MVC-Bench: Benchmarking Calibration of Medical Vision-Language Models

Ashshak Sharifdeen[1], Shihab Aaqil Ahamed[1], Ufaq Khan[1], Muhammad Akhtar Munir[1]
Sujair Ibrahim[2], Mohamed Rafeek Mareer Ahamed[3], Yutong Xie[1],
Imran Razzak[1], Muhammad Haris Khan[1]

[1]Mohamed bin Zayed University of AI
[2]Sabaragamuwa University of Sri Lanka
[3]Digital Platform Development, SLT PLC
{ashshak.sharifdeen, muhammad.haris}@mbzuai.ac.ae

Abstract

Reliable evaluation of vision-language models (VLMs) and medical vision-language models (Medical-VLMs) requires calibrated confidence, particularly under realistic clinical conditions. However, existing efforts mainly focused on improving accuracy, leaving calibration in the medical domain underexplored. To this end, we propose MVC-Bench, a calibration-centric benchmark for medical image classification with VLMs and Medical-VLMs. MVC-Bench assesses the calibration across three axes: (i) robustness to modality, backbone, and domain shift (ii) effectiveness of calibration strategies and prompt-tuning methods (iii) stability under prompt-template and random-seed variations. The benchmark covers eight different backbones, three medical modalities, including fundus imaging, histopathology, and chest X-ray under in-domain and domain shift settings. It compares post-hoc calibration, train-time calibration, and zero-shot inference methods, together with six prompt-tuning methods. Across more than 1638 controlled experiments, we report accuracy and Expected Calibration Error (ECE) as primary metrics, and further report results with complementary calibration measures, including Maximum Calibration Error (MCE) and Adaptive Calibration Error (ACE). We further investigate the underlying causes of miscalibration in VLMs and Medical-VLMs and propose a simple train-time calibration method, Multi-Class Margin (MCM) regularization, which achieves lowest ECE on 10 out of 12 settings in in-domain and remains competitive under domain shifts. Collectively, MVC-Bench provides a structured evaluation framework and actionable guidance for improving calibration in safety-critical medical workflows. Our Code:https://github.com/ashshaksharifdeen/MVC-Bench.

1 Introduction

VLMs leverage large-scale language supervision to align visual and textual representations, enabling strong zero/few-shot transfer with limited task-specific labels (Radford et al., 2021; Zhai et al., 2022; Li et al., 2022, 2023). Medical-VLMs extend this capability to clinical imaging, where annotations are scarce and costly. This capability has been explored across multiple medical modalities, including fundus imaging (Gulshan et al., 2016), histopathology (Campanella et al., 2019), and chest X-rays (Rajaraman et al., 2020). By aligning medical images with disease-aware prompts, medical-VLMs support disease-pattern recognition and transferable evaluation across heterogeneous medical datasets (Boecking et al., 2022; Wang et al., 2022; Zhang et al., 2023; Huang et al., 2023a; Ikezogwo et al., 2023a).Since full fine-tuning large VLMs is often costly, prompt tuning has become a practical parameter-efficient alternative, updating only a small set of prompt parameters while keeping the backbone frozen (Liu et al., 2023). This

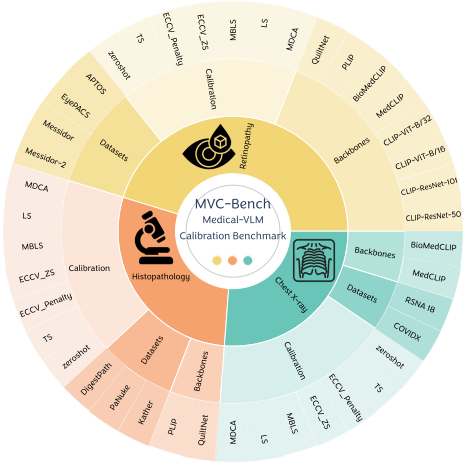


Figure 1: Overview of MVC-Bench, which evaluates calibration of VLMs and Medical-VLMs across three axes: robustness to modality, backbone, and domain shift; effectiveness of calibration and prompt-tuning strategies; and stability under prompt-template and random-seed variations. The benchmark spans DR, histopathology, and chest X-ray datasets, covering relevant backbones and calibration settings across more than 1638 controlled experiments.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 36128–36162
October 24-29, 2026 ©2026 Association for Computational Linguistics

low-cost adaptation is particularly valuable in medical imaging, where full fine-tuning is often infeasible (Hanif et al., 2024; Hussein et al., 2024).

Despite recent advances, prompt-tuning methods for VLMs mainly optimize accuracy while often overlooking confidence calibration (Sharifdeen et al., 2026). However, clinical deployment requires confidence scores that reflect the true likelihood of correctness, where a classifier is *well-calibrated* if predictions made with confidence p are correct approximately $p\%$ of the time (Guo et al., 2017). Miscalibration can cause *overconfidence*, where incorrect predictions receive high confidence and may miss human review, or *underconfidence*, where correct predictions receive unreliable confidence, undermining triage and trust (Minderer et al., 2021; Cao et al., 2024; Zhu et al., 2023b; Ao et al., 2023). In medical imaging, this problem is further amplified by class imbalance, heterogeneous acquisition, prevalence drift, and device or patient shifts (Koh et al., 2021). Therefore, calibration of VLMs and Medical-VLMs under realistic factors, including domain shift, prompt tuning, hard-prompt choice, and seed variability, remains insufficiently explored.

In this paper, we address this crucial gap with **MVC-Bench**, a comprehensive calibration-centric benchmark for VLM-based medical image classification. **MVC-Bench** organizes calibration evaluation along three dimensions: *robustness* across backbone, modality, and domain shift; *effectiveness* of calibration strategies and prompt-tuning methods; and *stability* under prompt-template and random-seed variations. This factorized design enables systematic analysis under clinically relevant conditions. To our knowledge, **MVC-Bench** is the first calibration-centric benchmark focused on prompt-tuned VLM and Medical-VLM image classification across these three medical modalities. We benchmark classical and VLM-aware calibration techniques across eight popular backbones CLIP-ResNet/ViT (Radford et al., 2021), PLIP (Huang et al., 2023b), QuiltNet (Ikezogwo et al., 2023a), MedCLIP (Wang et al., 2022), and BioMedCLIP (Zhang et al., 2023) and systematically compare prompt-tuning approaches by adopting CoOp as a widely used baseline and contrasting with more advanced prompt learners KgCoOp, MaPLe, PromptSRC, HiCroPL, and ProGrad (Zhou et al., 2022b; Khattak et al., 2023a,b; Yao et al., 2023; Zheng et al., 2025a; Zhu et al., 2023a). Our datasets span three modalities: (i) *fundus* (Aptos (APTOS, 2019), EyePACS (Kaggle, 2015), Messidor (Decencière et al., 2014a), Messidor-2 (Decencière et al., 2014b)), (ii) *histopathology* (Kather (Kather et al., 2019), PaNuke (Gamper et al., 2019), DigestPath (Da et al., 2022)), and (iii) *X-ray* (COVIDX (Rahman et al., 2021), RSNA 18 (rsn, 2018)).

Fig. 1 illustrates the coverage of our MVC-Bench. In total, we conduct over 1638 experiments organized around three dimensions: *robustness*, *effectiveness*, and *stability*. These dimensions are guided by seven research questions: Under *robustness*, we examine How domain shifts affect calibration, and Which backbones are better calibrated across medical modalities. Under *effectiveness*, we analyze the accuracy-calibration trade-offs, which calibration methods are consistently reliable across backbones, and how prompt-tuning families affect calibration. Under *stability*, we investigate How do different prompt initialization interact with Calibration, and How different seed initializations affect calibration performance?. We report accuracy and expected calibration error (ECE) (Pakdaman Naeini et al., 2015), with additional metrics in the supplementary.

Guided by these findings, we analyze miscalibration sources and observe that underconfidence often stems from weak inter-class separation, where the correct-class logit only slightly exceeds competing logits. We propose **Multi-Class Margin (MCM)** Regularization, a train-time regularizer that increases true-vs-rest margins while controlling margin variance to reduce underconfidence and spurious confidence spikes. MCM achieves the lowest ECE for most backbones and remains competitive under domain shifts. Overall, our contributions are a calibration-centric benchmark for VLMs and Medical-VLMs across robustness, effectiveness, and stability axes, a large-scale evaluation across modalities, backbones, calibration methods, and prompt-tuning strategies, and a simple Multi-Class Margin (MCM) regularizer that improves in-domain ECE while remaining competitive under domain shift.

