Answer Prior Collapse Control Through Counterbalanced Curricula for Robust Medical Vision Language Binary Inference Across Rare Findings

Agus Firmansyah[1], Hendra Wijaya[2], and Ilham Prakoso[3]

[1]Universitas Teuku Umar, Faculty of Economics and Business, Department of Business Information Systems, Jalan Alue Peunyareng, Ujong Tanoh Darat, Meulaboh 23615, Aceh Barat, Indonesia

[2]Universitas Musamus, Faculty of Economics and Business, Department of Management Information Systems, Jalan Kamizaun Mopah Lama, Merauke 99611, Papua Selatan, Indonesia

[3]Universitas Khairun, Faculty of Economics and Business, Department of Digital Business and Information Systems, Jalan Jusuf Abdulrahman, Gambesi, Ternate Selatan 97719, North Maluku, Indonesia

ABSTRACT

Medical vision-language models are evaluated through binary questions that appear simple on the surface. In practice, these questions can expose a model's tendency to rely on answer frequency, diagnostic prevalence, or prompt regularities instead of image-conditioned reasoning. Rare findings are especially vulnerable because the training stream may contain many more negative than positive cases, and a model can obtain acceptable average accuracy by leaning toward the majority answer. This paper presents an empirical study of answer-prior collapse in binary medical vision-language inference and evaluates counterbalanced curricula designed to reduce it. We construct a multi-domain dataset of 24,000 image-question examples covering chest radiography, retinal imaging, dermatology, and musculoskeletal radiography. The training distribution is deliberately skewed across finding prevalence, while the test distribution includes balanced, natural-prevalence, and rare-positive evaluation panels. We compare standard empirical-risk training with class reweighting, focal training, oversampling, hard-positive replay, and a proposed counterbalanced curriculum that alternates prevalence-realistic batches with balanced diagnostic contrast batches. The proposed curriculum reduces rare-positive miss rate from 38.7% to 19.4%, lowers answer-prior index from 0.42 to 0.17, and improves macro-balanced accuracy by 7.8 percentage points while preserving performance on common negative cases. The results show that curriculum structure, not only loss weighting, materially affects whether a model learns image-conditioned binary decisions for rare clinical findings.

1. Introduction

Binary medical image question answering is often treated as a convenient evaluation format. A model receives an image, a clinical question, and a restricted answer space. The answer is yes or no. This format is attractive because the output can be scored clearly, the response can be linked to a label, and the result can be summarized by accuracy. Yet the same format also creates a hidden weakness. A model can appear competent by learning that some answers are common and others are uncommon. When the finding is rare, the easiest route is often to answer no [1]. When the dataset contains many negative examples for a target condition, a model can score well without learning the visual boundary that separates subtle positives from true negatives.

The present paper studies this failure as answer-prior collapse. The term refers to a training outcome in which the model's binary answer becomes too strongly tied to

Article info:
Agus Firmansyah, Hendra Wijaya, and Ilham Prakoso. *Answer Prior Collapse Control Through Counterbalanced Curricula for Robust Medical Vision Language Binary Inference Across Rare Findings.* Center for Theoretical and Applied STEM Research (CTASR) . Licensed under Attribution - NonCommercial - No Derivatives 4.0 International (CC BY-NC-ND 4.0)

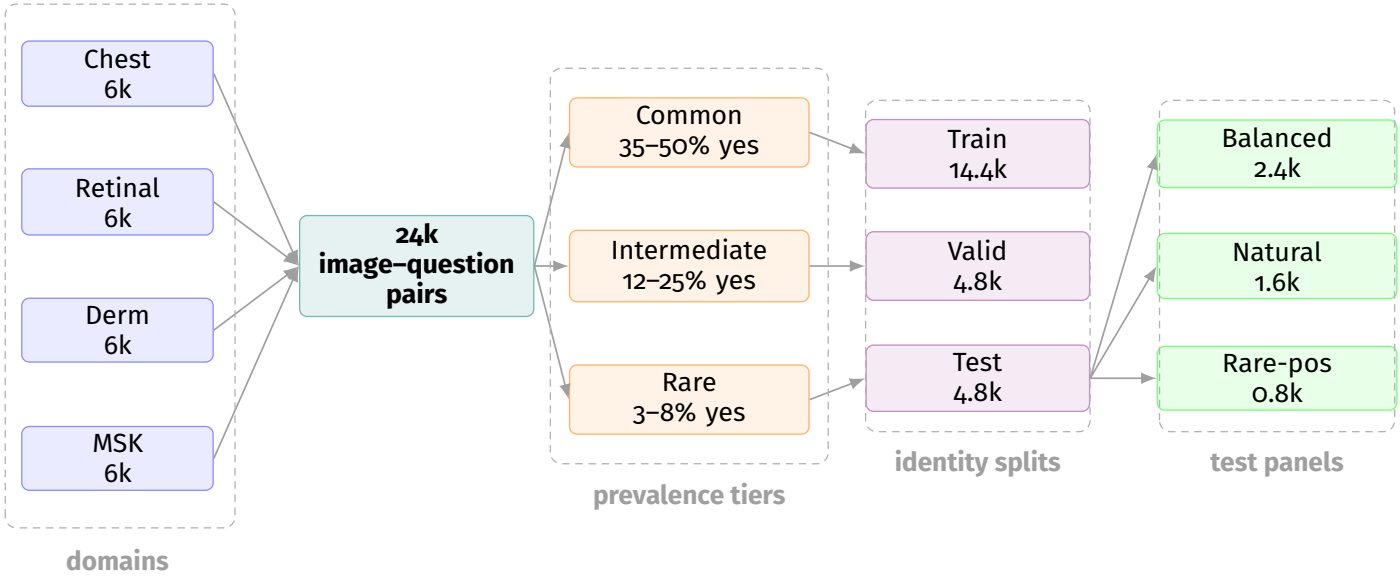


Figure 1 Multi-domain binary medical VQA corpus organized by imaging domain, target prevalence tier, identity-preserving split, and three held-out evaluation panels that separate visual discrimination from natural-prevalence scoring.

Table 1 Dataset composition across imaging domains.

Domain	Examples	Target Findings	Rare Positive Rate
Chest Radiography	6,000	9	3–8%
Retinal Imaging	6,000	9	3–8%
Dermatology	6,000	9	3–8%
Musculoskeletal Radiography	6,000	9	3–8%

the prior frequency of answers in the training distribution. The model does not necessarily answer the same way for every input. It may still respond to image quality, question wording, and finding category. The problem is more specific. For rare findings, it learns a conservative default that suppresses positive answers even when the image contains the relevant sign. This behavior is easy to miss if performance is reported only on a naturally imbalanced test set, because most examples are negative and a negative bias can look numerically respectable.

A useful medical vision-language model should not ignore prevalence. Disease probability matters, and some findings truly are uncommon [2]. The problem arises when prevalence overwhelms visible evidence. A binary question asks whether a finding is shown in the provided image, not whether it is common in the training corpus. The model should use prior knowledge as context, but the answer should remain responsive to the image. This distinction matters most for rare but important abnormalities. Missing a rare positive case can be more consequential than overcalling a common benign feature, depending on the clinical scenario.

This paper asks whether the structure of the training curriculum can reduce answer-prior collapse. The study does not rely on a new backbone architecture. It does not introduce a new interpretability tool. It does not use post-hoc calibration or evidence masking. Instead, it focuses on how training examples are sequenced and mixed. The central hypothesis is that ordinary minibatch training on an imbalanced stream teaches the model a strong answer prior before it has learned enough visual contrast for rare positives [3]. A curriculum that repeatedly interrupts prevalence-realistic batches with balanced diagnostic contrast batches may help the model retain the connection between visual evidence and the affirmative answer.

The work is motivated by a specific caution from prior medical vision-language consistency experiments. In a discussion of pure consistency training, Sadanandan and Behzadan (2026) described a collapse pattern in which a model learned to give the same answer across inputs because that shortcut satisfied the training objective while damaging discriminative behavior [4]. The present study examines a different form of collapse: not a universal constant answer, but a prevalence-shaped answer tendency that becomes especially harmful for rare positives. The lesson is that an objective can look reasonable while still allowing a shortcut unless the training design keeps the discriminative task anchored.

The empirical setup uses four imaging domains: chest

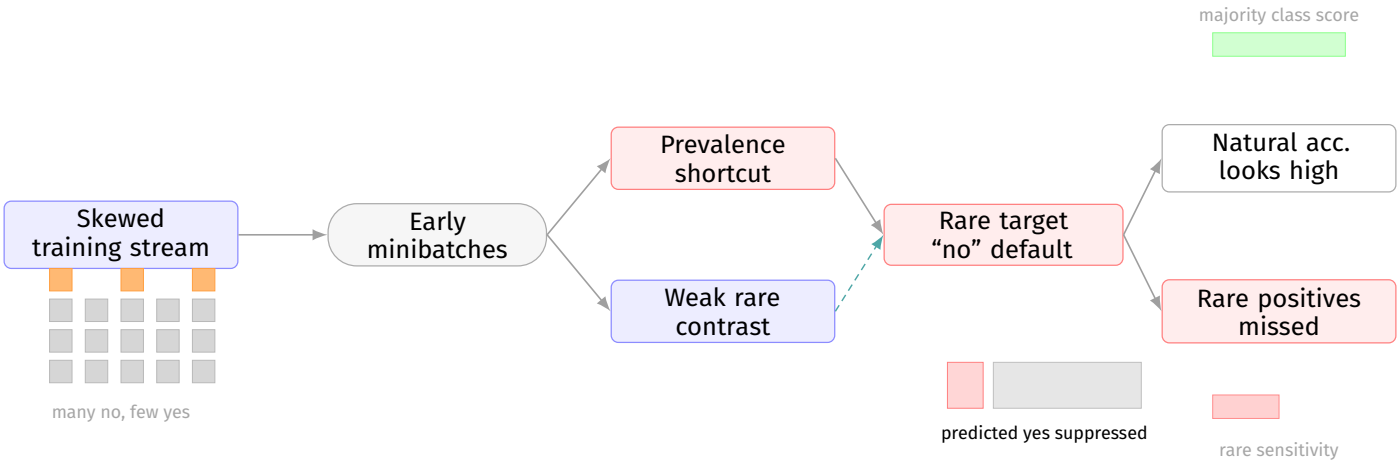


Figure 2 Answer-prior collapse occurs when an imbalanced binary training stream teaches a prevalence-shaped answer tendency before the model has learned stable visual contrast for rare positive findings.

