_{\mathrm{wrong}}$ on CPU and committed it with the code, together with a prediction of how the gain of ET-PRIOR over AN follows it. Among the declared predictions, the three about the diagnostic hold: the gain increases with $\widehat{M}_{\mathrm{wrong}}$ across the five coverages (declared Spearman ≥ 0.9; observed 1.0, Pearson 0.998), it is positive on all four seeds at $c = 0.2$ with a paired interval above zero ($[+0.009, +0.017]$), and it vanishes at $c = 1$. The first declared design examined one fixed category subset per pseudo-source; at $c = 0.2$ the three subsets cover at most 48 of the 80 categories and leave the prevalence of the rest unestimable, so before any checkpoint was written or validation label read we kept that version on record and revised it to per-image draws.

Results. Not treating unexamined labels as absent helps in both forms. MASK's gain over AN grows from 0.001 at $c = 0.75$ to 0.012 at $c = 0.2$, where it recovers 71% of the 0.017 that AN loses between $c = 1$ and $c = 0.2$, and ET-PRIOR gains 0.005 to 0.013 with every interval above zero. NONSENSE stays between 0.55 and 0.57, so the anchored range (Eq. 1) is at least 0.37 at every coverage.

Table 25: COCO at controlled coverage: macro AUC on val2017, mean over four seeds. M_{wrong} as estimated before training from the record ($\widehat{M}_{\text{wrong}}$) and as measured with the hidden labels. The last column is ET-PRIOR−AN with the 95% paired t-interval ($df = 3$).

	M_{wrong}						
c	record	hidden	AN	MASK	ET-PRIOR	NONSENSE	ET-PRIOR−AN
1.00	0	0	0.9489	0.9489	0.9489	0.5495	identical targets
0.75	0.0380	0.0379	0.9479	0.9488	0.9532	0.5537	$+0.0053$ $[+0.0036, +0.0070]$
0.50	0.0734	0.0734	0.9443	0.9475	0.9532	0.5614	$+0.0088$ $[+0.0069, +0.0108]$
0.35	0.0934	0.0934	0.9412	0.9472	0.9519	0.5678	$+0.0108$ $[+0.0074, +0.0141]$
0.20	0.1127	0.1125	0.9316	0.9439	0.9449	0.5702	$+0.0133$ $[+0.0094, +0.0172]$

Scored against the record. Rescoring the same checkpoints against val2017 labels hidden by the training rule, with unexamined entries read as absent, leaves every gain in place (Table 26): at $c = 0.2$, MASK−AN moves from $+0.0123$ to $+0.0120$ and ET-PRIOR−AN from $+0.0133$ to $+0.0132$, every seed stays positive at every coverage, and scoring the examined entries alone reproduces the complete-label values. Per-category AUC has to behave this way: when hiding ignores the label, a category's recorded negatives mix its true negatives with a random share q_ℓ of its positives, so its recorded AUC is $(1 - q_\ell)\,\text{AUC} + q_\ell/2$ in expectation (Menon et al., 2015), a shrinkage that preserves the order of arms on every category. The reversal of §7.3 needs an endpoint that scores recorded label sets as wholes in a closed candidate world; per-label ranking against a record hidden independently of the label cannot produce it.

Table 26: COCO checkpoints rescored: arm minus AN in macro AUC, mean over four seeds, against complete val2017 labels, against labels hidden by the training rule with unexamined entries as absent (recorded), and over examined entries only.