2 Related Work

Medical VLMs and Prompt Tuning. Medical imaging has increasingly adopted VLMs to reduce annotation demands and improve transfer across institutions and modalities. In radiology, MedCLIP,

BioViL/BioViL-T, BioMedCLIP, and CheXzero align medical images with free-text reports for zero-shot labeling and retrieval (Wang et al., 2022; Boecking et al., 2022; Zhang et al., 2023; Tiu et al., 2022), while pathology-focused models such as PLIP and Quilt/QuiltNet use language-supervised pretraining over slide tiles and pathology-specific corpora to improve recognition under data heterogeneity (Huang et al., 2023b; Ikezogwo et al., 2023b).

Prompt-tuning methods, beginning with CoOp (Zhou et al., 2022c) and later extended by KgCoOp, MaPLe, HiCroPL, PromptSRC, and ProGrad (Yao et al., 2023; Khattak et al., 2023a; Zheng et al., 2025a; Khattak et al., 2023b; Zhu et al., 2023a), adapt frozen VLMs through learnable prompts. However, although Medical-VLMs and prompt tuning often improve transfer or accuracy, their confidence calibration across modalities, institutions, prompts, and deployment shifts remains insufficiently understood, and prompt adaptation can even degrade calibration by shifting decision boundaries (Sharifdeen et al., 2026).

Calibration in Deep Neural Network and Medical Models. Calibration methods are commonly divided into post-hoc approaches, which adjust confidence without changing decision boundaries, and train-time approaches, which add calibration-aware objectives or regularizers. Representative post-hoc methods include Platt scaling, isotonic regression, binning-based calibration, temperature scaling, and multiclass calibration (Platt et al., 1999; Zadrozny and Elkan, 2002; Naeini and Cooper, 2016; Guo et al., 2017; Kull et al., 2019), while train-time methods include label smoothing, mixup, focal loss, class-balanced objectives, trainable calibration losses, and deep ensembles (Müller et al., 2019; Thulasidasan et al., 2019; Lin et al., 2017; Cui et al., 2019; Kumar et al., 2018; Lakshminarayanan et al., 2017).

In medical imaging, calibration is especially important for triage, escalation, and human-AI collaboration (Gawlikowski et al., 2023), but it is complicated by class imbalance, label noise, acquisition variation, and site shift (Karimi and Gholipour, 2020; Laves et al., 2020; Buda et al., 2018; Ovadia et al., 2019; Guo et al., 2017; Kull et al., 2017; Pakdaman Naeini et al., 2015; Lakshminarayanan et al., 2017). Most prior studies still focus on supervised CNN/ViT models with fixed label spaces (Minderer et al., 2021), leaving calibration in promptable, open-vocabulary Medical-VLMs under prompt tuning and domain shift relatively underexplored, despite the strong influence of prompts, class granularity, and template design on confidence reliability (Radford et al., 2021; Zhou et al., 2022b,a; Khattak et al., 2023a; Tiu et al., 2022; Boecking et al., 2022; Zhang et al., 2023).

3 MVC-Bench

Our **MVC-Bench** is formulated as a calibration stress-test protocol for VLMs and Medical-VLMs, rather than a single in-domain leaderboard. It organizes seven research questions into three axes: *robustness* to modality, backbone choice, and domain shift; *effectiveness* of calibration and prompt-tuning strategies; and *stability* under prompt-template and random-seed variation.

Dataset characteristics MVC-Bench covers fundus imaging (APTOS, EyePACS, Messidor, Messidor-2), histopathology (Kather, PaNuke, DigestPath), and CXR (COVIDX, RSNA18) (APTOS, 2019; Kaggle, 2015; Decencière et al., 2014b; Kather et al., 2019; Gamper et al., 2019; Da et al., 2022; Rahman et al., 2021; Stein et al., 2018). These datasets provide heterogeneous visual statistics, label spaces, and class priors for evaluating calibration under in-domain and shifted settings.

Implementation details. We report top-1 accuracy (Acc.) and 20-bin Expected Calibration Error (ECE) as the primary evaluation metrics, and provide complementary calibration analyses using Maximum Calibration Error (MCE) and Adaptive Calibration Error (ACE) in the appendix. Prompt tuning uses CoOp in a 16-shot setting with learning rate 0.005 and batch size 8, following official baseline implementations. All experiments are run over three seeds on an NVIDIA RTX A6000 GPU, with full hyperparameters in the *supplementary material*.

Vision-Language Models (VLMs). We evaluate both generic VLMs and Medical-VLMs to study calibration across model families and modalities. The generic backbones include CLIP-ViT-B/16, CLIP-ViT-B/32, CLIP-ResNet-50, and CLIP-ResNet-101, while the medical backbones include PLIP, QuiltNet, MedCLIP, and BioMedCLIP. We further analyze parameter-efficient prompt-tuning methods, including CoOp, KgCoOp, MaPLe, PromptSRC, ProGrad, and HiCroPL, to assess their impact on calibrated confidence.

Evaluated calibration baselines. For our study, we selected prominent and well-referenced calibra-

tion methods covering both train-time and post-hoc approaches: MDCA (Hebbalaguppe et al., 2022), LS (Müller et al., 2019), MBLS (Liu et al., 2022a), ECCV_ZS (Murugesan et al., 2024), ECCV_penalty (Murugesan et al., 2024), and Temperature Scaling (Guo et al., 2017). In addition, we report zero-shot inference for each frozen backbone as an uncalibrated reference baseline, without prompt tuning and post-hoc calibration.

4 Experiments and Discussion

4.1 Robustness: How do domain shifts affect calibration?

Domain shift is a central robustness challenge for clinical deployment of VLMs and Medical-VLMs, as changes in acquisition devices, patient populations, and clinical protocols can alter confidence reliability (Koh et al., 2021; Ghafoorian et al., 2017). These shifts manifest differently across modalities: fundus images vary across cameras and clinics (Galappaththige et al., 2024), histopathology is affected by staining variation (Tellez et al., 2019), and chest X-rays reflect differences in disease prevalence and demographics (Zech et al., 2018). Since such shifts may preserve accuracy while degrading calibration, we examine whether ECE-based calibration performance learned in in-domain settings remains stable under domain-shifted evaluation across modalities.

Experimental setup: We evaluate whether in-domain (ID) calibration predicts domain shift (DS) performance using ECE across eight backbones and seven calibration methods. The analysis covers three domains: retinopathy (ID: Messidor, excluding class 4; DS: APTOS, EyePACS, Messidor-2), chest X-ray (ID: RSNA18; DS: COVIDX), and histopathology (ID: PaNuke; DS: DigestPath, Kather). For each backbone, we rank calibration methods by ID ECE and compare them with DS rankings using Kendall's τ_b correlation.

Discussion: In Fig. 2, most retinopathy backbones show negative or near-zero τ_b, indicating weak calibration-rank transfer under external-domain evaluation. Although some backbones preserve rankings for particular datasets, this behavior is not consistent across shifts. Across all 30 ID-to-DS rank-transfer tests, the mean and median Kendall correlations are -0.0598 and -0.0476, respectively 17/30 correlations are negative, 13/30 are positive, and none shows significantly positive rank preservation at $p < 0.05$ (Appendix H). Complementary ACE and MCE analyses (Appendix K.1) show that rankings can remain stable for some closer shifts, but this stability is not universal across modalities, backbones, or metrics. Overall, **calibration-method rank transfer under external-domain shift is weak and metric-dependent.**

4.2 Robustness: Which VLM-based backbones are better calibrated?

Backbone selection is another key factor towards gauging calibration robustness because generic VLMs and Medical-VLMs encode different visual and semantic priors. We therefore compare generic and domain-specific backbones across DR, histopathology, and X-ray tasks to identify which model families provide more reliable confidence beyond accuracy.

Experimental setup: We assess calibration across VLM and Medical-VLM backbones on DR grading, histopathology, and X-ray tasks using seven calibration methods, reporting the mean and standard deviation of ECE.