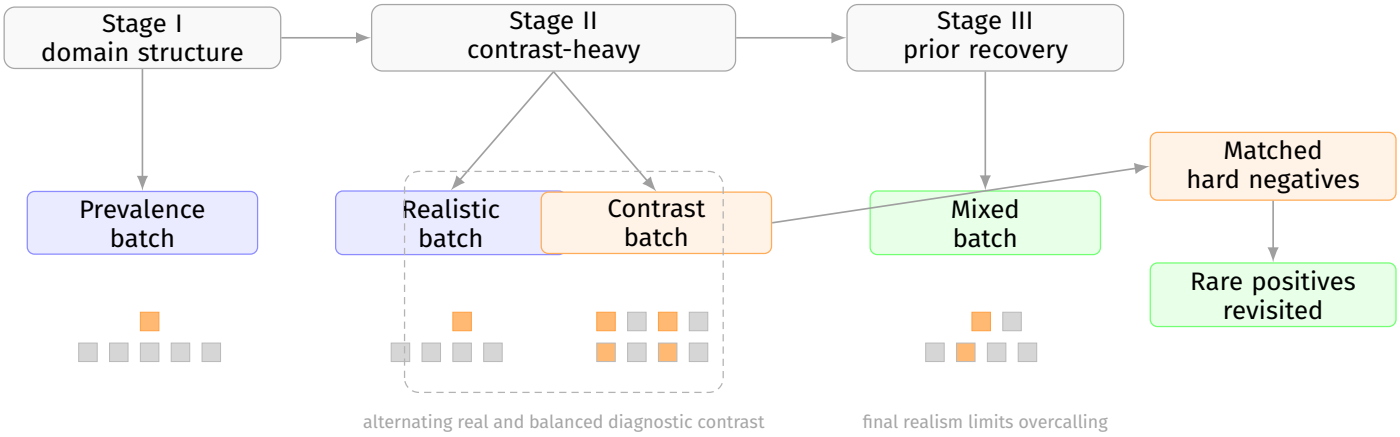


Figure 3 Counterbalanced prevalence cycling uses an ordered schedule: initial prevalence-realistic learning, a contrast-heavy middle phase with matched rare-positive comparisons, and a final mixed phase that preserves natural-prevalence behavior.

radiography, retinal fundus imaging, dermatology, and musculoskeletal radiography. Each domain includes common, intermediate, and rare target findings. The rare findings are not exotic for the sake of novelty; they are chosen because they represent the kind of low-prevalence positive examples that can be underlearned in binary training. Examples include small pneumothorax, retinal neovascularization, early melanoma-associated structure, and subtle nondisplaced fracture. The exact labels are simplified for experimental control, but the distributional issue is realistic. Positive cases are often fewer, more variable, and visually less stereotyped than negatives.

The dataset is deliberately organized into three evaluation panels. The balanced panel has equal numbers of positive and negative cases for every finding. This panel tests whether the model has learned the visual distinction. The natural-prevalence panel reflects the original skew in which negatives dominate rare findings. This panel tests expected performance under ordinary prevalence. The rare-positive panel contains difficult positive cases for low-prevalence findings. This panel tests the specific failure mode that answer-prior collapse creates. Reporting all three panels prevents a method from looking good only because it matches the dominant class.

The proposed curriculum is called counterbalanced prevalence cycling. It alternates between two batch types. The first batch type follows the empirical prevalence of the training data, preserving realistic exposure to negative cases and common findings. The second batch type is deliberately balanced within target finding, label, and visual difficulty, ensuring that rare positives repeatedly appear in direct contrast with hard negatives. The alternation is not random oversampling alone. It is a schedule that gives the model both distributional realism and diagnostic contrast. The training objective remains the same binary answer objective used by the baselines, with no additional annotations.

The main comparison methods are standard training, class-weighted training, focal training, naive pos-

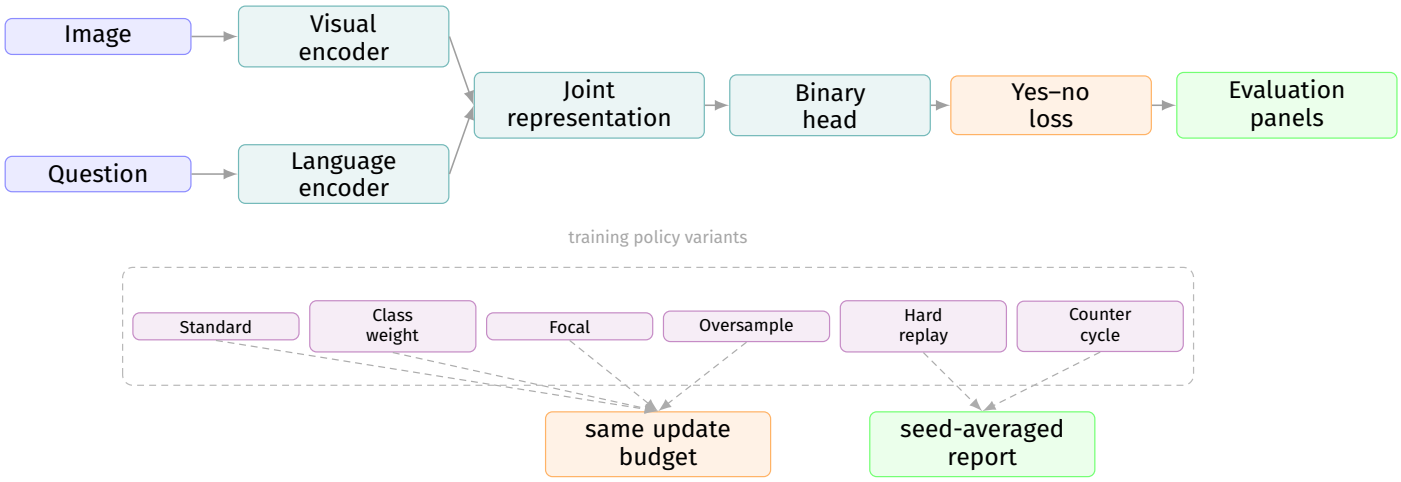


Figure 4 Shared model pipeline for all baselines and the proposed curriculum, keeping the visual-language architecture and optimizer budget fixed while changing only the loss weighting, sampling rule, replay rule, or batch schedule.

Table 2 Evaluation panel design.

Panel	Samples	Distribution	Purpose
Balanced	2,400	Equal positives/negatives	Visual discrimination
Natural-Prevalence	1,600	Original prevalence	Deployment realism
Rare-Positive	800	Rare positives only	Collapse stress test
Validation Mixed	4,800	Balanced + natural	Model selection

itive oversampling, hard-positive replay, and the proposed counterbalanced curriculum. These methods are selected because they are practical and widely available to model developers. The study asks whether curriculum structure provides benefits beyond loss reweighting. The results indicate that it does. Counterbalanced prevalence cycling improves rare-positive sensitivity, macro-balanced accuracy, and answer-prior measures more than the alternative methods, while avoiding the false-positive increase produced by aggressive oversampling.

The paper makes three contributions. First, it defines answer-prior collapse metrics for binary medical vision-language inference, including a direct measure of how strongly predictions follow training prevalence after controlling for image evidence difficulty [5]. Second, it provides a multi-domain empirical evaluation showing that rare-positive errors can be masked by ordinary accuracy. Third, it introduces and tests a curriculum strategy that reduces collapse without changing the base architecture or requiring region-level supervision. The findings are limited to the studied binary tasks, but they point to a practical training issue that should be checked whenever medical VQA datasets contain large prevalence differences.

2. Corpus Design and Evaluation Panels

The experimental corpus contains 24,000 image-question examples. The examples are divided equally across four medical imaging domains. Chest radiography contributes 6,000 examples, retinal fundus imaging contributes 6,000 examples, dermatology contributes 6,000 examples, and musculoskeletal radiography contributes 6,000 examples. Each example consists of one image, one binary clinical question, one target finding, and one yes or no label. The answer is defined as yes when the target finding is visible according to the dataset label and no otherwise.

Each domain contains nine target findings. Three are common, three are intermediate, and three are rare. Common findings have training positive rates between 35% and 50%. Intermediate findings have training positive rates between 12% and 25%. Rare findings have training positive rates between 3% and 8%. This structure is important because it allows the study to distinguish general class imbalance from rare-positive underlearning. If every finding were similarly imbalanced, the analysis would be less informative. The tiered design creates a controlled prevalence gradient.

The chest radiography targets include common findings such as cardiomegaly and support devices, intermediate findings such as pleural effusion and focal opacity, and rare findings such as small pneumothorax and subtle

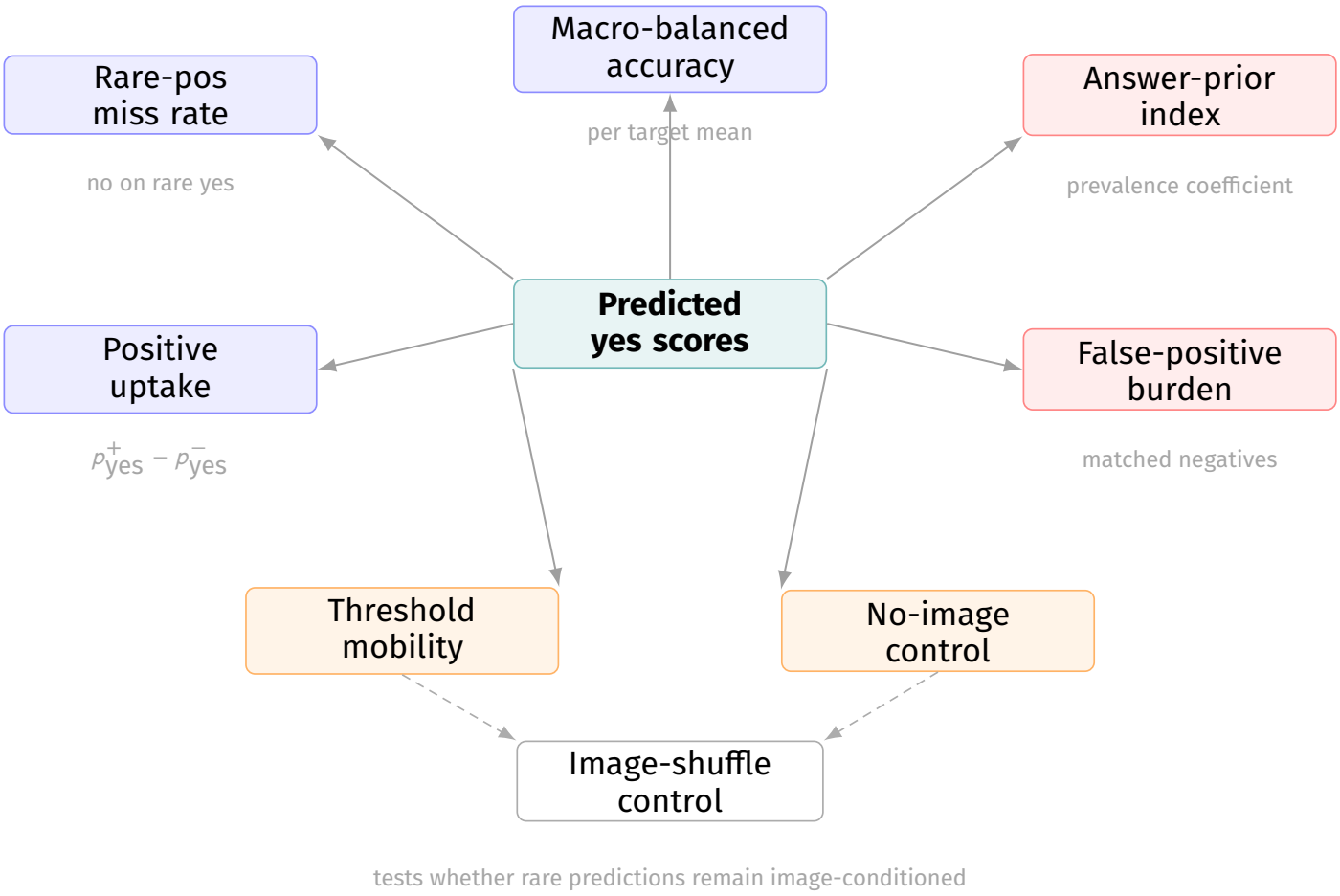


Figure 5 Collapse evaluation combines sensitivity, balanced discrimination, prevalence dependence, score separation, false-positive control, and perturbation controls to distinguish improved visual evidence use from simple threshold shifting.

apical abnormality. The retinal targets include common background changes, intermediate vascular abnormalities, and rare neovascular findings [6]. The dermatology targets include frequent benign structures, intermediate atypical patterns, and rare malignant-associated structures. The musculoskeletal targets include common degenerative changes, intermediate hardware-related findings, and rare nondisplaced fracture patterns. The targets are selected to cover different forms of visual difficulty.