	MASK−AN			ET-PRIOR−AN		
c	complete	recorded	examined	complete	recorded	examined
0.75	$+0.001$	$+0.001$	$+0.001$	$+0.005$	$+0.005$	$+0.005$
0.50	$+0.003$	$+0.003$	$+0.003$	$+0.009$	$+0.009$	$+0.009$
0.35	$+0.006$	$+0.006$	$+0.006$	$+0.011$	$+0.011$	$+0.011$
0.20	$+0.012$	$+0.012$	$+0.012$	$+0.013$	$+0.013$	$+0.013$

A similarity-quantile rule on the record. Detectors that flag negatives by embedding similarity remove what the record could have protected. As a stand-in for the per-anchor quantile rules that such detectors learn (Balmaseda et al., 2025), we flag, for each of the first 5,000 training images, the 5% of candidate categories (all but the examined positives) with the highest cosine to the image on the frozen features, under the training masks of four seeds (Table 27). As coverage rises from 0.2 to 0.8, the share of flags that land on examined, recorded-negative entries, which the record certifies, rises from 14% to 73%, the share the knowledge-graph rule reached on the pathology corpus (Table 1). Restricted to unexamined entries, the rule keeps a precision of 32–34% present categories at every coverage, nine times the 3.6% of a random unexamined entry. The record and the detector are complementary: the record withholds the certified flags, and the detector concentrates the undecidable ones.

Table 27: Similarity-quantile flags on COCO training features (5,000 images, four seeds pooled): the share of flags on entries the record certifies as negative, the share of flags on present categories, the same share among flags on unexamined entries, and the present share of a random unexamined entry.

c	Flags	Certified	Present, all	Present, unexamined	Random unexamined
0.2	91,276	14.3%	28.8%	33.6%	3.6%
0.5	84,387	40.1%	19.7%	32.8%	3.7%
0.8	81,278	73.2%	8.5%	31.9%	3.6%

L NATURAL RECORDS: LVIS, OPEN IMAGES, CHEXPERT, AND VINDR-CXR

LVIS against COCO. LVIS (Gupta et al., 2019) annotates COCO images with 1,203 categories in a federated design: for each category a positive set is annotated exhaustively and a negative set is verified absent, and the category is unexamined in every other image; a not-exhaustive flag marks images whose positives may be incomplete and is kept unexamined here. On the 100,170 images that LVIS v1 train shares with COCO train2017, 63 COCO categories map to exactly one LVIS category by name or listed synonym. Table 28 audits their image–category entries against COCO's instance annotations. The record examines 2.39% of the entries; of the 6,160,015 unexamined ones, 152,343 hold a COCO-annotated object, a lower bound of 2.47% since COCO's own misses are not verified; the 54,492 verified absences disagree with a COCO annotation 111 times. Examination depends on the label: LVIS examined 37.6% of the entries COCO annotates as present and 0.97% of the rest, so the prevalence among examined entries, 0.64 by LVIS's own labels, is far above the share of unexamined entries that hold an object.

Table 28: LVIS v1 train against COCO train2017 on their 100,170 shared images and 63 one-to-one categories.

Image–category entries	6,310,710
Examined by LVIS (present or verified absent)	150,695 (2.39%)
Unexamined (C1)	6,160,015
with a COCO-annotated object	152,343 (2.47%, lower bound)
Verified absent yet COCO-annotated present	111 of 54,492 (0.20%)
LVIS examination rate, entries COCO annotates present	91,680 of 244,023 (37.6%)
LVIS examination rate, other entries	59,015 of 6,066,687 (0.97%)

Training on the LVIS record. We train the image–text head of App. K, with the same loss, prompts, optimizer, epochs, and seeds, on the frozen CLIP ViT-B/16 features of the 84,001 shared images that carry at least one recorded entry for the 63 categories, with LVIS's record as the only training label, and score the 5,000 images of COCO val2017 against their complete instance labels. AN, MASK, and ET-PRIOR are defined as on COCO; MASK ignores 97.2% of the entries, and ET-PRIOR fills them with the examined prevalence, 0.62 on average. ET-AUDIT fills them with the per-category prevalence estimated by the 1% audit below, and ORACLE, a reference rather than a method, fills them with their true labels from COCO's annotations of the training images. Table 29: masking beats the closed world on every seed in macro AP and macro AUC; the examined prevalence as a target costs 0.14 AP, its prior being 25 times the unexamined one, and the audited prevalence removes that cost; the true labels add 0.024 AP over the closed world and 0.013 ($[+0.006, +0.020]$) over masking. With a two-layer classifier head trained from scratch on the same record (59 exact-name categories, two seeds) and held-out training images scored against COCO's annotations, the closed world loses more, 0.53 against 0.67 for masking.