Discussion: Fig. 3 shows clear calibration disparities across VLM and Medical-VLM backbones. Domain-specific models generally achieve lower mean ECE than generic CLIP variants, with BioMedCLIP-X-ray outperforming MedCLIP-X-ray and PLIP/QuiltNet variants showing strong calibration in their aligned modalities. However, MedCLIP-X-ray is an important counterexample, showing that medical-domain pretraining alone does not guarantee calibrated confidence. Although MedCLIP and BioMedCLIP are both medically pretrained, they differ substantially in pretraining scale and data coverage. BioMedCLIP is trained on the much broader PMC-15M biomedical image-text corpus(Zhang et al., 2023). Thus, calibration appears to depend not only on whether pretraining is domain-specific, but also on its scale, coverage, label-space alignment, and resulting confidence geometry. ACE and MCE analyses lead to the same broader conclusion and likewise identify MedCLIP-X-ray as a particularly poorly calibrated backbone (Appendix K.2). Overall, **medical/domain-aligned pretraining generally improves calibration, but it is not sufficient by itself to guarantee reliable confidence.**

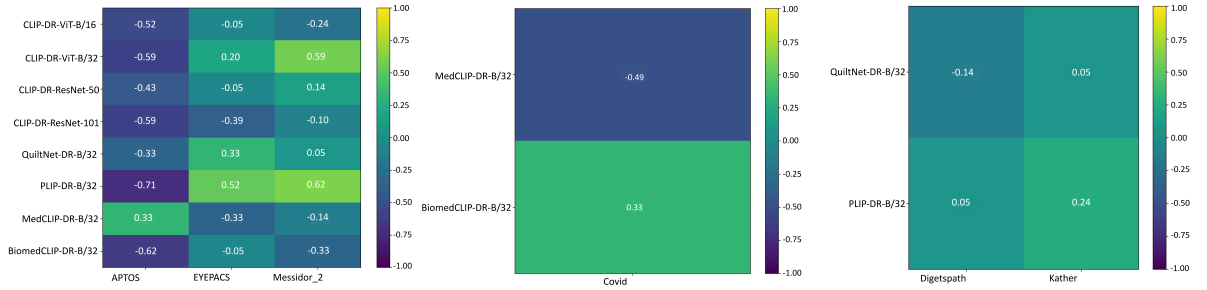


Figure 2: Kendall rank correlation (τ) between ECE-based method rankings on the In-domain(ID) source and each Domain shift(DS) dataset is shown for every backbone (rows = backbones, columns = DS sets). Cells are colored by τ (-1 to 1) and annotated with values: higher/positive τ means rankings (and thus relative calibration quality) are preserved under shift, while near-zero/negative τ indicates rank reversals of the calibration under shift.

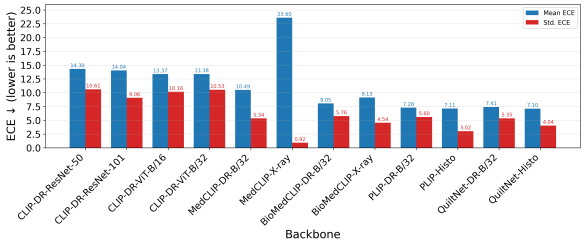


Figure 3: Comparison of mean and Std. of ECE for various VLM-based backbones across different medical modality inputs.

4.3 Effectiveness: How does calibration vary with accuracy for different calibration methods?

Calibration effectiveness cannot be inferred from accuracy alone, because medical models may remain confidently wrong under class imbalance, heterogeneous acquisition, or prevalence shift. We therefore analyze the accuracy-calibration relationship across calibration methods and backbones using Pearson correlation between Acc. and ECE.

Experimental setup: We compute Pearson correlation $r(\text{Acc}, \text{ECE})$ across datasets for each calibration method and backbone to examine the accuracy–calibration relationship. Negative values indicate that higher accuracy corresponds to lower ECE, while positive values indicate that higher accuracy is associated with higher ECE.

Discussion: Fig. 4 shows that the accuracy-ECE relationship is highly method and backbone-dependent. TS, MDCA, and MBLS often show negative correlations, indicating that accuracy and calibration can improve together, whereas ECCV_Penalty consistently shows positive correlations, suggesting an accuracy-calibration trade-off. CLIP backbones exhibit mixed behaviour, while X-ray backbones show extreme ± 1 values due to limited dataset pairs. Overall, **there is no universal accuracy-calibration trade-off, calibration effectiveness depends strongly on both the method and backbone.** Complementary ACE and MCE analyses show the same overall behavior: the relationship between predictive accuracy and calibration remains method and backbone dependent, confirming that higher accuracy does not generally imply better calibration (Appendix K.3).

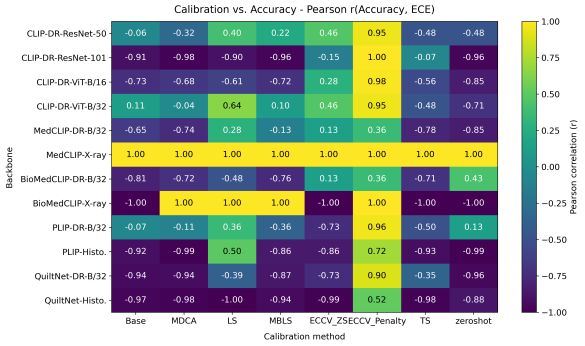


Figure 4: Pearson correlation matrix between accuracy and ECE evaluated across diverse calibration methods and backbones.

4.4 Effectiveness: Which calibration methods show consistent calibration performance across various backbones?

A practical calibration method should remain reliable across backbones, modalities, and deployment shifts without requiring extensive re-tuning. We therefore evaluate whether each calibration strategy maintains low and consistent ECE across VLM and Medical-VLM backbones in both ID and DS settings.

Experimental setup: In Fig. 5, we evaluate seven calibration methods across 12 backbones and three medical modalities. The ID panel reports ECE for each backbone-method pair, while the DS panel reports the median backbone-level $\Delta\text{ECE} = \text{ECE}_{\text{DS}} - \text{ECE}_{\text{ID}}$ across the corresponding shifted datasets to assess robustness.

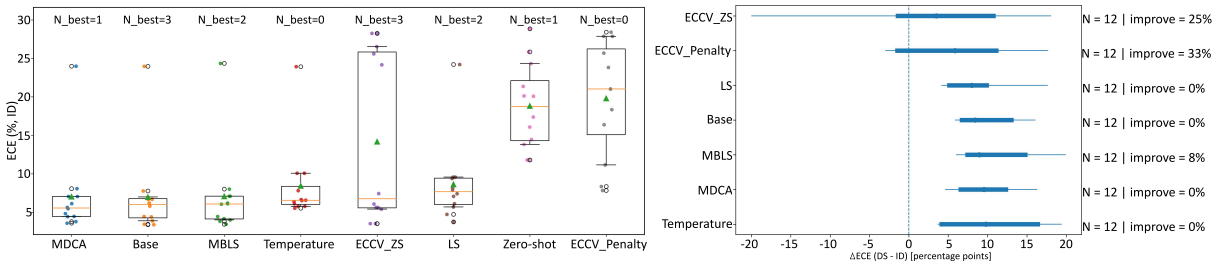


Figure 5: Left (ID): Boxplots of ECE (%) per method across backbones (IQR boxes, 10–90% whiskers, ▲ mean, colored dots = backbone points, hollow circles = outliers). N_{best} shows the number of times it performs best across backbones. Lower, tighter boxes = more consistent ID calibration. Right (DS): Forest plot summarizing how each method's ECE changes from ID to DS across backbones (dot = median, thick bar = IQR, thin line = 10–90%, dashed line at zero change), right text reports N backbones and improve % (share of backbones that get lower ECE in DS).

Discussion: In ID settings, MBLS, Base, and MDCA show the most consistent calibration with low ECE and smaller spread, while LS and TS perform moderately, and ECCV_ZS, ECCV_Penalty, and Zero-shot show higher ECE variability. N_{best} further favors Base as the most frequent lowest-ECE method. Under DS, ΔECE is mostly positive, indicating degraded calibration. Only ECCV_ZS and ECCV_Penalty improve in a subset of cases, likely due to their use of pretrained knowledge for shift-aware calibration. Overall, **no method is consistently reliable across both ID and DS settings, highlighting the need for calibration methods specifically designed for robust medical deployment under domain shift.** This conclusion is consistent under ACE and MCE: calibration method rankings and dispersion remain backbone and shift-dependent, with no existing baseline providing uniformly reliable calibration across ID and DS settings (Appendix K.4).

4.5 Effectiveness: How do different prompt-tuning methods interact with Calibration?

Recent work has studied how prompt tuning effectively affects calibration and global logit distributions in natural image benchmarks (Liu et al., 2023). However, in medical imaging, this interaction is more complex due to device heterogeneity, prevalence shifts, ambiguous class boundaries, context length, text-image alignment, and backbone pretraining. These factors can significantly alter confidence reliability even when accuracy remains high.

Experimental setup: We compare six prompt-tuning families CoOp, KgCoOp, MaPLe, PromptSRC, HiCroPL, and ProGrad on DR classification across APTOS, EyePACS, Messidor, and Messidor-2. Using the same backbone, data splits, and no additional calibration method, only the prompt-tuning strategy is varied. The figure reports mean Accuracy and ECE across datasets with standard deviation error bars, where higher Accuracy and lower ECE indicate better performance.

Discussion: Fig. 6 shows a clear accuracy–calibration trade-off. PromptSRC achieves the highest accuracy (∼72%) but has poor calibration (ECE ≈11.5%) and high variability, while ProGrad performs worst in both accuracy (∼66%) and ECE (∼23%). In contrast, CoOp, HiCroPL, and KgCoOp provide the best balance, maintaining competitive accuracy (∼68–75%) with low ECE (∼3.5–4.8%) and smaller variance. MaPLe shows intermediate performance. Overall, **high accuracy alone does not guarantee reliable confidence, highlighting the need for prompt-tuning methods that explicitly preserve calibration.** ACE and MCE exhibit the same broad prompt-family trend, with HiCroPL, CoOp, and KgCoOp providing comparatively stronger calibration than several higher-error prompt learners (Appendix K.5). Thus, the observed accuracy calibration mismatch is not specific to ECE.