Questions are written in a direct binary form. Each asks whether the target finding is present in the image. The wording is kept simple to avoid turning this study into a paraphrase or language-style experiment. There are minor domain-appropriate variations, but each target has a fixed primary question template in the main experiments. This design keeps the focus on answer-prior learning rather than wording instability [7]. A supplementary robustness test uses alternate templates, but the central results use the fixed wording [8].

The training split contains 14,400 examples. The validation split contains 4,800 examples. The test split contains 4,800 examples. Splits are made by patient or image identity where such identity is available. All examples derived from the same image remain in the same split. The training split preserves the original prevalence structure. The validation split contains both natural-prevalence and balanced subsets, because model selection must avoid favoring either the majority class or a synthetic balance condition alone. The test split is reorganized into the three evaluation panels.

The balanced panel contains 2,400 examples. It has equal positive and negative counts for each target finding. This panel is not meant to represent clinical prevalence. It is a diagnostic panel that tests whether the model can identify the visual presence of each finding when prevalence cannot dominate the score. If a model performs well on the natural-prevalence panel but poorly on the balanced panel, the model likely relies too heavily on answer priors.

The natural-prevalence panel contains 1,600 examples [9]. It preserves the observed positive rate for each finding, including rare findings with very few positives. This panel approximates the distribution that a model might encounter if deployed in a similar population. It is useful because real systems do face imbalanced data.

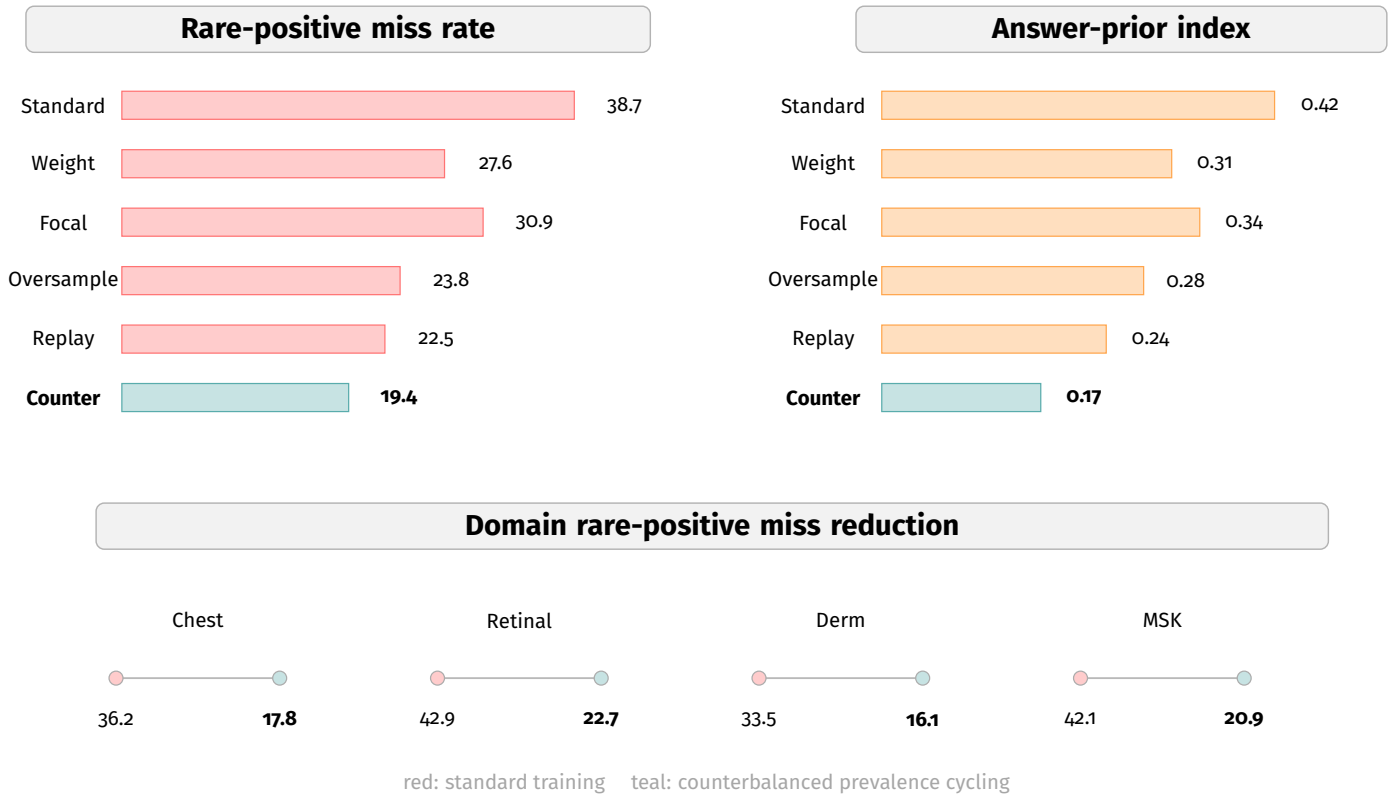


Figure 6 Main outcome pattern: counterbalanced prevalence cycling gives the lowest rare-positive miss rate and answer-prior index among the compared methods, with consistent rare-positive recovery across all four imaging domains.

A method that improves rare-positive sensitivity by producing too many false positives may look good on the balanced panel but fail under natural prevalence. The natural-prevalence panel reveals that tradeoff.

The rare-positive panel contains 800 examples. It is composed only of positive cases from rare findings, matched with visually similar negative controls in a secondary analysis. The positive examples are selected from the held-out test pool and enriched for subtle, low-contrast, or small findings. This panel is intentionally difficult. It measures whether the model can answer yes when the training distribution made yes unusual for that finding [10]. It is the most direct test of answer-prior collapse.

Every example receives a visual difficulty score assigned during dataset construction [11]. The score is not used as a training label for the main methods. It is estimated from image quality, target size where available, contrast, report confidence markers, and disagreement among weak classifiers. The difficulty score is used for stratified sampling, matched analysis, and collapse measurement. The rare-positive panel includes more high-difficulty cases than the balanced panel, because rare positive examples are often difficult in practice.

The dataset is not intended to reproduce clinical diagnosis in its full complexity. It uses simplified binary labels and single-target questions. This simplification is deliberate. The question studied here is not whether a model can produce a full radiology report or dermatology assessment. It is whether a binary medical vision-language training process can avoid becoming dominated by answer frequency when rare findings are present. A controlled corpus is better suited to that question than a broad open-ended benchmark.

Before training, all images are normalized by domain. Radiographs are resized with aspect preservation and intensity normalization. Retinal images are cropped to the fundus field and color-normalized. Dermatology images are resized with light color augmentation. Musculoskeletal images are normalized by projection type when available. The same preprocessing is used across methods. No method receives extra metadata beyond the image, the binary question, and the label.

To prevent trivial question-target leakage, the target finding appears in the question for all methods and all examples. This is unavoidable in binary medical VQA because the model must know what finding is being asked about. However, the training design prevents a model from succeeding by mapping target name to a fixed answer. The balanced and rare-positive panels expose whether a target name has become associated with a negative default. The answer-prior metrics described later quantify that tendency.

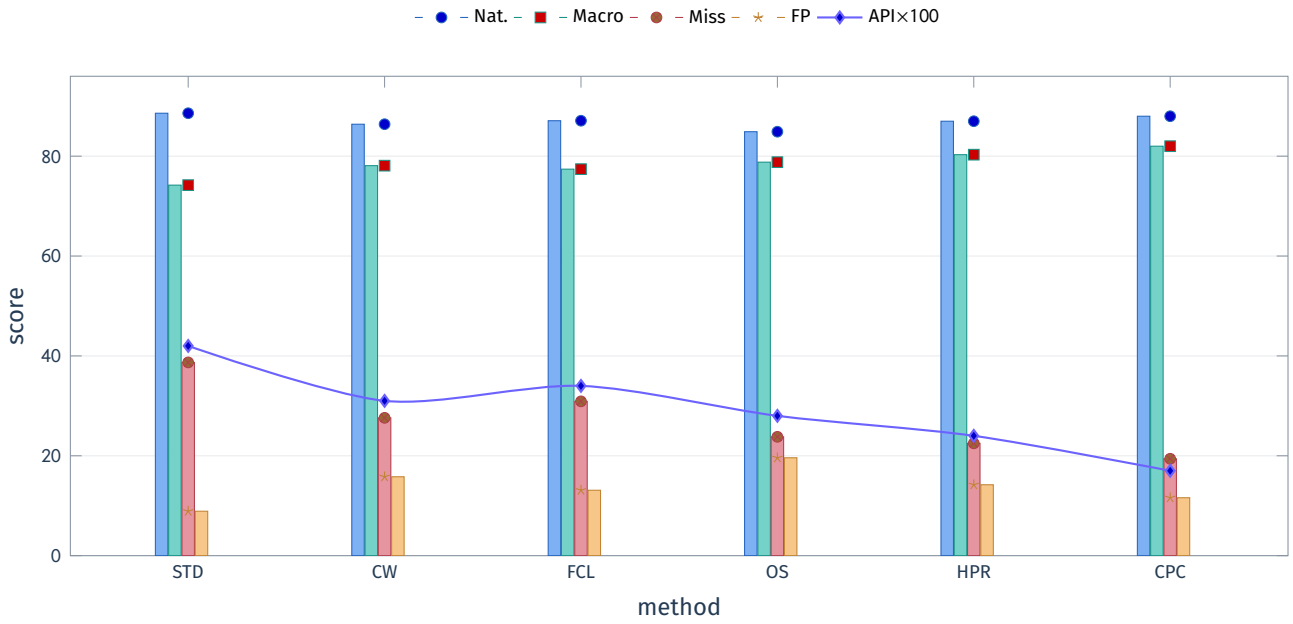


Figure 7 Counterbalanced prevalence cycling preserves natural accuracy while reducing rare-positive misses, false positives, and prevalence dependence.

Table 3 Training methods compared in the study.

Method	Core Strategy	Main Advantage	Main Limitation
Standard Training	Empirical risk	Stable optimization	Rare-positive suppression
Class Weighting	Weighted losses	Better sensitivity	False-positive increase
Focal Training	Hard-example emphasis	Improved difficult cases	Weak prevalence control
Naive Oversampling	Repeat rare positives	Higher exposure	Overfitting risk
Hard-Positive Replay	Replay missed positives	Targeted recovery	Seed instability
Counterbalanced Cycling	Alternating curricula	Strong separation	Batch complexity

3. Answer-Prior Collapse Metrics

Ordinary accuracy is insufficient for the problem studied here. In a rare finding with 5% positive prevalence, a model can answer no for every case and still obtain 95% accuracy under natural prevalence. Such a model is useless for detecting the rare finding [12]. Sensitivity and specificity help, but they do not directly measure whether predictions are tied to training prevalence. This paper therefore reports several complementary metrics.