Table 29: Heads trained on the LVIS record, scored on COCO val2017's complete labels over 63 categories: macro AP and macro AUC per seed, and the difference from AN with the 95% paired t-interval ($df = 4$). ORACLE fills unexamined entries with their true labels and is a reference, not a method.

	Arm	3407	3507	3607	3707	3807	Mean	−AN
AP	AN	0.6973	0.6978	0.6991	0.6999	0.6991	0.6986	
	MASK	0.7086	0.7093	0.7173	0.7105	0.7033	0.7098	$+0.0112$ $[+0.0050, +0.0174]$
	ET-PRIOR	0.5312	0.5887	0.5706	0.5844	0.5180	0.5586	-0.1400 $[-0.1796, -0.1005]$
	ET-AUDIT	0.7001	0.6977	0.7024	0.6990	0.7037	0.7006	$+0.0020$ $[-0.0009, +0.0048]$
	ORACLE	0.7237	0.7193	0.7234	0.7212	0.7245	0.7224	$+0.0238$ $[+0.0210, +0.0267]$
AUC	AN	0.9409	0.9382	0.9410	0.9409	0.9392	0.9401	
	MASK	0.9597	0.9582	0.9593	0.9599	0.9566	0.9587	$+0.0187$ $[+0.0175, +0.0198]$
	ET-PRIOR	0.9384	0.9405	0.9370	0.9382	0.9333	0.9375	-0.0026 $[-0.0063, +0.0012]$
	ORACLE	0.9595	0.9604	0.9600	0.9593	0.9599	0.9598	$+0.0197$ $[+0.0177, +0.0218]$

Auditing the record. Where examination depends on the label, the record's prevalence is not the unexamined one, and a random audit is the cheapest replacement. Drawing 1%, 3%, or 5% of the $6{,}160{,}015$ unexamined entries at random, in five independent draws each, and reading COCO's annotation for the drawn entries, puts the annotated-positive rate of the unexamined entries at 0.0249, 0.0247, and 0.0248, averaged over categories, against 0.0247 over all of them (Table 28) and 0.62 for the examined prevalence; every draw lies between 0.0244 and 0.0252. Table 30 repeats the audit where the hidden entries are known. A 5% audit lands within 0.01 of the hidden-positive rate on every source, with an interval that covers it, whether the record overstates the rate 25-fold (LVIS) or understates it 9-fold (the testbed with positives hidden preferentially); where hiding ignores the label, the record already has it (the testbed and VinDr-CXR). The audit needs no model and no assumption about why entries went unexamined. As a training target on LVIS (ET-AUDIT, Table 29), the audited prevalence removes the 0.14 AP cost of the record's own prevalence ($+0.002$ against the closed world, $[-0.001, +0.005]$); masking and the hidden labels themselves remain above it.

Table 30: Auditing the unexamined entries. *True*: the share of unexamined entries that are positive (LVIS: COCO-annotated), with a 95% bootstrap interval over rows; *record*: the examined prevalence, the prior a record-only estimate uses (Table 29 uses its per-category values, which average 0.62); *audit*: the same share on a random 5% of the unexamined entries, with a 95% bootstrap interval over the drawn entries. On the testbed this is the coordinate-level rate, not the pair-level wrong share of App. J.