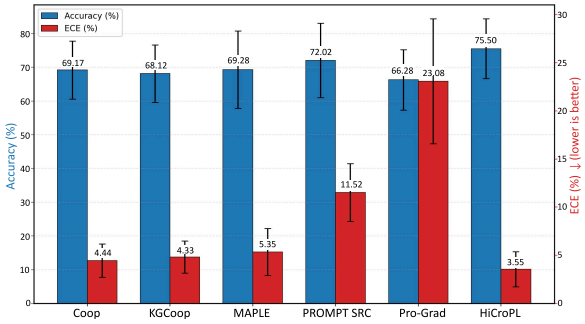


Figure 6: Comparison of Acc and ECE for different prompt-tuning methods.

4.6 Stability: How does hard-prompt initialization affect calibration?

Prompt-template selection is an important stability factor because Generic prompts may improve calibration in natural-image settings (Yoon et al., 2024; Sharifdeen et al., 2025; Ahamed et al., 2025), but medical tasks require careful phrasing due to fine-grained ontologies and variable terminology. Since subtle wording changes can alter confidence without changing accuracy, we benchmark natural-image-inspired, noisy, and clinically tailored prompts to study their effect on medical VLM calibration.

Experimental setup: We study the effect of hard-prompt initialization on calibration using MaPLe-ViT-B/16 (Khattak et al., 2023a) for diabetic retinopathy classification. We evaluate Accuracy and ECE under seven hard-prompt templates (P1-P7) with different phrasing, context, and domain specificity. Prompt-template details are provided in the supplementary material.

Discussion: Fig. 7a reports the ECE standard deviation across datasets for each hard-prompt template, with mean ECE μ shown as a side label. Templates P1–P4 achieve low and stable ECE, with *"a photo of a"* giving the most stable calibration. In contrast, the noisy P7 prompt shows high variance, indicating that calibration is highly sensitive to hard-prompt initialization. **This highlights the need for prompt-tuning strategies that remain robust to prompt wording and initialization choices.** The same sensitivity is observed with ACE and MCE, where early templates such as P1-P2 remain comparatively stable while P6-P7 show substantially larger variability(Appendix K.6).

4.7 Stability: How do different seed initializations affect calibration performance?

Random-seed variation is another important source of calibration instability in medical imaging, where limited data, long-tailed prevalence, label noise, and subtle findings can make confidence estimates sensitive to initialization. This instability can be further amplified under prevalence and acquisition shifts (Guo et al., 2017). To assess reproducibility, we run each setup across three seeds and report the standard deviation of ECE and accuracy, allowing us to identify backbones with stable calibration estimates and those that are more seed-sensitive.

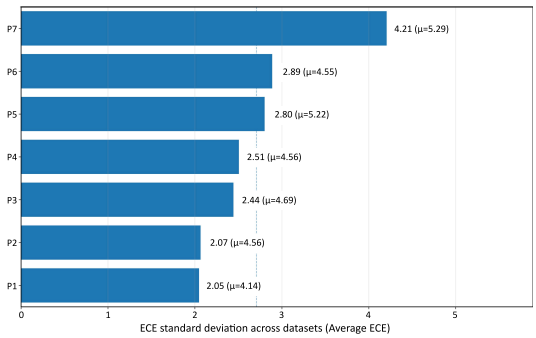
Experimental setup: We assess random-seed sensitivity by running each backbone with the base calibration method across three seeds, $s \in 1, 2, 3$, while keeping data splits, hyperparameters, and preprocessing fixed. Seed variability is measured as the standard deviation of ECE across seeds, with lower Std. ECE indicating more stable calibration.

Discussion: Fig. 7b shows that calibration stability varies across seeds, with lower Std. ECE indicating more stable backbones. BioMedCLIP-DR-B/32, PLIP-DR-B/32, and CLIP-DR-ViT-B/32 show low seed variability, with Std. ECE values of 0.25, 0.47, and 0.49, respectively, indicating consistent calibration. In contrast, MedCLIP-DR-B/32 and MaPLe-DR-ViT/16 show higher variability, with Std. ECE values of 2.49 and 1.45, making them more seed-sensitive. Overall, **calibration performance and reproducibility depend on both seed initialization and backbone choice.** ACE and MCE likewise confirm that seed sensitivity is strongly backbone-dependent, with MedCLIP variants showing particularly large variation compared with several other backbones (Appendix K.7).

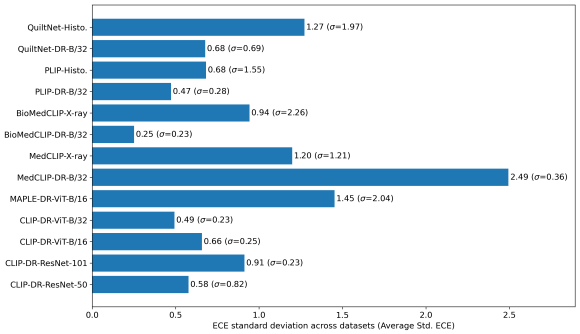
5 Multi-Class Margin Regularization (MCM)

Motivation: Findings from research questions 4.1, 4.2, and 4.4 show that ID calibration often fails to transfer to DS settings, increasing ECE, while backbone and calibration-method performance remain inconsistent. To understand this, we analyze DR reliability diagrams in Fig. A1b and observe underconfidence across two backbones, both with and without MBLS (Liu et al., 2022a). Fig. 8b further shows that CE and CE+MBLS concentrate most samples in the low-margin region ($\bar{m} \leq 1$), where the correct-class logit only slightly exceeds competing logits. This weak inter-class separation produces uncertain confidence and underconfidence. MBLS reduces extreme margins to limit overconfidence but does not sufficiently lift small margins, leaving logits concentrated in the low-margin region.

Multi-Class Margin Regularization: Our proposed **MCM** introduces a true-class-vs-rest margin objective that reshapes the inter-class logit distribution. It shifts samples from low-margin regions ($\bar{m} \leq 1$) toward moderate margins ($\approx$ 2–4), improving class separation while controlling margin dispersion to produce reliable confidence. For logits $\mathbf{z}_i \in \mathbb{R}^C$ and label y_i, we define the true-vs-rest margin as $m_{i,k} = z_{i,y_i} - z_{i,k}$ for $k \neq y_i$, the per-

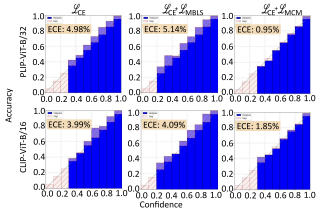


(a) Comparison of Std. and mean of ECE (μ) across datasets for seven different hard-prompt templates (P1-P7). Templates P1-P4 have low and stable ECE across datasets, while P7 has high variance.

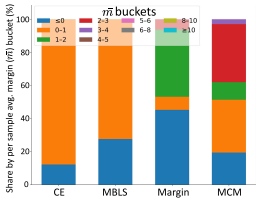


(b) Comparison of seed variability, i.e., Std. of ECE and Acc (σ), across different backbones and three random seeds. Some backbones show stable calibration, while others are seed-sensitive.

Figure 7: Prompt-template sensitivity and random-seed sensitivity of calibration.



(a) Comparison of Calibration error across 3 regularizers on two backbones,



(b) Comparison of per-sample average true class-vs-rest of the class margin across different methods.

Figure 8: Miscalibration analyses.

sample average margin as $\bar{m}_i = \frac{1}{C-1}\sum_{k\neq y_i} m_{i,k}$, and the batch mean as $\mu = \frac{1}{B}\sum_{i=1}^{B}\bar{m}_i$. The all-pairs margin variance is then computed as $\sigma^2_{\text{all-pairs}} = \frac{1}{B(C-1)}\sum_{i=1}^{B}\sum_{k\neq y_i}(m_{i,k}-\mu)^2$, and the MCM loss is defined as:

$$\mathcal{L}_{\text{MCM}} = -\alpha\,\mu + \beta\,\sigma^2_{\text{all-pairs}}, \qquad \alpha,\beta>0, \quad (1)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{MCM}}. \quad (2)$$

The μ term, weighted by α, increases the margin between the ground-truth logit and competing class logits, directly reducing underconfidence. The $\sigma^2_{\text{all-pairs}}$ term, weighted by β, controls margin dispersion and prevents spurious confidence spikes by stabilizing the margin distribution. As shown in Fig. A1b, MCM substantially reduces ECE across both backbones compared with existing methods. Unlike the concurrent MARGIN loss (Sharifdeen et al., 2026), which only considers the ground-truth versus second-highest logit margin, MCM uses all true-vs-rest margins. Fig. 8b shows that MARGIN reduces samples in the $(0\text{–}1)$ bucket but remains unstable across other margin regions, including the ≤ 0 bucket where incorrect classes dominate. In contrast, MCM maintains a more stable inter-class margin distribution across all buckets.