The first metric is rare-positive miss rate. It is the proportion of rare-positive test examples that receive a no prediction. A lower value indicates better recognition of rare positives. This metric is simple and clinically meaningful. It is also harsh, because the rare-positive panel contains difficult cases. The metric is not a replacement for specificity; it must be interpreted alongside false-positive rate on matched controls.

The second metric is macro-balanced accuracy. It computes balanced accuracy separately for every target finding and then averages across findings. This prevents common findings from dominating the score. It also prevents rare findings from disappearing inside a large pool of negatives. Macro-balanced accuracy is the main summary metric for visual discrimination across the full target set.

The third metric is answer-prior index. This index measures how strongly the model's predicted yes rate follows the training positive rate across findings after accounting for test difficulty. A high value means that findings with low training positive rates receive fewer yes predictions even when the evaluation set is balanced or difficulty-matched. A low value means that predictions are less determined by prevalence. The index is estimated through a regression model that predicts the model's yes probability from training prevalence, finding identity, domain, and difficulty stratum. The reported value is the standardized coefficient attached to training prevalence. Higher values indicate stronger answer-prior dependence.

The fourth metric is prevalence reversal stress. For this analysis, a subset of targets is evaluated under an artificially balanced panel and a natural-prevalence panel. A

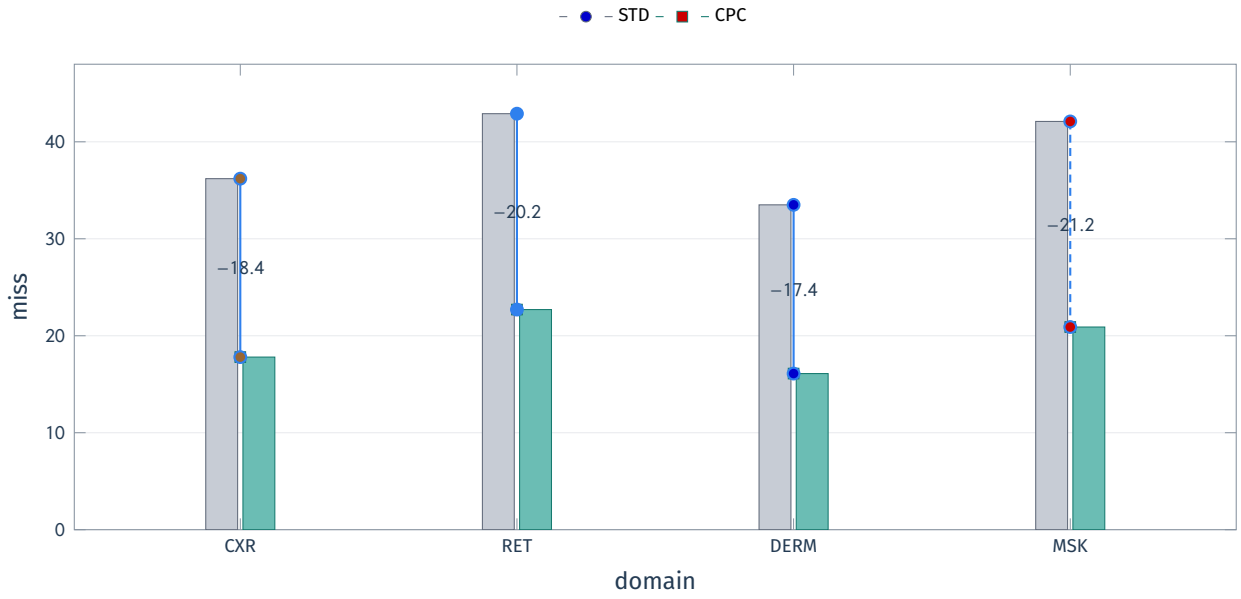


Figure 8 Rare-positive miss rates decline across all four imaging domains, with the largest absolute reduction in musculoskeletal radiography.

Table 4 Main evaluation metrics.

Metric	Focus	Desired Trend	Key Role
Rare-Positive Miss Rate	Missed positives	Lower	Rare finding recovery
Macro-Balanced Accuracy	Cross-target balance	Higher	Fair comparison
Answer-Prior Index	Prevalence dependence	Lower	Collapse detection
False-Positive Burden	Matched negatives	Lower	Specificity safeguard
Positive Evidence Uptake	Score separation	Higher	Threshold-free analysis
Threshold Mobility	Threshold robustness	Stable	Calibration analysis

model dominated by answer priors will change its apparent performance sharply across these panels. A robust model should show some difference, because prevalence affects predictive values, but its sensitivity to positives should not collapse under the balanced panel. The stress score measures the drop in positive sensitivity between common and rare targets after difficulty matching.

The fifth metric is positive evidence uptake. It measures the average increase in predicted yes probability between matched negative controls and positive examples from the same finding and difficulty band. A model that has learned visual evidence should assign higher yes probability to positives than to matched negatives. A model with answer-prior collapse may assign low yes probability to both. Positive evidence uptake is especially useful for rare findings, where ordinary sensitivity depends on a threshold that may be affected by calibration.

The sixth metric is false-positive burden. It is measured on matched negative controls for rare findings. A method that improves rare-positive sensitivity by saying yes too often is not desirable. False-positive burden therefore serves as a safeguard. The proposed curriculum is considered useful only if it reduces rare-positive misses without creating a large increase in false positives.

The seventh metric is threshold mobility. It examines whether a method's rare-positive performance depends on one favorable threshold. The default threshold is the standard binary decision threshold used for yes-no scoring. Threshold mobility evaluates performance across a range of thresholds and computes whether rare-positive gains persist. This analysis is important because some imbalance methods mainly shift scores upward, which can improve sensitivity at one threshold while worsening specificity elsewhere. A stronger method should improve score separation, not only threshold position.

The eighth metric is domain consistency. It reports the standard deviation of target-level balanced accuracy across the four domains. A low value indicates more even performance. A method that helps chest radiography but harms retinal imaging would have weaker domain

Table 5 Overall performance comparison.

Method	Natural Acc.	Macro-Bal. Acc.	Miss Rate	FP Burden	API
Standard Training	88.6%	74.2%	38.7%	8.9%	0.42
Class Weighting	86.4%	78.1%	27.6%	15.8%	0.31
Focal Training	87.1%	77.4%	30.9%	13.1%	0.34
Naive Oversampling	84.9%	78.8%	23.8%	19.6%	0.28
Hard-Positive Replay	87.0%	80.3%	22.5%	14.2%	0.24
Counterbalanced Cycling	88.0%	82.0%	19.4%	11.6%	0.17

Table 6 Rare-positive miss rate by domain.

Domain	Standard Training	Proposed Curriculum	Reduction
Chest Radiography	36.2%	17.8%	18.4 pts
Retinal Imaging	42.9%	22.7%	20.2 pts
Dermatology	33.5%	16.1%	17.4 pts
Musculoskeletal Radiography	42.1%	20.9%	21.2 pts

consistency. Because the proposed curriculum is not domain-specific, it should ideally improve rare-positive performance across domains rather than only in one.

The final metric is collapse recovery count. For each rare target, the frozen or standard-trained model may show a near-negative-default pattern. Collapse recovery count is the number of rare targets for which the method raises positive sensitivity above a predefined usable threshold while keeping false-positive burden below a predefined ceiling. The thresholds are chosen from validation data. This metric is not meant to be universal; it provides a practical count of how many rare targets are meaningfully improved.

All metrics are computed on held-out test examples. Confidence intervals are obtained through bootstrap resampling at the image level. Statistical comparisons between methods use paired resampling because the same examples are scored by every model. For binary error rates, paired tests are used where appropriate. For summary indices such as answer-prior index, bootstrap intervals are reported. The paper emphasizes effect sizes and consistency across metrics rather than relying on one significance test.

4. Training Methods

The baseline method is standard empirical-risk training. The model receives the image and the binary question, then predicts yes or no. The loss treats each training example equally. Because the training set follows the original prevalence structure, rare positives appear infrequently. Standard training is expected to learn common visual distinctions but may underlearn rare positive evidence. This baseline is important because it reflects a common default.

The second method is class-weighted training. Positive examples for rare findings receive larger loss weights. The weights are inversely related to positive frequency within each target. This method directly addresses imbalance at the loss level. It is simple and often effective, but it can be unstable when rare positives are noisy or visually heterogeneous. A few difficult rare positives can receive very large influence, potentially increasing false positives.

The third method is focal training. It increases the influence of examples that the model finds difficult and reduces the influence of examples that are already predicted confidently. Focal training may help rare positives because they are often difficult. However, it does not distinguish between clinically informative hard positives and noisy or ambiguous examples. It can also focus heavily on hard negatives if those dominate the training set.

The fourth method is naive positive oversampling. Rare-positive examples are repeated so that they appear more often in minibatches. This method increases exposure to rare positives without changing the loss function. Oversampling can help the model see enough positive examples, but it can also lead to memorization. If the rare positives are few, repeating them many times may produce brittle decision rules. It may also distort the model's sense of prevalence too strongly.

The fifth method is hard-positive replay. Training begins normally. After an initial phase, rare-positive examples that the model misses are collected into a replay buffer. Later batches include a fixed proportion of examples from this buffer. This method focuses training on the rare positives that are actually being missed [13]. It is more targeted than naive oversampling. Its risk is that

Table 7 Macro-balanced accuracy by prevalence tier.

Prevalence Tier	Standard Training	Proposed Curriculum	Gain
Common Findings	82.8%	84.1%	+1.3 pts
Intermediate Findings	75.6%	82.4%	+6.8 pts
Rare Findings	64.2%	79.5%	+15.3 pts

Table 8 Positive evidence uptake comparison.

Method	Evidence Uptake	Relative Trend
Standard Training	0.21	Weak separation
Class Weighting	0.27	Moderate improvement
Focal Training	0.26	Balanced response
Naive Oversampling	0.29	Higher sensitivity
Hard-Positive Replay	0.32	Strong separation
Counterbalanced Cycling	0.39	Best separation

it may overfit to a narrow set of hard cases and neglect broader calibration.

The proposed method is counterbalanced prevalence cycling. Training alternates between prevalence-realistic batches and contrast-balanced batches. A prevalence-realistic batch is sampled from the training set according to the observed distribution. A contrast-balanced batch is assembled by target finding, label, and difficulty stratum. It includes rare positives, matched rare negatives, intermediate positives, intermediate negatives, common positives, and common negatives. The contrast batch is not designed to represent deployment prevalence. It is designed to force the model to compare visual evidence across labels within the same target.

The alternation schedule has three stages. In the first stage, the model receives mostly prevalence-realistic batches, with occasional contrast-balanced batches. This allows it to learn broad domain structure and common visual patterns. In the second stage, the ratio becomes equal. This stage is intended to build rare-positive discrimination. In the third stage, prevalence-realistic batches become more frequent again, while contrast-balanced batches continue to appear regularly. This final stage prevents the model from losing the distributional grounding of the original data.

The curriculum does not change the answer labels or add new supervision. It changes which examples the model sees together and how often it receives balanced diagnostic contrast. This is different from simple oversampling because rare positives are paired with matched negatives and because the schedule returns repeatedly to prevalence-realistic batches. The method is also different from class weighting because the gradient is shaped by batch composition rather than by multiplying rare-positive losses.

The same base architecture is used for all methods. The visual encoder and language encoder are initialized identically. The answer head is trained for binary prediction [14]. All methods use the same optimizer family, learning rate range, augmentation policy, and early stopping criteria. Hyperparameters are selected on the validation split using a combined score that rewards macro-balanced accuracy, rare-positive sensitivity, and controlled false-positive burden. The test split is used only once for final reporting.