Source	True	Record	Audit, 5%
LVIS record against COCO	0.0247 $[0.0246, 0.0249]$	0.608	0.0250 $[0.0245, 0.0255]$
Testbed, positives hidden preferentially	0.244 $[0.241, 0.246]$	0.028	0.245 $[0.233, 0.257]$
Testbed, hiding ignores the label	0.137 $[0.134, 0.139]$	0.135	0.135 $[0.125, 0.145]$
VinDr-CXR, half hidden at random	0.144 $[0.141, 0.148]$	0.143	0.137 $[0.122, 0.152]$

Open Images pilot. With the human-verified image labels of Open Images V7 (Kuznetsova et al., 2020), we train the same head for 12 labels on $2{,}048$ images drawn from the validation split, with verified-present and verified-absent labels as the record and everything else unexamined, and read macro AP on the verified entries of $1{,}024$ test-split images (Table 31). Masking unexamined entries scores 0.90 against 0.75 for the closed world on both seeds, and the prevalence target 0.82.

Table 31: Open Images pilot: macro AP over 12 labels on the verified entries of $1{,}024$ test images, two seeds.

Arm	3407	3507	Mean
AN	0.7499	0.7492	0.7495
MASK	0.8953	0.8961	0.8957
ET-PRIOR	0.8385	0.8068	0.8227

CheXpert and VinDr-CXR. CheXpert (Irvin et al., 2019) labels 13 findings from reports as positive, negative, uncertain, or blank, and a blank means the finding was not mentioned. Reading blanks as unexamined, the $223{,}414$ training studies give corpus rates over ordered pairs of $\rho_{\mathrm{C1}} = 0.732$, $\rho_{\mathrm{C2}} = 0.142$, and $\rho_{\mathrm{fab}} = 0.874$ with uncertain labels also read as unexamined, and 0.797, 0.112, and 0.908 with uncertain read as present; the 234 validation studies are fully labelled and give $\rho_{\mathrm{C1}} = 0$. Mapping blanks to zero, the usual practice, is therefore the substitution $S_i := \mathcal{L}$ on the 73% of candidate pairs that are undecidable. On VinDr-CXR (Nguyen et al., 2022), where three radiologists read each of $15{,}000$ images for 14 findings, hiding reader entries at random and estimating the fabricated share of candidate pairs from the observed entries alone, with a 12-class latent-class model, recovers it within 0.0003, 0.002, 0.012, and 0.015 at coverage 0.75, 0.50, 0.35, and 0.20; when the hiding depends on the label (positives observed at rate 0.81 against 0.48), the same model overstates the share by 0.069, the counterpart of the understatement on the testbed.

M PROVENANCE AND EVALUATION BUDGET

A provenance scan re-derives 67 run-backed cells from run artifacts (Table 2, the values plotted in Figure 1b, the loss-side fractions of the same-group-positive arms in Table 12, the REACH-7 probe, the per-seed held-out contrasts of Table 15, and the testbed endpoints, closed-world rescorings, and stain-consistent rescoring of App. J): none is inconsistent and all five cross-cell assertions hold; three cells are skipped because they belong to external baselines without a run directory. Printed SDs are checked to be sample SDs, and a retrained RAND-33 seed-3407 run reproduces the metrics file of an earlier seed-3407 run byte for byte (identical SHA-256). All 40 historical runs share one code revision, but two evaluation-side data products changed during the campaign; re-evaluating frozen checkpoints against frozen manifests retired 20 of the 40, including the run that had held the best held-out number; an older consecutive-seed held-out batch is retired as well. The held-out budget was fixed in advance at ten reads and is spent. The per-coordinate audit of App. H re-analyses stored scores and uses no read; the arms trained after the budget was spent (random, observation, and keep-all corrections) are read on the profile probes only. The eight-seed arms (Tables 17 and 18), the COCO runs (App. K), and the natural-record runs (App. L) are not yet part of the scan. A launcher error that made some of the later runs fall back to the default seed 3407 was caught from the seeds recorded in their metrics files; the affected runs were set aside and rerun with the intended seeds. The testbed and its rescorings use no held-out read, and the COCO, LVIS, Open Images, CheXpert, and VinDr-CXR analyses read only public labels.