Results: Tab. 1 provides a compact comparison of MCM against Base and ECCV_ZS, which were selected as representative baselines based on their stronger ID consistency and comparatively favorable DS behavior in Fig. 5. Across the 12 modality-specific settings, MCM achieves the lowest ECE in 10/12 ID configurations while remaining competitive under DS. Complete per-setting comparisons against all applicable calibration baselines, including Base, MDCA, LS, MBLS, ECCV_ZS, ECCV_Penalty, and Temperature Scaling, are reported in Appendix Tabs. A8 and A9. Comparisons across additional prompt-tuning methods are also provided in the supplementary material. Fig. 9 extends the comparison to all evaluated calibration methods. MCM achieves the highest frequency of best calibration performance in ID, obtaining the lowest ECE in 10/12 settings. Under DS, MCM achieves the lowest ECE in 4/12 settings, while other methods, particularly ECCV_ZS, remain competitive in individual configurations. The broader DS distributions further show that strong ID calibration does not guarantee uniform robustness after distribution shift. Overall, the figure supports the conclusion that MCM provides strong and consistent ID calibration gains while its DS advantage is more setting-dependent. **Statistical significance:** A paired Wilcoxon analysis across the 12 matched configurations further supports the ID gains of MCM, with statistically significant improvements over ECCV_ZS ($p = 0.0212$) and ECCV_Penalty ($p = 0.0005$). Under DS, MCM remains competitive but is not uniformly superior.

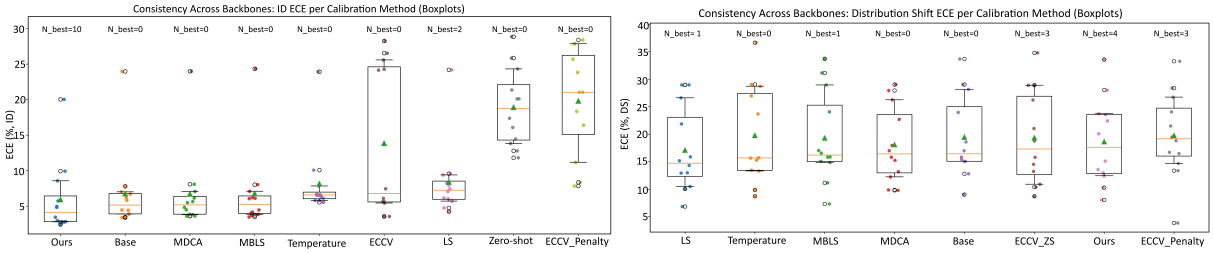


Figure 9: ECE (%) across backbones for each calibration method under ID (top) and DS (bottom) settings. Boxplots show the interquartile range (IQR), whiskers indicate the 10–90% range, and individual backbone results are overlaid as colored markers. N_{best} denotes the number of backbones for which a method achieves the lowest ECE. Lower medians and tighter distributions indicate more consistent calibration across architectures.

Full statistical results are provided in Appendix I.

Hyperparameter transfer: The MCM hyperparameters are selected using validation data only and then fixed across all remaining backbones, modalities, and DS evaluations. A four-fold DR Leave One Dataset Out analysis independently selects the same $(\alpha, \beta) = (0.1, 0.01)$ setting in every fold, supporting its transferability without target-dataset tuning (Appendix G.1).

Failure cases: MCM primarily targets low-margin underconfidence and is therefore not uniformly effective across modalities. For PLIP-Histo. and QuiltNet-Histo., the modality-specific reliability patterns indicate a different calibration regime, with overconfidence more prominent than the low-margin underconfidence targeted by MCM (Appendix Fig. A2). Consequently, MCM increases ECE in these cases.

Ablation: Tab. 2 shows that using either $-\alpha\mu$ or $+\beta\sigma^2_{\text{all-pairs}}$ alone worsens calibration, increasing ECE to 5.45% and 6.37%. Combining both terms reduces ECE to 3.43% while preserving accuracy at 69.05%, confirming that margin expansion and margin control are complementary for calibration.

Backbone	Metric	ID			DS		
		Base	ECCV_ ZS	Ours (MCM)	Base	ECCV_ ZS	Ours (MCM)
CLIP-DR-ResNet-50	Acc.	67.87	72.09	67.71	52.29	27.98	51.30
	ECE	4.40	28.26	**2.96**	15.07	18.77	**10.26**
CLIP-DR-ResNet-101	Acc.	68.29	67.69	68.19	41.91	59.63	51.73
	ECE	5.82	25.59	**4.82**	16.99	34.79	**15.12**
CLIP-DR-ViT-B/16	Acc.	69.18	68.56	69.05	32.90	38.51	33.35
	ECE	4.44	26.54	**3.43**	33.68	**15.79**	20.08
CLIP-DR-ViT-B/32	Acc.	68.13	72.09	67.79	43.58	29.03	46.59
	ECE	3.89	28.26	**2.77**	15.83	14.53	**12.44**
MedCLIP-DR-B/32	Acc.	59.82	59.85	59.74	52.69	59.91	59.83
	ECE	6.69	5.63	**4.94**	23.95	26.23	**23.54**
MedCLIP-X-ray	Acc.	65.77	65.96	65.83	79.15	79.02	79.15
	ECE	23.97	24.17	**20.03**	29.03	28.91	**28.04**
BioMedCLIP-DR-B/32	Acc.	67.47	59.85	66.80	59.45	59.11	59.26
	ECE	3.88	5.63	**2.82**	15.06	21.03	**12.94**
BioMedCLIP-X-ray	Acc.	73.56	72.94	73.70	74.56	73.63	78.84
	ECE	7.78	7.44	**5.71**	9.00	10.38	**8.02**
PLIP-DR-B/32	Acc.	68.05	67.79	67.97	55.40	54.27	47.16
	ECE	3.38	3.53	**2.75**	12.79	**10.88**	13.56
PLIP-Histo.	Acc.	85.43	86.27	85.30	52.83	65.35	53.88
	ECE	7.00	**6.08**	9.92	18.58	**13.24**	22.43
QuiltNet-DR-B/32	Acc.	66.54	67.79	66.49	40.08	50.50	27.92
	ECE	3.45	3.53	**2.40**	15.42	**8.69**	23.72
QuiltNet-Histo.	Acc.	86.73	87.87	86.87	46.46	44.93	46.76
	ECE	6.21	**5.42**	8.55	**28.15**	28.94	33.60

Table 1: Overall average accuracy and ECE across 12 settings covering three modalities. ID: in-domain and DS: domain shift (out-domain).

Loss Term	Metric	Aptos	eyepacs	messidor	messidor_2	Avg
CoOp (Zhou et al., 2022b)	Acc.	77.90	74.64	59.50	64.66	69.18
	ECE	3.99	2.25	5.29	6.24	4.44
MBLS (Liu et al., 2022a)	Acc.	77.94	74.64	59.53	64.62	69.18
	ECE	4.09	2.25	5.31	6.13	4.45
$-\alpha\,\mu$	Acc.	74.82	73.72	59.83	62.79	67.54
	ECE	4.81	9.32	4.98	2.71	5.45
$+\beta\,\sigma^2_{\text{all-pairs}}$	Acc.	78.26	74.49	58.42	64.62	68.94
	ECE	8.03	2.98	5.61	8.87	6.37
$-\alpha\,\mu + \beta\,\sigma^2_{\text{all-pairs}}$	**Acc.**	77.50	74.53	59.72	64.43	69.05
	ECE	1.85	3.25	5.09	3.53	**3.43**

Table 2: Loss Component analyses in our MCM.

6 Conclusion

We introduced **MVC-Bench**, a large-scale benchmark for evaluating calibration in VLMs and Medical-VLMs under clinically relevant conditions. Our analysis shows that calibration is sensitive to DS, modality, and backbone choice. ID calibration does not reliably transfer, and medically aligned backbones are generally better calibrated than generic CLIP variants. It demonstrates that accuracy and calibration are not universally aligned, and no calibration or prompt-tuning method is consistently optimal across all settings. It shows that calibration varies with hard-prompt initialization and random seeds, highlighting the limitations of prompt initialization and the sensitivity of seed evaluation. Motivated by these findings, we proposed **Multi-Class Margin Regularization (MCM)**, which reduces ID calibration error and remains competitive under DS. Overall, MVC-Bench reveals that calibration should be treated as a first-class requirement for safe medical VLM deployment, evaluated through robustness, adaptation effectiveness, and stability under realistic variation.