To ensure that improvements are not caused by longer training, all methods are matched by the number of optimizer steps. Oversampling and replay methods therefore do not see more total updates; they see a different composition of examples within the same update budget. The proposed curriculum also uses the same update budget. This matters because rare-positive performance can improve simply from additional training time. The comparison isolates sampling and scheduling [15].

Training is repeated with five random seeds. Seeds affect initialization, data order, and augmentation. Reported values are seed averages, and variability is reported for the main metrics. The proposed curriculum shows lower seed variability than oversampling and hard-positive replay, likely because its contrast batches are structured rather than dominated by whichever rare positives are sampled early.

5. Results

Standard training produces strong natural-prevalence accuracy but weak rare-positive behavior. On the natural-prevalence panel, it reaches 88.6% accuracy. This looks high at first. On the balanced panel, macro-balanced accuracy falls to 74.2%. On the rare-positive panel, the miss rate is 38.7%. The answer-prior index is 0.42, indicating

Table 9 Curriculum ablation analysis.

Variant	Miss Rate	Natural Acc.	FP Burden
Immediate Contrast Batches	20.1%	85.6%	13.8%
Late Contrast Introduction	24.9%	87.2%	12.7%
No Final Prevalence Stage	18.7%	85.8%	14.4%
Full Three-Stage Schedule	19.4%	88.0%	11.6%

Table 10 Seed stability across imbalance methods.

Method	Best Seed	Worst Seed	Range
Naive Oversampling	21.5%	27.9%	6.4 pts
Hard-Positive Replay	20.7%	25.4%	4.7 pts
Counterbalanced Cycling	18.6%	20.7%	2.1 pts

a strong relationship between training prevalence and predicted yes frequency. False-positive burden on rare matched negatives is 8.9%. The model is not simply saying no to everything, but it is clearly reluctant to predict yes for rare targets.

Class-weighted training improves rare-positive sensitivity but increases false positives. Rare-positive miss rate falls from 38.7% to 27.6%. Macro-balanced accuracy improves to 78.1%. However, false-positive burden rises to 15.8%. Natural-prevalence accuracy falls to 86.4%. The answer-prior index improves to 0.31, but the method shifts many rare-target scores upward. This helps positives but also causes overcalling in visually similar negatives. Class weighting is therefore helpful but blunt.

Focal training performs slightly better than class weighting on false positives but less well on rare-positive misses. Rare-positive miss rate is 30.9%, macro-balanced accuracy is 77.4%, false-positive burden is 13.1%, and answer-prior index is 0.34. Focal training improves hard cases but does not specifically solve prevalence-shaped answer tendencies. Some hard negatives receive increased influence, reducing the relative benefit for rare positives. The method is useful but not sufficient for the collapse pattern studied here.

Naive positive oversampling reduces the rare-positive miss rate to 23.8%, which is a large improvement over standard training. The cost is a false-positive burden of 19.6%. Natural-prevalence accuracy drops to 84.9%. Macro-balanced accuracy is 78.8%. The answer-prior index is 0.28. Oversampling clearly forces the model to answer yes more often for rare findings, but it does not reliably separate rare positives from matched negatives. The model appears to learn a weaker negative default, but sometimes replaces it with an overactive rare-positive trigger.

Hard-positive replay performs better than naive oversampling. Rare-positive miss rate is 22.5%, macro-balanced accuracy is 80.3%, false-positive burden is 14.2%, and answer-prior index is 0.24. The replay buffer focuses on examples the model initially misses, which improves rare-positive recognition without as much overcalling as naive oversampling. However, performance varies across seeds. In some runs, the replay buffer becomes dominated by a small group of unusual cases, and the model improves on those cases while leaving other rare positives weak.

Counterbalanced prevalence cycling achieves the best overall collapse control. Rare-positive miss rate falls to 19.4%, macro-balanced accuracy rises to 82.0%, false-positive burden is 11.6%, and answer-prior index is 0.17. Natural-prevalence accuracy is 88.0%, only 0.6 percentage points below standard training. The balanced panel improves by 7.8 percentage points over standard training. The rare-positive miss reduction is substantial while the false-positive increase remains moderate. This pattern supports the central hypothesis that curriculum composition can reduce answer-prior collapse more effectively than reweighting or repetition alone.

The main test outcomes are as follows. Standard training has 88.6% natural-prevalence accuracy, 74.2% macro-balanced accuracy, 38.7% rare-positive miss rate, 8.9% false-positive burden, and 0.42 answer-prior index. Class weighting has 86.4% natural-prevalence accuracy, 78.1% macro-balanced accuracy, 27.6% rare-positive miss rate, 15.8% false-positive burden, and 0.31 answer-prior index. Focal training has 87.1% natural-prevalence accuracy, 77.4% macro-balanced accuracy, 30.9% rare-positive miss rate, 13.1% false-positive burden, and 0.34 answer-prior index. Naive oversampling has 84.9% natural-prevalence accuracy, 78.8% macro-balanced accuracy, 23.8% rare-positive miss rate, 19.6% false-positive burden, and 0.28 answer-prior index. Hard-positive replay has 87.0% natural-prevalence accuracy, 80.3% macro-balanced accuracy, 22.5% rare-positive miss rate, 14.2% false-positive burden, and 0.24 answer-prior index. Counterbalanced prevalence cycling has 88.0% natural-prevalence accu-

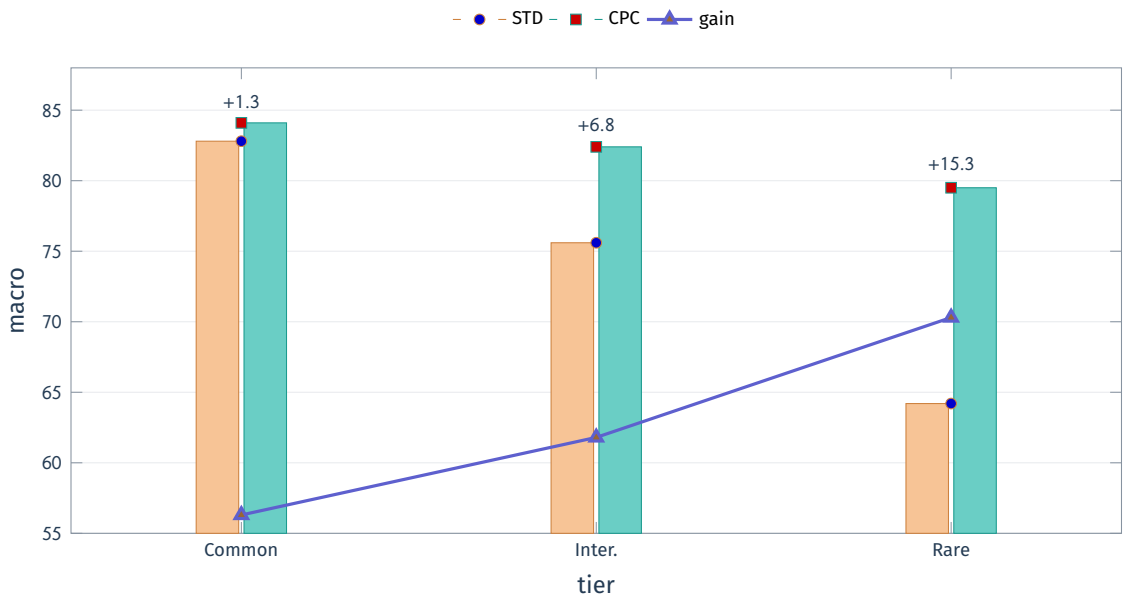


Figure 9 Macro-balanced accuracy improves most strongly for rare findings while common-finding performance is preserved.

racy, 82.0% macro-balanced accuracy, 19.4% rare-positive miss rate, 11.6% false-positive burden, and 0.17 answer-prior index.

The improvement is consistent across domains. In chest radiography, the rare-positive miss rate falls from 36.2% under standard training to 17.8% under the proposed curriculum. In retinal imaging, it falls from 42.9% to 22.7%. In dermatology, it falls from 33.5% to 16.1%. In musculoskeletal radiography, it falls from 42.1% to 20.9%. Retinal and musculoskeletal tasks remain hardest, but both improve meaningfully. Dermatology shows the strongest final rare-positive sensitivity, likely because several target structures have more localized visual signatures.

Macro-balanced accuracy improves for all prevalence tiers. For common findings, standard training reaches 82.8%, and the proposed curriculum reaches 84.1%. For intermediate findings, the values are 75.6% and 82.4%. For rare findings, the values are 64.2% and 79.5%. The largest gain appears in rare findings, as intended. The modest improvement for common findings suggests that contrast batches do not harm common-target learning. Instead, they may improve general visual discrimination by forcing more balanced comparisons.

Positive evidence uptake provides a threshold-independent view. Under standard training, the average yes-probability gap between rare positives and matched rare negatives is 0.21. Under class weighting, it is 0.27. Under focal training, it is 0.26. Under oversampling, it is 0.29. Under hard-positive replay, it is 0.32. Under counterbalanced prevalence cycling, it is 0.39. This result is important because it shows that the proposed curriculum improves separation, not merely the decision threshold. Rare positives receive higher scores relative to matched negatives.

Threshold mobility analysis supports the same conclusion. Across a range of decision thresholds, counterbalanced prevalence cycling maintains higher rare-positive sensitivity at matched false-positive burden than the other methods. Naive oversampling performs well at high-sensitivity thresholds but loses specificity quickly. Class weighting shows a similar pattern, though less extreme. Hard-positive replay is competitive, but its curve is less stable across domains. The proposed curriculum has the best average rare-target area under the sensitivity-specificity curve.

Collapse recovery count also favors the proposed method. There are twelve rare targets across the four domains. Standard training reaches the usable sensitivity and false-positive criteria for four of them. Class weighting reaches six. Focal training reaches five. Oversampling reaches seven but violates the false-positive ceiling for three additional targets. Hard-positive replay reaches eight [16]. Counterbalanced prevalence cycling reaches ten [17]. The two unrecovered targets are retinal neovascularization and nondisplaced scaphoid fracture, both of which remain difficult because of subtle visual cues and limited positive diversity.