N FROZEN-FEATURE PROBE

A linear probe on frozen features, trained and evaluated identically for each encoder on the clean-gold validation slides, gives a mean AUC over 27 stain–lesion rows of 0.906 for the in-domain backbone (DINO teacher weights), 0.874 for PLIP (Huang et al., 2023), and 0.840 for CLIP (Radford et al., 2021). Paired slide-cluster bootstrap intervals exclude zero for all three pairwise differences.

O EXTENDED RELATED WORK

Negatives in contrastive and metric learning. Contrastive objectives (van den Oord et al., 2018; Chen et al., 2020; He et al., 2020; Khosla et al., 2020; Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023) assume that negatives are trustworthy; Arora et al. (2019) show that collisions between a positive pair and its negatives directly bound downstream performance, and analyses of the loss geometry show that the repulsive term alone sets the uniformity of the learned sphere and the effective hardness of each negative (Wang and Isola, 2020; Wang and Liu, 2021; Yeh et al., 2022). Methods that relax the assumption estimate which negatives to cancel, reweight, or down-sample from the learned representation (Huynh et al., 2022; Chen et al., 2022; Chuang et al., 2020; Robinson et al., 2021; Kalantidis et al., 2020; Dwibedi et al., 2021; Zheng et al., 2021), and dense retrieval faces the same issue when hard negatives are mined from unlabelled passages (Xiong et al., 2021; Qu et al., 2021). Recent vision–language work relabels or filters false negatives with a pretrained matcher, learned per-anchor thresholds, or a learned batch-construction policy (Byun et al., 2024; Bulat et al., 2024; Balmaseda et al., 2025; Kim et al., 2025); medical image–text methods replace hard InfoNCE targets by label- or report-similarity soft targets (Wang et al., 2022b; Ko and Park, 2025) or make the debiasing prior sample-specific (Wang et al., 2023), and the last reports that removing or reweighting suspected false negatives helps classification but not grounding or retrieval, a metric dependence we observe in a sharper form (§7). That *which* negatives a loss sees governs what it learns is established in metric learning (Schroff et al., 2015; Sohn, 2016; Wu et al., 2017; Shrivastava et al., 2016), where Musgrave et al. (2020) showed that many reported gains vanish under matched protocols. Our diagnostic reads the relevant uncertainty from the recorded observation process rather than from the model, so it is available before training and fixed during it; our matched-fraction and matched-key controls are in the spirit of that reality check.

Partial, positive–unlabelled, and noisy multi-label supervision. Classifiers trained under weak or partial labels (Sun et al., 2010; Bucak et al., 2011; Yu et al., 2014; Xie and Huang, 2018; Durand et al., 2019; Huynh and Elhamifar, 2020; Ben-Baruch et al., 2022; Kim et al., 2022) or a single positive per image (Cole et al., 2021) face the same missingness, which the multi-label surveys treat