7 Limitations

While MVC-Bench provides a large-scale calibration study of VLMs and Medical-VLMs, its scope is limited to medical image classification. We do not evaluate whether the same calibration behaviour extends to segmentation, detection, report generation, retrieval, or medical VQA, which require different uncertainty definitions and evaluation protocols. Furthermore, MCM targets low-margin underconfidence and is therefore not a universal solution to all calibration failures. Its improvements are less consistent in histopathology, particularly for PLIP-Histo. and QuiltNet-Histo., where overconfidence appears more prominent. Although we use a single hyperparameter setting across backbones and modalities, further work is needed to understand modality-specific margin behavior and calibration objectives under broader clinical shifts.

References

2018. Rsna pneumonia detection challenge. https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge. Accessed 2025-11-03.

Shihab Aaqil Ahamed, Udaya SKP Thanthrige, Ranga Rodrigo, and Muhammad Haris Khan. 2025. A-tpt: Angular diversity calibration properties for test-time prompt tuning of vision-language models. *arXiv preprint arXiv:2510.26441*.

Shuang Ao, Stefan Rueger, and Advaith Siddharthan. 2023. Two sides of miscalibration: Identifying over and under-confidence prediction for network calibration. *Preprint*, arXiv:2308.03172.

APTOS. 2019. APTOS 2019 Blindness Detection. https://www.kaggle.com/competitions/aptos2019-blindness-detection/data.

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

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

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

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309.

Zongjing Cao, Yan Li, Dong-Ho Kim, and Byeong-Seok Shin. 2024. Deep neural network confidence calibration from stochastic weight averaging. *Electronics*, 13(3).

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, and 1 others. 2022. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical image analysis*, 80:102485.

Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordoñez-Varela, Pascale Massin, Ali Erginay, and 1 others. 2014a. Feedback on a publicly distributed image database: The messidor database. *Image Analysis & Stereology*, pages 231–234.

Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, and 1 others. 2014b. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pages 231–234.

Chamuditha Jayanga Galappaththige, Gayal Kuruppu, and Muhammad Haris Khan. 2024. Generalizing to unseen domains in diabetic retinopathy classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7685–7695.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. 2019. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European congress on digital pathology*, pages 11–19. Springer.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(S1):1513–1589.

Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles R. G. Guttmann, Frank-Erik de Leeuw, Clare M. Tempany, Bram van Ginneken, Andriy Fedorov, Purang Abolmaesumi, Bram Platel, and William M. Wells. 2017. *Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation*, page 516–524. Springer International Publishing.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, and 1 others. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Asif Hanif, Fahad Shamshad, Muhammad Awais, Muzammal Naseer, Fahad Shahbaz Khan, Karthik Nandakumar, Salman Khan, and Rao Muhammad Anwer.

2024. Baple: Backdoor attacks on medical foundational models using prompt learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 443–453. Springer.

Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. 2022. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16081–16090.

Zhi Huang, Federico Bianchi, Mert Yüksekgönül, Thomas J Montine, and James Zou. 2023a. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29(9):2307–2316.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. 2023b. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316.

Noor Hussein, Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. 2024. Promptsmooth: Certifying robustness of medical vision-language models via prompt learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 698–708. Springer.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023a. Quilt-1m: One million image-text pairs for histopathology. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 37995–38017.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023b. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017.

Kaggle. 2015. Diabetic Retinopathy Detection. https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data.

Davood Karimi and Ali Gholipour. 2020. Improving calibration and out-of-distribution detection in medical image segmentation with convolutional neural networks. *arXiv preprint arXiv:2004.06569*.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, and 1 others. 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023a. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023b. Self-regulating prompts: Foundational model adaptation without forgetting. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15190–15200.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, and 1 others. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR.

Meelis Kull, Telmo Silva Filho, and Peter Flach. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pages 623–631. PMLR.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. 2018. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814. PMLR.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Max-Heinrich Laves, Sontje Ihler, Jacob F. Fast, Lüder A. Kahrs, and Tobias Ortmaier. 2020. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL)*, volume 121 of *Proceedings of Machine Learning Research*, pages 393–412. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. 2022a. The devil is in the margin: Margin-based label smoothing for network calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 80–88.

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. 2022b. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 80–88.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. 2024. Robust calibration of large vision-language adapters. In *European Conference on Computer Vision*, pages 147–165. Springer.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Mahdi Pakdaman Naeini and Gregory F Cooper. 2016. Binary classifier calibration using an ensemble of near isotonic regression models. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 360–369. IEEE.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR workshops*, volume 2.

Stephen Obadinma, Hongyu Guo, and Xiao-Dan Zhu. 2021. Class-wise calibration: A case study on covid-19 hate speech. *Proceedings of the Canadian Conference on Artificial Intelligence*.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*.

John Platt and 1 others. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, and 1 others. 2021. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319.

Sivaramakrishnan Rajaraman, Sudhir Sornapudi, Philip O Alderson, Les R Folio, and Sameer K Antani. 2020. Analyzing inter-reader variability affecting deep ensemble learning for covid-19 detection in chest radiographs. *PloS one*, 15(11):e0242301.

Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, and Muhammad Haris Khan. 2025. O-tpt: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19942–19951.

Ashshak Sharifdeen, Fahad Shamshad, Muhammad Akhtar Munir, Abhishek Basu, Mohamed Insaf Ismithdeen, Jeyapriyan Jeyamohan, Chathurika Sewwandi Silva, Karthik Nandakumar, and Muhammad Haris Khan. 2026. Towards calibrating prompt tuning of vision-language models. *Preprint*, arXiv:2602.19024.

Anouk Stein, Carol Wu, Chris Carr, George Shih, Jamie Dulkowski, J Kalpathy-Cramer, and et al. 2018. Rsna pneumonia detection challenge. Mountain View: Kaggle.

David Tellez, Geert J. S. Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in neural information processing systems*, 32.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1412.

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

Hantao Yao, Rui Zhang, and Changsheng Xu. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. 2024. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. *arXiv preprint arXiv:2403.14119*.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony Beardsworth Costa, Joseph J. Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133.

Ziyue Zhang, Tianjiao Wang, Yanwu Xu, Xiaosong Yang, Kang Zhou, Yutong Wang, Yunlu Zhang, Kailong Huang, Dong Tian, Eric Xing, and 1 others. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from 15 million scientific image–text pairs. *Preprint*, arXiv:2303.00915.

Hao Zheng, Shunzhi Yang, Zhuoxin He, Jinfeng Yang, and Zhenhua Huang. 2025a. Hierarchical cross-modal prompt learning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1891–1901.

Hao Zheng, Shunzhi Yang, Zhuoxin He, Jinfeng Yang, and Zhenhua Huang. 2025b. Hierarchical cross-modal prompt learning for vision-language models. *Preprint*, arXiv:2507.14976.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16816–16825.

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

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

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. 2023a. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15659–15669.

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. 2023b. Rethinking confidence calibration for failure prediction. *Preprint*, arXiv:2303.02970.

Appendices

In this supplementary material, we provide the following:

1. **Evaluation metrics and protocol** (Sec. A)
2. **Evaluated calibration baselines** (Sec. B)
3. **Loss component analyses** (Sec. C)
4. **Prompt-learning Comparision** (Sec. D)
5. **Prompt-template details** (Sec. E)
6. **Reliability diagram for other backbones** (Sec. F)
7. **Hyperparameter sweep** (Sec. G)
8. **Statistical Analysis of Domain-Shift Rank Transfer** (Sec. H)
9. **Paired Statistical Comparison of MCM** (Sec. I)
10. **Scale of the Experimental Evaluation**(Sec. K)
11. **Other Metrics Analysis on Research Questions**(Sec. J)
12. **Brier Score Analyses**(Sec. L)
13. **Weighted Subset ECE Analyses**(Sec. M)
14. **Results**(Sec. N)
15. **Declaration of LLM usage**(Sec. O)

A Evaluation metrics and protocol

The following metrics are used to evaluate the models' performance: Accuracy (Acc.), Expected Calibration Error (ECE), Maximum Calibration Error (MCE), Adaptive Calibration Error (ACE), and reliability diagram.

Accuracy (Acc) Proportion of correctly classified samples, indicating overall predictive performance.