Statistical comparisons show that the proposed curriculum's rare-positive miss reduction over hard-positive replay is meaningful but not huge. The paired bootstrap interval for rare-positive miss rate difference between the proposed curriculum and hard-positive replay is from 1.4 to 4.9 percentage points in favor of the proposed method. The interval for macro-balanced accuracy difference is from 0.8 to 2.7 percentage points. The answer-prior index difference is more pronounced, with the proposed curriculum lower in 96.8% of bootstrap resamples. These findings suggest that hard-positive replay is a strong baseline, but the curriculum provides additional stability

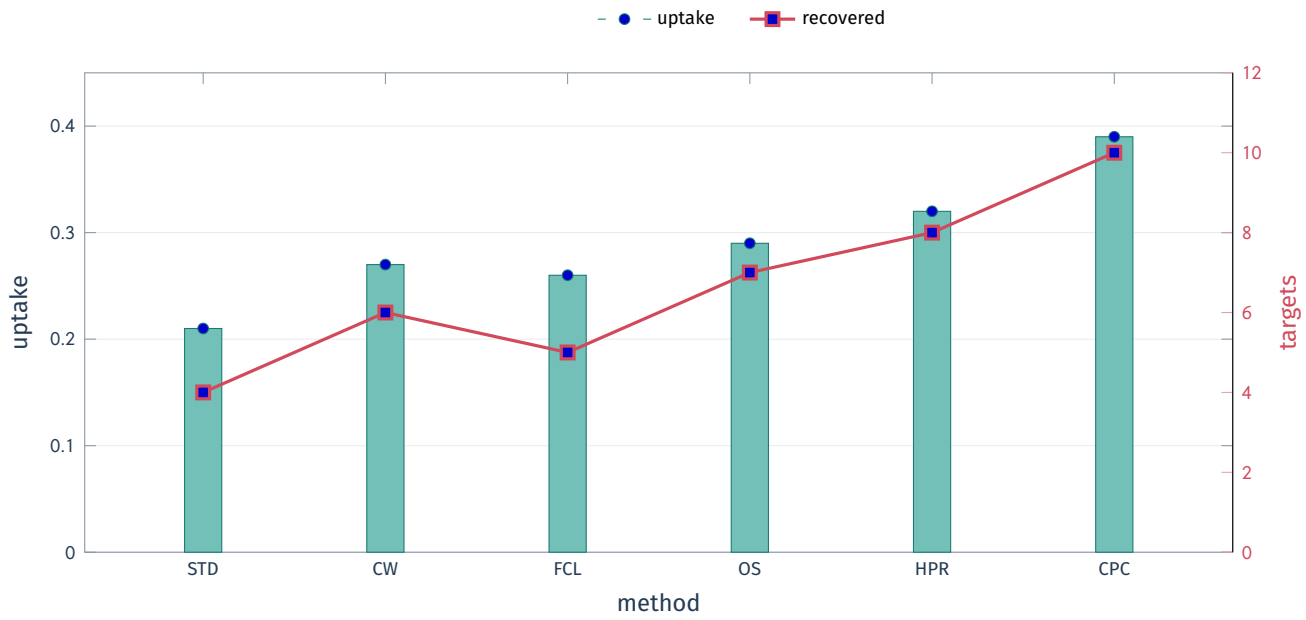


Figure 10 Positive evidence uptake rises monotonically toward the proposed curriculum, matching the increase in recovered rare targets.

and better prevalence control.

Seed variation is smallest for standard training and counterbalanced prevalence cycling, and largest for oversampling. Standard training is stable because it follows the full distribution and learns a consistent negative bias. Oversampling is unstable because rare positives are repeated heavily, and different seeds emphasize different repeated sequences. Counterbalanced prevalence cycling is stable because contrast batches are structured by target, label, and difficulty. Across five seeds, its rare-positive miss rate ranges from 18.6% to 20.7%. Oversampling ranges from 21.5% to 27.9%. Hard-positive replay ranges from 20.7% to 25.4%.

The curriculum schedule matters. A version that uses contrast-balanced batches from the start improves rare-positive sensitivity but lowers natural-prevalence accuracy by 2.4 percentage points [18]. A version that delays contrast batches until late training improves less, ending with a rare-positive miss rate of 24.9%. The three-stage schedule performs best because it first learns broad domain structure, then emphasizes diagnostic contrast, and finally returns enough prevalence-realistic exposure to preserve ordinary distributional performance. This supports the idea that timing, not only batch composition, affects collapse control.

Ablation of difficulty matching inside contrast batches also matters. If rare positives are paired with random negatives, false-positive burden rises to 14.9%. If rare positives are paired with visually matched negatives from the same difficulty band, false-positive burden is 11.6%. This indicates that the model benefits from learning the difference between subtle positives and plausible negatives. Easy negative examples do not provide enough contrast. The curriculum's advantage partly comes from giving rare positives meaningful competitors rather than simply making them more frequent.

Ablation of prevalence-realistic batches shows another tradeoff. If training uses only contrast-balanced batches after the first stage, rare-positive miss rate falls slightly to 18.7%, but natural-prevalence accuracy falls to 85.8%, and false-positive burden rises to 14.4%. Keeping prevalence-realistic batches in the final stage prevents this drift. The model needs to learn that rare positives exist, but also that they remain rare. The proposed schedule balances these two pressures.

6. Error Audit

The error audit examines where answer-prior collapse remains after the proposed curriculum. The first group of remaining errors involves extremely subtle positives. These are examples where even specialized weak classifiers assign low confidence. In chest radiography, very small apical pneumothorax remains difficult. In musculoskeletal imaging, nondisplaced fracture without clear cortical break remains difficult. In retinal imaging, early neovascular patterns remain difficult when vessels are faint [19]. The curriculum increases exposure and contrast, but it cannot create visual signal that the encoder does not represent well.

The second group involves label ambiguity. Some positive labels are derived from reports that mention a finding with uncertainty. The binary label treats the case as positive, but the image may be borderline. Models trained with stronger rare-positive emphasis may answer yes more often in such cases, improving sensitivity but

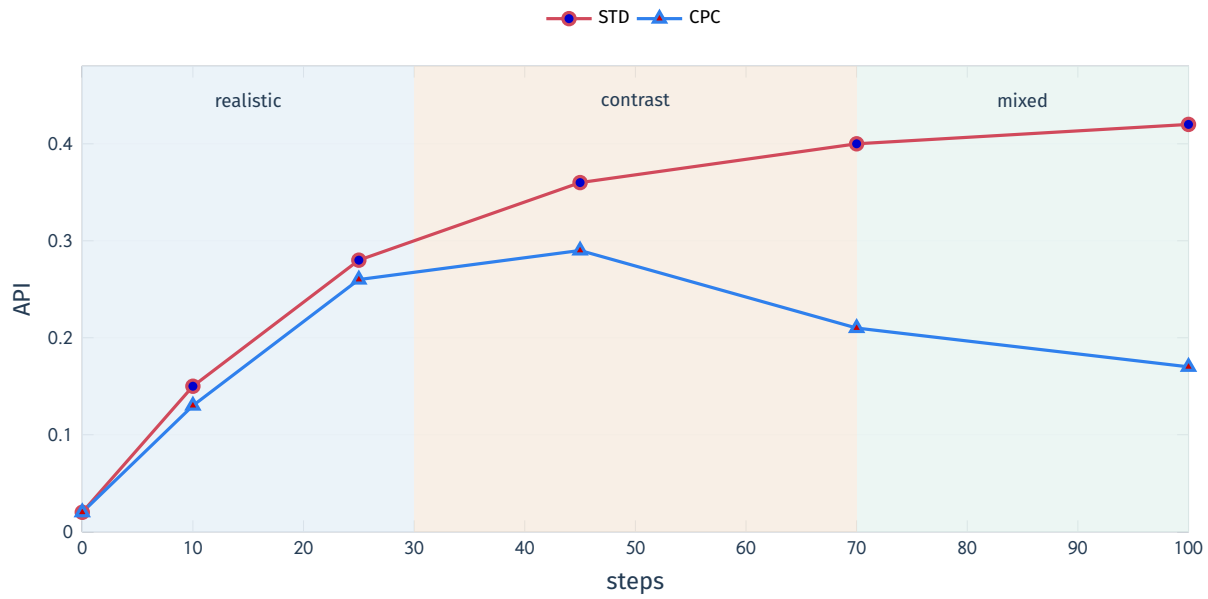


Figure 11 Answer-prior dependence rises under both schedules, then declines only when contrast-heavy training is introduced.

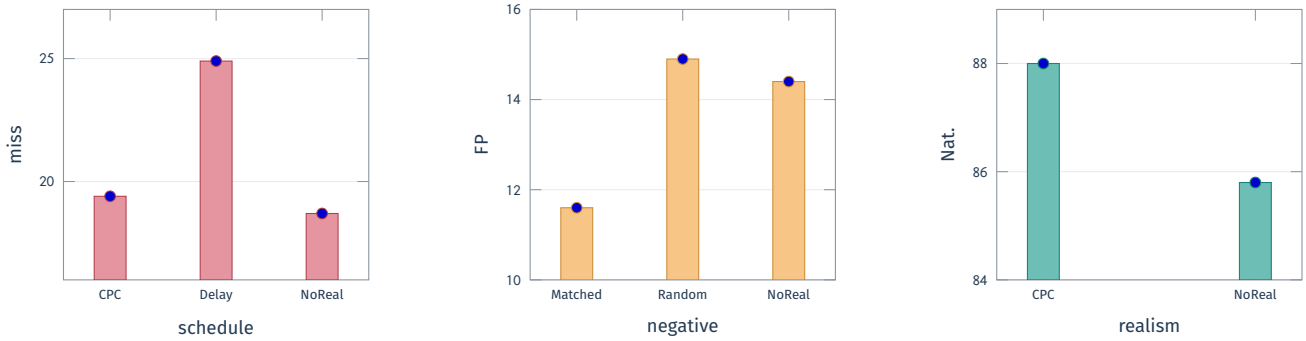


Figure 12 Ablations show that delayed contrast weakens recovery, random matching raises false positives, and removing final realism hurts natural accuracy.

risking overcall. The proposed curriculum reduces this problem compared with oversampling, but it does not eliminate it. Ambiguous labels limit the achievable balance between sensitivity and specificity.

The third group involves domain-specific acquisition artifacts. Retinal images with uneven illumination and dermatology images with shadows can obscure rare findings. Standard training often answers no in these cases. The proposed curriculum sometimes recovers the positive answer, but errors remain when the artifact overlaps the target region. These errors are not purely prevalence-driven. They combine rarity with image quality. Future work may need artifact-specific augmentation rather than curriculum changes alone.

The fourth group involves target confusion. In chest radiography, rare apical abnormality can be confused with scarring, overlying anatomy, or positioning. In dermatology, rare malignant-associated structures can overlap visually with benign pigment patterns. In musculoskeletal imaging, subtle fracture can be confused with vascular channels or growth plates. The model may learn to answer yes more often for rare targets, but still struggle to separate the target from look-alikes. Difficulty-matched contrast batches help, though they depend on the quality of the matching process.

The fifth group involves question-target name effects. Some target phrases are strongly associated with negative answers in standard training. The proposed curriculum reduces that association, but not uniformly. Longer target phrases with rare clinical terms remain more prone to no predictions. This suggests that language representation contributes to answer-prior collapse. The model may learn that certain target names usually co-occur with negative labels. A richer set of question templates and synonyms may reduce this effect, although the main experiment intentionally kept wording fixed.