as a property of the label matrix rather than of the annotation protocol (Zhang and Zhou, 2014; Liu et al., 2022); Kim et al. (2022) show that the false negatives implied by treating missing labels as absent are exactly the large-loss terms, and asymmetric losses down-weight the resulting negative mass without deciding which entries are unexamined (Ridnik et al., 2021). Methods that combine this setting with contrastive objectives bootstrap missing-label semantics from the model (Chen et al., 2023), reject negative-loss terms the model flags (Song et al., 2024), or disambiguate partial labels contrastively (Wang et al., 2022a); all derive their evidence from the model being trained, so the criterion moves during training and cannot be audited beforehand. Our C1 clause is the positive–unlabelled ambiguity (Elkan and Noto, 2008; du Plessis et al., 2014; Kiryo et al., 2017; Bekker and Davis, 2020) restricted to recorded coordinates; class-prior-aware risk estimators for single-positive labels (Liu et al., 2025), entropy-based treatment of unannotated labels (Zhou et al., 2022a), and positive–unlabelled contrastive objectives (Acharya et al., 2024) estimate the relevant prior rather than reading which coordinates were examined. Partially labelled segmentation marginalizes over the classes a dataset did not annotate (Shi et al., 2021), the closest analogue of an observation-aware loss we know of. The observation mask is a missing-data indicator (Rubin, 1976; Little and Rubin, 2020; Mohan and Pearl, 2021), and its dependence on stain is an instance of structured missingness (Mitra et al., 2023). Label-noise estimation and benchmark audits (Natarajan et al., 2013; Frénay and Verleysen, 2014; Northcutt et al., 2021b; Beyer et al., 2020; Northcutt et al., 2021a) quantify errors in *recorded* entries; we quantify supervision manufactured from *unrecorded* ones.

Sampling and exposure. Resampling, importance weighting, and decoupled schedules change a loss's effective distribution (Chawla et al., 2002; Buda et al., 2018; Cui et al., 2019; Kang et al., 2020; Byrd and Lipton, 2019; Zhang et al., 2023); we measure how they change the *composition* of mined negatives. Implicit-feedback recommendation treats unobserved items as negatives (Hu et al., 2008; Rendle et al., 2009), and exposure modelling (Liang et al., 2016; Schnabel et al., 2016) records who could have clicked; the observation mask plays that role here. That data missing not at random cannot be corrected by reweighting the observed entries alone is the standard result in that literature and in econometrics (Heckman, 1979; Marlin and Zemel, 2009; Steck, 2010), and our C1 clause is its multi-label contrastive counterpart.

Controls for evaluation validity. The meaningless-text control follows control tasks and randomized-label diagnostics (Zhang et al., 2017; Hewitt and Liang, 2019; Pimentel et al., 2020; Voita and Titov, 2020; Belinkov, 2022), applied to an evaluation instrument rather than to a probe. It is the training-side counterpart of an evaluation-side observation: zero-shot scores move substantially with the wording of the prompt set (Zhou et al., 2022b; Menon and Vondrick, 2023), and Roth et al. (2023) obtain competitive accuracy from class descriptors made of random words, which shows that a prompt-based score can be carried by structure rather than meaning. We push the control one step earlier, into training, so that the floor of the metric is measured rather than argued. That a matched-vocabulary score survives word-order scrambling echoes bag-of-words behaviour in language and vision–language models (Sinha et al., 2021; Yuksekgonul et al., 2023) and benchmarks that reward artifacts and shortcuts (Gururangan et al., 2018; Geirhos et al., 2020; Lapuschkin et al., 2019; DeGrave et al., 2021; Torralba and Efros, 2011; Recht et al., 2019). Reported gains that do not survive a matched protocol, a tuned baseline, or a re-implementation are a recurring finding across subfields (Melis et al., 2018; Ferrari Dacrema et al., 2019; Musgrave et al., 2020; Liao et al., 2021; Raji et al., 2021; Bowman and Dahl, 2021), and have prompted calls for empirical rigour and for reporting score distributions rather than single runs (Sculley et al., 2018; Lipton and Steinhardt, 2019; Reimers and Gurevych, 2017). Our reporting protocol and seed audit sit with prior calls for variance accounting and provenance (Henderson et al., 2018; Dodge et al., 2019; Bouthillier et al., 2021; Pineau et al., 2021; Picard, 2021; Kapoor and Narayanan, 2023; Gebru et al., 2021; Sambasivan et al., 2021); what we add is that the evaluation vocabulary and the removal fraction are also provenance items. Medical vision–language pre-training (Zhang et al., 2022; Boecking et al., 2022; Tiu et al., 2022; Huang et al., 2023; Lu et al., 2024) is the setting of our training experiments; nothing in the diagnostic is specific to it.