Expected calibration error (ECE) (Naeini et al., 2015) The goal of calibration is to ensure the model's predicted probabilities accurately reflect the true likelihood of accuracy. If the outputs of the model are perfectly calibrated, then

$$\mathbb{P}\left(\hat{X} = X \mid \hat{P} = p\right) = p, \forall p \in [0, 1] \quad (3)$$

where X is the ground truth label, $\hat{X}$ is the predicted class label with its predicted confidence $\hat{P} = p$ that need to be calibrated. The ECE metric is computed as the weighted average of the absolute difference between the accuracy and the mean confidence within each bin. Mathematically, ECE is formulated as follows:

$$\mathrm{ECE} = \sum_{z=1}^{Z} \frac{|C_z|}{M} \left|\mathrm{acc}\left(C_z\right) - \mathrm{conf}\left(C_z\right)\right|, \quad (4)$$

where Z denotes the total number of bins. C_z defines the image set with predicted confidence falling into the z-th bin. $|C_z|$ is the number of images within this bin, and M is the total number of predictions across all the probability bins. Furthermore, $\mathrm{acc}\left(C_z\right)$, and $\mathrm{conf}\left(C_z\right)$ represent the accuracy of the predictions and average prediction confidence of all associated with bin z, respectively.

Maximum Calibration Error (MCE) (Naeini et al., 2015) This metric captures the maximum discrepancy between confidence and accuracy across all bins, highlighting the worst-case deviation. It is defined as:

$$\mathrm{MCE} = \max_{z \in \{1,\ldots,Z\}} \left|\mathrm{acc}(C_z) - \mathrm{conf}(C_z)\right|,$$

where the terms are as previously defined.

Adaptive Calibration Error (ACE) (Nixon et al., 2019) This metric addresses some limitations of ECE by using an adaptive binning approach, ensuring that each bin has an equal number of samples. The ACE is calculated as:

$$\mathrm{ACE} = \frac{1}{Z} \sum_{z=1}^{Z} \left|\mathrm{acc}(C_z) - \mathrm{conf}(C_z)\right|,$$

where Z is the number of bins, and other terms are as previously defined.

Reliability diagram also called the calibration curve, provides a qualitative description of calibration. With better calibration, the points will lie closer to the main diagonal that extends from the bottom left to the top right of the reliability diagram. The points below the diagonal indicate that the model is overconfident, with predicted probabilities that are too large. Those above the diagonal indicate that the model is underconfident, and the predicted probabilities are too small.

B Evaluated calibration baselines

The following mix of post-hoc and training-time calibration baselines is used in this benchmarking study:

Multi-class difference in confidence and accuracy (MDCA) (Hebbalaguppe et al., 2022) penalizes mismatches between the model's confidence and accuracy to capture class interactions. By adding this auxiliary loss, the model's predicted probabilities better align with its actual accuracy across multiple classes, thereby improving overall calibration.

Label smoothing (LS) (Müller et al., 2019) combats overfitting and overconfidence by replacing hard labels with a smoothed probability distribution. This discourages the model from assigning full probability to a single class, leading to more moderate and calibrated predictions.

Margin-based label smoothing (MBLS) (Liu et al., 2022b) is a class-balanced loss based on the number of samples to counter prevalence skew (e.g., the dominance of "No DR"), mitigating calibration bias induced by class imbalance.

ECCV_ZS (Murugesan et al., 2024) shifts the direction of the prior-weighted logit vector to match the correct category encoded in the one-hot label, and the logit norm normalizes logit values according to the ZS logit range.

ECCV_penalty (Murugesan et al., 2024) when the constraint is not satisfied, i.e., there exist logit magnitudes outside the zero-shot logit range, the value of the penalty term increases, backpropagating gradients to modify the logit values according to the enforced constraint.

Temperature scaling (Guo et al., 2017) adjusts the model's logits by a scalar, learnable temperature parameter $\tau > 0$ before the softmax layer, optimized on a held-out validation set by minimizing the neg-

ative log-likelihood (NLL). A zero-shot variant can be applied in scenarios where a validation set is not available.

C Loss component analyses

As shown in Fig. A1a, the inclusion of the mean term ($-\alpha\mu$) with CE loss drives a larger inter-class margin range. However, this explicit expansion leads to the overconfidence issue observed in Fig. A1b. In contrast, the variance term ($\beta\sigma^2_{\text{all-pairs}}$) restricts the margin range ($\approx$ 0–1), resulting in underconfidence. Combining both terms (MCM) balances these effects, producing a stable margin field and lower ECE.

D Prompt-learning Comparison

We compare five prompt-learning baselines (MaPLe, KgCoOp, PromptSRC, ProGrad, and HiCroPL) (Zhou et al., 2022b; Khattak et al., 2023a,b; Yao et al., 2023; Zhu et al., 2023a; Zheng et al., 2025b) on diabetic retinopathy, histopathology, and chest X-ray benchmarks using Accuracy and Expected Calibration Error (ECE). We report each method and its calibrated variant (suffix "_Ours") in Table. A1.

Our calibration reduces *average* ECE across each method, while retaining high accuracy. Per method, the average ECE drops from $12.44 \rightarrow 10.02$ for MaPLe, from $10.57 \rightarrow 8.28$ for KgCoOp from $15.64 \rightarrow 12.62$ for PromptSRC, from $18.59 \rightarrow 15.41$ for ProGrad, and from $5.45 \rightarrow 3.74$ for HiCroPL. Accuracy remains effectively unchanged: from $59.21 \rightarrow 58.14$ for MaPLe, from $60.08 \rightarrow 63.79$ for KgCoOp, from $60.98 \rightarrow 62.15$ for PromptSRC, from $57.59 \rightarrow 57.43$ for ProGrad, and from $72.09 \rightarrow 73.10$ for HiCroPL. (Table A1). Furthermore, Table. A2 shows that our method consistently reduces ECE in domain shift (DS) while retaining accuracy.

Our method provides substantial calibration gains without sacrificing accuracy cost, making it a practical drop-in for prompt-based VLM calibration.

E Prompt-template details

We consider seven hard prompt templates (P1-P7) that differ in phrasing, context, and domain. We have used the prompt template as follows: *"a photo of a", "a photo of a retina with", "a color fundus photograph showing diabetic retinopathy severity level as", "Classify diabetic retinopathy severity from this fundus photo of a", "You are classifying diabetic retinopathy on a fundus photo of a", "This picture illustrates diabetic retinopathy severity of a", "The DR level is".*

F Reliability diagram for other backbones

Fig.A2 shows that most of the other backbones have underconfidence issues across the modalities.

G Hyperparameter sweep

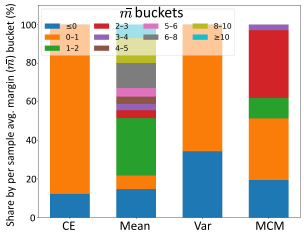
We tuned MCM on CLIP-ViT-B/16 using a DR benchmark(CLIP-DR-ViT-B/16), averaging three seeds. We swept $\alpha \in \{0.1, 0.2, 0.3\}$ and $\beta \in \{0.01, 0.05\}$ (Table A3). The best calibration was at (α, β)=(0.1,0.01) with Acc. 69.05 and ECE 3.43. This setting remains optimal when averaging over four DR datasets.

At fixed $\beta = 0.01$, increasing α raised accuracy but worsened calibration: $(0.1, 0.01) \rightarrow (0.3, 0.01)$: Acc. (69.05→70.49) alongside ECE (3.43→6.79). At fixed α, moving from $\beta = 0.01$ to 0.05 generally increased ECE and reduced accuracy (e.g., $\alpha = 0.1$: Acc $69.05 \rightarrow 68.29$, ECE $3.43 \rightarrow 5.73$); the exception was $\alpha = 0.3$, where both β settings remained poorly calibrated.

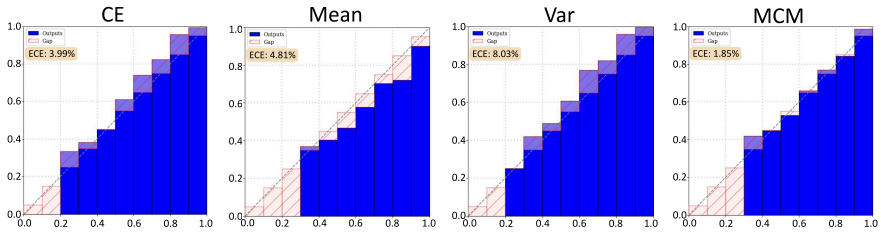
We select (0.1,0.01) as the default: it achieves the *lowest* ECE without compromising accuracy, providing the best overall balance. To verify that this choice does not depend on the target evaluation dataset, we additionally conduct a Leave One Dataset Out (LODO) hyperparameter-transfer analysis in Sec. G.1, where the same $(\alpha, \beta) = (0.1, 0.01)$ setting is independently selected in all four DR folds.