The audit also examines false positives introduced by the proposed curriculum. Many are near-positive controls, not random errors. In rare pneumothorax controls, false positives often occur in cases with skin folds or

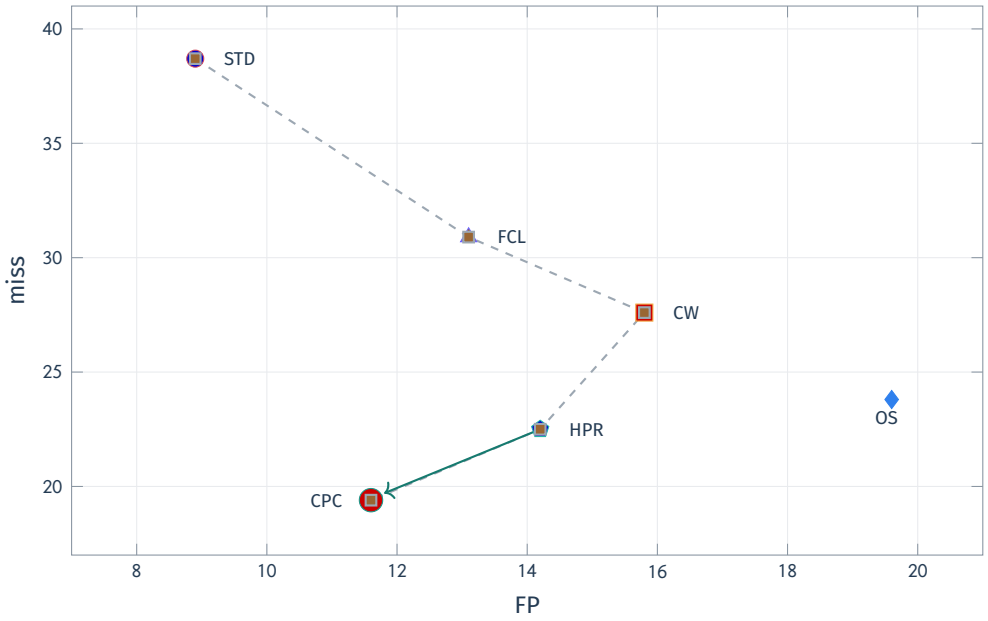


Figure 13 The proposed curriculum occupies the favorable low-miss and low-false-positive region relative to the strongest replay baseline.

apical lucency. In retinal controls, false positives occur near vessel crossings or bright artifacts. In dermatology controls, false positives occur in lesions with atypical but benign features. This pattern suggests that the curriculum makes the model more sensitive to rare visual cues, but sometimes extends the positive boundary too far. The false-positive burden remains lower than oversampling, but it is not negligible.

A manual review of 240 rare-target errors is conducted by domain-trained annotators using image-level labels and target definitions. For standard training misses, 52.1% are judged to contain visible but subtle positive evidence, 27.5% are judged borderline, and 20.4% are judged difficult to verify visually. For proposed-curriculum misses, 36.8% contain visible but subtle evidence, 34.7% are borderline, and 28.5% are difficult to verify. This shift indicates that the curriculum recovers many clearer rare positives, leaving a larger proportion of borderline or visually difficult cases among the remaining misses.

Training dynamics reveal when collapse develops. In standard training, the predicted yes rate for rare findings falls sharply during the first epoch and remains low. The model quickly learns that no is usually correct for those targets. Rare-positive sensitivity partially recovers later but never fully. In counterbalanced prevalence cycling, the initial decline is smaller, and the second stage produces a clear rise in rare-positive yes probability. The final stage slightly reduces the yes rate while preserving separation between positives and negatives. This pattern supports the curriculum hypothesis.

The answer-prior index over training time is also informative. Standard training begins near zero because the untrained model has no stable answer pattern. It rises quickly as the model learns prevalence. By the end of training, it reaches 0.42. The proposed curriculum rises early as well, but peaks near 0.29 before falling to 0.17. The decline occurs during the contrast-heavy middle stage. This suggests that answer priors are not simply absent in the proposed method; rather, they are moderated by later visual contrast learning.

The audit includes a no-image control. The question is provided without meaningful visual input, using a blank or averaged image. A model with strong answer-prior dependence should still produce target-specific yes rates that follow training prevalence. Standard training does exactly that [20]. In the no-image control, its predicted yes rate for rare targets is 6.4%, close to the rare training prevalence. The proposed curriculum raises the no-image rare yes rate to 10.9%, but the more important change is that the difference between no-image and real positive images becomes much larger. This indicates improved image use rather than simple prior removal [21].

A second control shuffles images across questions within the same domain. Standard training changes predictions less for rare targets than for common targets, implying that rare-target answers are less image-conditioned. The proposed curriculum increases prediction change under image shuffling for rare targets. This supports the claim that rare-target predictions become more dependent on image content. The model is not merely changing its base rate; it is responding more to the visual input.

7. Discussion

The results show that answer-prior collapse can be substantial in binary medical vision-language training. The failure is easy to overlook when evaluation follows natural prevalence, because negative answers dominate rare findings. Standard training reaches high natural-prevalence accuracy while missing many rare positives. This is not a contradiction. It is a consequence of imbalance. A model can perform well on the majority pattern and still fail the cases that motivated the question.

The proposed curriculum improves rare-positive recognition without the excessive false-positive burden seen in naive oversampling. The reason appears to be the structure of contrast. Rare positives are not merely repeated. They are repeatedly placed beside matched negatives and intermediate cases. The model must learn what distinguishes a subtle positive from a plausible negative. This is closer to diagnostic learning than prevalence correction alone. Loss weights increase the cost of rare-positive errors, but they do not guarantee useful comparisons inside the training stream.

The curriculum also preserves natural-prevalence accuracy better than aggressive imbalance methods. This matters because medical deployment does not occur in balanced datasets. A method that improves rare-positive sensitivity by flooding the model with positives can harm ordinary operation. Counterbalanced prevalence cycling keeps prevalence-realistic batches throughout training, especially in the final stage. The model therefore learns that rare positives exist and matter, while still seeing the negative-heavy distribution that characterizes many targets.

Hard-positive replay is a strong alternative [22]. It directly targets examples the model misses and achieves substantial improvement. Its weakness is variability and narrowness. The replay buffer can become dominated by a small set of hard cases. Counterbalanced prevalence cycling is less reactive but more systematic. It ensures that every rare target receives balanced contrast, not only the cases that enter a buffer early. In practice, a hybrid of the two may be useful: a structured curriculum with a limited replay component for persistent misses.

The answer-prior index is useful because it captures something that sensitivity alone does not. A model can improve sensitivity by moving all rare-target scores upward. The answer-prior index asks whether predictions still follow training prevalence after difficulty adjustment. The proposed curriculum lowers this index more than the alternatives. This suggests that it reduces the model's reliance on prevalence as a shortcut. It does not remove prevalence information entirely, and that may be appropriate. The goal is not to make the model ignorant of rarity. The goal is to prevent rarity from suppressing visible positive evidence.

The no-image and image-shuffle controls strengthen the interpretation. Standard training produces rare-target predictions that remain close to training prevalence even when meaningful image information is absent. The proposed curriculum increases the gap between no-image predictions and real positive images. It also increases sensitivity to image shuffling for rare targets. These controls suggest that the curriculum improves image conditioning. This is important because a method could otherwise appear to help simply by changing thresholds.

The findings also suggest that rare-positive learning is not only a data quantity problem. More rare-positive exposure helps, but the form of exposure matters. Oversampling increases quantity and improves sensitivity, yet it also increases false positives substantially. Counterbalanced contrast improves separation. This distinction is relevant for dataset construction. If developers collect additional rare positives, they should also collect matched hard negatives. A rare-positive dataset without appropriate controls may teach the model an overly broad positive concept.

The study has implications for benchmark reporting. Binary medical VQA benchmarks should not report only overall accuracy. They should report prevalence-tier metrics, rare-positive sensitivity, matched false-positive burden, and macro-balanced accuracy. If a benchmark includes rare findings, it should include balanced evaluation slices or the model's rare-positive behavior will be hidden [23]. A high score on a naturally imbalanced benchmark may say more about prevalence matching than about visual recognition.

The work also has implications for model selection. Early stopping on natural-prevalence validation accuracy can select a checkpoint with strong answer-prior collapse. In the training dynamics, standard validation accuracy peaks while rare-positive sensitivity remains poor. A validation score that includes rare-positive sensitivity and false-positive burden selects different checkpoints. Developers should therefore align validation criteria with the clinical failure modes they want to avoid.

The proposed curriculum is practical because it does not require extra annotations beyond labels and target categories. Difficulty matching improves results, but approximate difficulty strata can be built from existing signals such as image quality, preliminary model confidence, or target size when available. The method can be implemented with ordinary dataloaders. This practicality is important because many medical AI teams may not have access to dense annotations or model internals.

The curriculum is not a complete answer to rare finding recognition. Some rare positives remain missed because they are visually subtle, ambiguous, or poorly represented by the encoder. Some false positives appear because the model becomes more sensitive to rare-like patterns. Better image representations, higher-quality labels, and domain-specific augmentation remain necessary. The contribution of the curriculum is to reduce a training shortcut, not to solve all rare disease detection.

One subtle issue is that prevalence is clinically meaningful. A model that ignores prevalence completely may overcall rare findings. The proposed method does not aim for prevalence erasure. Its final stage intentionally returns to prevalence-realistic batches. The desired model should know that some findings are rare but should still answer yes when the image supports the finding. This balance is difficult. The results suggest that alternating distributional realism with diagnostic contrast is one workable approach.

Another issue is fairness across domains and patient groups. Rare findings may be even rarer in some subpopulations, and positive examples may differ visually across demographic or acquisition groups. If rare-positive examples are not diverse, a curriculum can overfit to the available positives. The present study balances domains but does not fully analyze demographic variation. Future work should examine whether curriculum methods improve or worsen subgroup performance. Rare-positive recovery should not benefit only the most represented patient groups.

The findings also connect to human learning. Clinicians learn both prevalence and contrast [24]. They know that some findings are uncommon, but they also study positive cases and mimics. A training stream that mostly shows negatives for rare findings may not provide enough contrast for a model to learn the visual concept. Counterbalanced prevalence cycling is a machine-learning version of alternating routine exposure with focused case review. The analogy is imperfect, but it helps explain why the schedule may work.

8. Limitations and Future Work

The first limitation is that the corpus uses simplified binary questions. Medical interpretation often involves multiple findings, severity, location, uncertainty, and temporal comparison. Answer-prior collapse may appear differently in those settings. A model generating a report could express a rare finding with hedged language rather than a simple yes or no error. Future work should extend collapse metrics to structured reports and multi-label outputs.

The second limitation is that prevalence tiers are constructed for controlled evaluation. Real clinical prevalence varies by institution, referral pattern, age group, scanner type, and season. A curriculum tuned to one prevalence structure may need adjustment elsewhere. The proposed method should be tested on external datasets with different prevalence patterns. A robust curriculum should reduce collapse without assuming that the training prevalence exactly matches deployment prevalence.

The third limitation is label quality. Rare positives are often harder to label confidently. Some may be subtle, uncertain, or inconsistently reported. Increasing their training influence can amplify label noise. The proposed curriculum reduces overfitting compared with oversampling, but it does not remove noise. Future work should combine counterbalanced curricula with uncertainty-aware labeling, expert adjudication, or soft labels for borderline cases.

The fourth limitation is that difficulty scores are approximate. The curriculum uses difficulty matching in contrast-balanced batches, but difficulty is not a perfect concept. Some examples are difficult because of image quality, some because of subtle pathology, and some because labels are ambiguous. Better difficulty estimation could improve batch construction. Active learning may help identify rare positives and hard negatives that provide the most useful contrast.

The fifth limitation is that the study uses a single general training objective across all domains. Some domains may benefit from specialized augmentation or architecture changes [25]. Retinal imaging and musculoskeletal radiography remain challenging after the curriculum. Their rare findings may require higher resolution, local crops, or specialized encoders. Curriculum design can improve learning from available examples, but it cannot fully compensate for inadequate visual representation.

The sixth limitation is that the method assumes target categories are known during sampling. This is true for binary VQA datasets, where each question names a target. It may be less straightforward for open-ended user questions or generated reports. A deployment-oriented system would need to map questions to target categories or use a more flexible sampling strategy. Future work could use language embeddings to group related questions for contrast-balanced training.

The seventh limitation is computational. Although the method does not add model parameters, constructing balanced contrast batches with difficulty matching requires indexing and sampling infrastructure. This is modest compared with model training, but it is more complex than ordinary random batching. In very large datasets, efficient batch assembly may require precomputed strata and periodic refreshes.

The eighth limitation is that the study does not evaluate downstream clinical decisions. Rare-positive sensitivity is important, but the acceptable tradeoff with false positives depends on clinical context. A screening tool, a triage assistant, and an educational VQA system may require different operating points. The reported metrics describe model behavior, not clinical utility [26]. Future studies should connect collapse control to decision-analytic outcomes.

Future work should investigate adaptive curricula. The schedule in this paper is fixed after validation [27]. A more adaptive method could monitor rare-positive uptake during training and adjust the frequency of contrast batches automatically. If a rare target remains weak, the curriculum could increase matched contrast for that target. If false positives rise, it could add harder negatives. Such adaptation may outperform a fixed schedule.

Another direction is combining curriculum learning with representation diagnostics. The present study observes output behavior but does not inspect internal features. A diagnostic method could identify whether rare positives occupy a separable region in representation space. If they do not, the problem may require better visual training rather than sampling changes. If they do, the problem may be mainly threshold or prior dominance. Distinguishing these cases would make curriculum design more targeted.

Future work should also examine multilingual and institution-specific question wording. The main experiment keeps wording simple to isolate answer priors. In practice, rare findings may be asked about in varied language. Some terms may appear rarely, compounding the rarity of the visual finding. A model may then face both visual rarity and linguistic rarity. Counterbalanced curricula could be extended to balance target, label, difficulty, and wording form together.

Another useful extension is prospective evaluation. A model trained with counterbalanced prevalence cycling could be tested on a stream of real clinical questions and images, with rare-positive alerts reviewed by specialists. Such a study would reveal whether reduced answer-prior index translates into better real-world case finding. It would also measure review burden from false positives.

Finally, the curriculum could be paired with data collection. The error audit identifies rare targets that remain unrecovered. These targets are candidates for targeted data acquisition. Instead of collecting more data uniformly, developers could collect rare positives and matched negatives for the specific targets where collapse persists. Training curricula and data collection strategies should be designed together.

9. Conclusion

This paper studied answer-prior collapse in binary medical vision-language inference. The central problem is that a model trained on imbalanced data can learn to answer according to finding prevalence rather than image evidence, especially for rare positive cases. Overall accuracy under natural prevalence can hide this behavior because negative examples dominate rare findings.

A multi-domain corpus was used to compare standard training, class weighting, focal training, positive oversampling, hard-positive replay, and counterbalanced prevalence cycling. The proposed curriculum alternates prevalence-realistic batches with balanced diagnostic contrast batches. It does not require new annotations or architectural changes. It changes the training sequence so that rare positives are repeatedly compared with matched negatives while the model continues to see the original distribution.

The empirical results show that counterbalanced prevalence cycling reduces rare-positive miss rate, improves macro-balanced accuracy, lowers answer-prior index, and preserves natural-prevalence accuracy better than aggressive oversampling. Hard-positive replay is a strong baseline, but the proposed curriculum is more stable and provides stronger prevalence control. The gains are consistent across chest radiography, retinal imaging, dermatology, and musculoskeletal radiography, though difficult rare targets remain.

The study also shows that rare-positive evaluation requires more than ordinary accuracy. Metrics such as rare-positive miss rate, false-positive burden, positive evidence uptake, and answer-prior index reveal failures that natural-prevalence accuracy can conceal. Medical VQA benchmarks that include rare findings should report such metrics if they aim to evaluate image-conditioned reasoning rather than majority-answer behavior.

The main conclusion is practical. When binary medical vision-language datasets contain strong prevalence imbalance, training schedule design can affect whether the model learns rare positive evidence. Counterbalanced curricula provide a simple way to reduce answer-prior collapse while maintaining exposure to realistic distributions. They should be viewed as part of a broader strategy that also includes better labels, matched negatives, rare-case data collection, and clinically meaningful validation.

References

[1] S. Salehi, Y. Singh, K. K. Horst, Q. A. Hathaway, and B. J. Erickson, "Agentic ai and large language models in radiology: Opportunities and hallucination challenges.," *Bioengineering (Basel, Switzerland)*, vol. 12, pp. 1303–1303, 11 2025.

[2] C. A. Munro, J. L. Jefferys, E. W. Gower, B. Munoz, C. G. Lyketsos, L. Keay, K. A. Turano, K. Bandeen-Roche, and S. K. West, "Predictors of lane-change errors in older drivers," *Journal of the American Geriatrics Society*, vol. 58, pp. 457–464, 3 2010.

[3] V. Mehta, C. Sharma, and K. Thiyagarajan, "Large language models and 3d vision for intelligent robotic perception and autonomy.," *Sensors (Basel, Switzerland)*, vol. 25, pp. 6394–6394, 10 2025.

[4] B. Sadanandan and V. Behzadan, "Mechanistically guided lora improves paraphrase consistency in medical vision-language models," *arXiv preprint arXiv:2603.00148*, 2026.

[5] W.-C. Lin, J. S. Chen, M. F. Chiang, and M. R. Hribar, "Applications of artificial intelligence to electronic health record data in ophthalmology.," *Translational vision science & technology*, vol. 9, pp. 13–13, 2 2020.

[6] D. J. Ramsey, J. T. Kwan, and A. Sharma, "Keeping an eye on the diabetic foot: The connection between diabetic eye disease and wound healing in the

lower extremity.," *World journal of diabetes*, vol. 13, pp. 1035–1048, 12 2022.

[7] N. Shi, Z. Liu, Z. Wan, G. Yang, Y. Shi, P. Wang, and X. Liu, "A vision-language model-based approach for lung cancer diagnosis using lossless 3d ct images: evaluation of gpt-4.1 and gpt-4o for patient-level malignancy assessment.," *Health information science and systems*, vol. 14, pp. 16–16, 12 2025.

[8] M. Chincarini, M. Signorini, N. Faccioli, and R. P. Mucelli, "Colorectal cancer vs chronic diverticular disease: differential diagnosisat ct colonography," *Insights into imaging*, vol. 8, pp. 224–224, 2 2017.

[9] A. Knight, T. Gschwind, P. Galer, G. A. Worrell, B. Litt, I. Soltesz, and S. Beniczky, "Artificial intelligence in epilepsy phenotyping.," *Epilepsia*, vol. 66 Suppl 3, pp. 39–52, 1 2024.

[10] S. Wang, X. Zhou, C. Li, S. Wang, Y. Li, T. Tan, and H. Zheng, "Generative artificial intelligence in medical imaging: Foundations, progress, and clinical translation.," *Research (Washington, D.C.)*, vol. 8, pp. 1029–1029, 12 2025.

[11] Y. Voskobiynyk and J. T. Paz, "Ai-nalyzing mouse behavior to combat epilepsy.," *Epilepsy currents*, vol. 23, pp. 315–317, 7 2023.

[12] V. Ahumada-Newhart, H. J. Kashyap, T. Hwu, Y. Tian, L. Mirzakhanian, M. Minton, S. Seader, S. Hedden, D. Moore, J. L. Krichmar, and J. S. Eccles, "Evaluation of the toyota human support robot (hsr) for social interaction and learning.," *International journal of technology, knowledge and society*, vol. 19, pp. 21–52, 5 2023.

[13] J. Wu, Y. Wang, Z. Zhong, W. Liao, N. Trayanova, Z. Jiao, and H. X. Bai, "Vision-language foundation model for 3d medical imaging," *npj Artificial Intelligence*, vol. 1, 8 2025.

[14] C. Fu and Q. Chen, "The future of pharmaceuticals: Artificial intelligence in drug discovery and development.," *Journal of pharmaceutical analysis*, vol. 15, pp. 101248–101248, 2 2025.

[15] L. Zühlke and M. E. Engel, "The importance of awareness and education in prevention and control of rhd.," *Global heart*, vol. 8, pp. 235–239, 9 2013.

[16] J. Morgan-Daniel, X. Goodman, S. G. Franklin, K. Bartley, M. N. Noe, and J. Pionke, "Medical library association diversity and inclusion task force report," *Journal of the Medical Library Association : JMLA*, vol. 109, pp. 141–153, 1 2021.

[17] T. M. Mengistu, T. Kim, and J.-W. Lin, "A survey on heterogeneity taxonomy, security and privacy preservation in the integration of iot, wireless sensor networks and federated learning.," *Sensors (Basel, Switzerland)*, vol. 24, pp. 968–968, 2 2024.

[18] K. Zhang, P. Gilley, N. Abdoli, T. C. Thai, Y. Chen, L. Dockery, K. Moore, R. S. Mannel, Q. Tang, and Y. Qiu, "Parameter efficient fine-tunning of foundation model to facilitate tumor response prediction for ovarian cancer patients.," *BMC medical imaging*, vol. 25, pp. 504–504, 12 2025.

[19] P. Cao, Z. Yang, L. Sun, Y. Liang, M. Q. Yang, and R. Guan, "Image captioning with bidirectional semantic attention-based guiding of long short-term memory," *Neural processing letters*, vol. 50, pp. 103–119, 1 2019.

[20] L. Xiao, J. Dong, E. Zhao, N. Hu, J. Zeng, Z. Wu, Y. Yang, J. Wang, Q. Xu, H. Shi, Z. Jin, L. Hu, and Y. Dong, "The association between multimorbidity and intrinsic capacity decline among older adults: the mediating role of frailty.," *BMC public health*, vol. 26, pp. 372–372, 11 2025.

[21] A. Kasagga, A. Sapkota, G. Changaramkumarath, J. M. Abucha, M. M. Wollel, N. Somannagari, M. Y. Husami, K. T. Hailu, and E. Kasagga, "Performance of chatgpt and large language models on medical licensing exams worldwide: A systematic review and network meta-analysis with meta-regression.," *Cureus*, vol. 17, pp. e94300–e94300, 10 2025.

[22] D. S. Shrestha, A. M. Straus, R. K. Sah, H. H. Khanal, R. Gautam, B. D. Paudel, T. M. Cornelius, and R. R. Love, "Illness and treatment beliefs in kathmandu valley nepalis under hypertensive care.," *PLOS global public health*, vol. 5, pp. e0004270–e0004270, 6 2025.

[23] S. Satheakeerthy, D. Jesudason, B. Bahrami, S. Bacchi, Y. M. Lee, R. Casson, M. Sun, and W. Chan, "Zero-shot llm-based visual acuity extraction: a pilot study.," *BMC ophthalmology*, vol. 25, pp. 359–359, 7 2025.

[24] W. Hsu, C. Baumgartner, and T. M. Deserno, "Notable papers and new directions in sensors, signals, and imaging informatics," *Yearbook of medical informatics*, vol. 30, pp. 150–158, 9 2021.

[25] E. Bösl and S. Samida, "Next generation sequencing: Challenges for science and society," *TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, vol. 30, pp. 10–52, 7 2021.

[26] L. F. Nakayama, L. Z. Ribeiro, F. Novaes, I. A. Miyawaki, A. E. Miyawaki, J. A. E. de Oliveira, T. Oliveira, F. K. Malerbi, C. V. S. Regatieri, L. A. Celi, and P. S. Silva, "Artificial intelligence for telemedicine diabetic retinopathy screening: a review.," *Annals of medicine*, vol. 55, pp. 2258149–2258149, 9 2023.

[27] M. Bunn, D. Khanna, E. Farmer, E. Esbrook, H. Ellis, A. Richard, and S. Weine, "Rethinking mental

healthcare for refugees.," *SSM. Mental health*, vol. 3, pp. 100196–100196, 2 2023.