G.1 Leave One Dataset Out Hyperparameter Transfer

To verify that the selected MCM hyperparameters are not the result of calibration-oriented tuning on the final evaluation datasets, we perform a Leave One Dataset Out (LODO) analysis across the four DR datasets. For each fold, the target dataset is completely excluded from hyperparameter selection. We select (α, β) using only the validation splits of the remaining three DR datasets and then evaluate the selected setting on the held out target dataset. No final test labels or shifted-domain test labels are used during hyperparameter selection. As shown in Table A4, all four LODO folds inde-



(a) Each loss component analysis on margin range



(b) Each loss component analysis through the Reliability diagram

Figure A1: Loss components analysis

Method	Metric	APTOS	EyePACS	Messidor	Messidor-2	Pannuke	Kather	DigestPath	RSNA18	COVIDX	Average
MaPLe	Acc.	80.60	77.58	57.83	61.10	57.94	52.05	45.14	32.53	68.10	59.21
	ECE	3.97	3.02	5.86	8.56	10.77	16.21	32.39	11.55	19.60	12.44
MaPLe_Ours	Acc.	81.79	77.58	55.36	64.81	58.57	51.84	47.23	32.79	53.25	58.14
	ECE	3.09	2.92	5.46	7.66	10.58	14.45	27.70	7.08	11.26	10.02
KgCoOp	Acc.	76.31	73.99	58.20	63.97	57.94	52.05	45.14	43.88	69.21	60.08
	ECE	3.85	3.45	4.82	7.18	10.77	16.21	32.39	10.07	6.38	10.57
KgCoOp_Ours	Acc.	75.99	74.30	57.58	63.91	61.79	58.03	68.11	44.26	70.11	63.79
	ECE	2.68	3.63	4.34	4.91	16.79	8.91	18.01	9.69	5.57	8.28
PromptSRC	Acc.	84.42	77.85	60.61	65.19	53.48	79.59	12.31	40.78	74.55	60.98
	ECE	15.99	10.14	9.63	10.32	17.46	9.18	52.73	1.11	14.19	15.64
PromptSRC_Ours	Acc.	84.71	78.04	63.47	69.99	53.36	79.45	14.65	40.36	75.30	62.15
	ECE	9.24	3.74	7.60	7.38	18.37	1.87	50.44	1.12	13.84	12.62
ProGrad	Acc.	73.13	73.81	54.92	63.24	59.20	50.36	39.11	40.38	64.12	57.59
	ECE	27.67	27.16	13.72	23.77	19.15	4.75	30.86	9.04	11.16	18.59
ProGrad_Ours	Acc.	74.15	73.82	56.17	63.95	61.01	56.76	24.04	40.64	66.31	57.43
	ECE	27.78	23.95	13.33	22.57	5.96	4.03	23.56	8.59	8.94	15.41
HiCroPL	Acc.	86.78	77.57	66.20	71.45	73.16	85.29	72.57	40.44	75.32	72.09
	ECE	5.12	1.05	4.69	3.33	4.02	8.95	2.95	3.25	15.65	5.45
HiCroPL_Ours	Acc.	86.87	77.81	66.75	71.58	74.93	85.00	74.78	41.83	78.31	73.10
	ECE	1.36	1.83	4.14	2.01	3.91	1.42	2.26	1.99	14.75	3.74

Table A1: Accuracy (Acc.) and calibration (ECE) on diabetic retinopathy, histopathology, and chest X-ray benchmarks for different prompt-learning methods, with and without our calibration (suffix "_Ours"). Our calibration consistently lowers ECE while maintaining accuracy.

pendently select the same hyperparameter setting, $(\alpha, \beta) = (0.1, 0.01)$, as the default MCM configuration. Consequently, LODO-MCM exactly reproduces the Default-MCM ECE on every held-out dataset, with an average ECE of 3.43 compared with 4.44 for Base. This analysis shows that the selected hyperparameters are stable across the DR datasets and that the reported MCM results are not obtained by tuning (α, β) on the corresponding target evaluation dataset. The same selected setting is subsequently fixed across the remaining backbones, modalities, and domain-shift experiments without per-backbone or per-modality retuning.

H Statistical Analysis of Domain-Shift Rank Transfer

To further quantify the stability of calibration-method rankings under domain shift, we aggregate the Kendall rank-correlation results across all ID-to-DS evaluations. As summarized in Table A5, the analysis contains 30 ID-to-DS rank-transfer tests. Overall rank preservation is weak, with a mean Kendall correlation of -0.0598 and a median correlation of -0.0476. Among the 30 tests, 17 produce negative correlations, whereas 13 produce positive correlations, and none show significantly positive rank preservation at $p < 0.05$.

These results should not be interpreted causally as identifying a single source of distribution shift. The external datasets jointly differ in acquisition conditions, patient populations, class prevalence,

Method	Metric	APTOS	EyePACS	Messidor-2	Kather	DigestPath	COVIDX	Average
MaPLe	Acc.	46.64	67.61	57.09	20.61	36.95	67.01	49.32
	ECE	12.62	31.95	19.34	8.73	41.41	12.06	21.02
MaPLe_Ours	Acc.	46.11	68.55	57.19	20.58	37.22	65.31	49.16
	ECE	11.63	26.44	17.63	7.85	21.15	11.03	15.96
KgCoOp	Acc.	7.88	3.66	10.07	23.86	33.59	47.82	21.15
	ECE	54.06	60.10	42.84	59.06	52.39	14.10	47.09
KgCoOp_Ours	Acc.	31.34	40.50	35.47	22.83	20.69	47.21	33.01
	ECE	26.51	13.43	10.72	52.74	50.47	13.33	27.87
PromptSRC	Acc.	50.03	59.58	53.77	15.20	15.96	74.57	44.85
	ECE	33.73	23.82	13.54	4.91	53.02	14.03	23.84
PromptSRC_Ours	Acc.	49.23	58.88	54.05	15.90	15.94	75.29	44.88
	ECE	31.38	18.69	11.23	2.18	52.55	13.73	21.63
ProGrad	Acc.	44.03	69.18	53.17	26.19	11.64	64.70	44.82
	ECE	12.13	23.02	17.85	50.03	72.84	7.21	30.51
ProGrad_Ours	Acc.	43.38	68.92	52.27	25.98	13.43	67.72	45.28
	ECE	13.63	22.22	11.96	46.56	70.05	6.88	28.55
HiCroPL	Acc.	48.44	63.04	49.43	20.61	36.95	62.12	46.77
	ECE	24.04	6.96	3.31	8.73	41.41	3.53	14.66
HiCroPL_Ours	Acc.	47.57	62.96	49.77	26.03	42.65	62.19	48.53
	ECE	22.10	4.96	2.63	7.17	31.96	3.25	12.01

Table A2: Accuracy (Acc.) and calibration (ECE) under distribution shift. Messidor, Pannuke, and RSNA18 are the datasets used to train the model for the respective modalities' DS evaluation.

(α, β)	Acc.	ECE
(0.1, 0.01)	69.05	3.43
(0.2, 0.01)	70.15	4.15
(0.3, 0.01)	70.49	6.79
(0.1, 0.05)	68.29	5.73
(0.2, 0.05)	69.12	5.90
(0.3, 0.05)	69.69	6.53

Table A3: **Hyperparameter settings** we used α was chosen from 0.1 to 0.3, and β from 0.01 to 0.05. The backbone used was CLIP-ViT-B/16. Experiments were run across three seeds on a DR dataset. The average across 4 datasets indicated that the optimal hyperparameters were $(\alpha, \beta) = (0.1, 0.01)$.

dataset characteristics, and other distributional factors. Rather, the analysis evaluates whether calibration methods selected from in-domain validation retain their relative ranking when transferred to realistic external-domain evaluation. The absence of significantly positive rank preservation indicates that in-domain calibration rankings are generally not reliable predictors of their relative performance under such combined distributional changes.

I Paired Statistical Comparison of MCM

We further compare MCM against each calibration baseline using matched backbone–modality configurations. For every configuration, we compute

$$\Delta\text{ECE} = \text{ECE}_{\text{baseline}} - \text{ECE}_{\text{MCM}},$$

such that a positive value indicates lower ECE for MCM. We then apply a paired Wilcoxon signed-rank test across the 12 matched backbone–modality settings. Table A6 summarizes the resulting mean and median differences, the number of configurations in which MCM obtains lower or higher ECE, and the corresponding Wilcoxon p-values. The paired analysis provides additional statistical support for the MCM results. In ID settings, MCM obtains lower ECE in 10 of 12 matched configurations against Base, MDCA, LS, MBLS, ECCV_ZS, and TS, and in 11 of 12 configurations against ECCV_Penalty. The improvements are statistically significant against ECCV_ZS ($p = 0.0212$), ECCV_Penalty ($p = 0.0005$), and TS ($p = 0.0134$). Improvements relative to Base, MDCA, LS, and MBLS are numerically consistent but do not reach the $p < 0.05$ threshold. Under DS, MCM wins in 6–9 of 12 configurations depending on the baseline, but none of the paired comparisons are statistically significant. These results support our main conclusion that MCM provides consistent ID calibration gains while remaining competi-

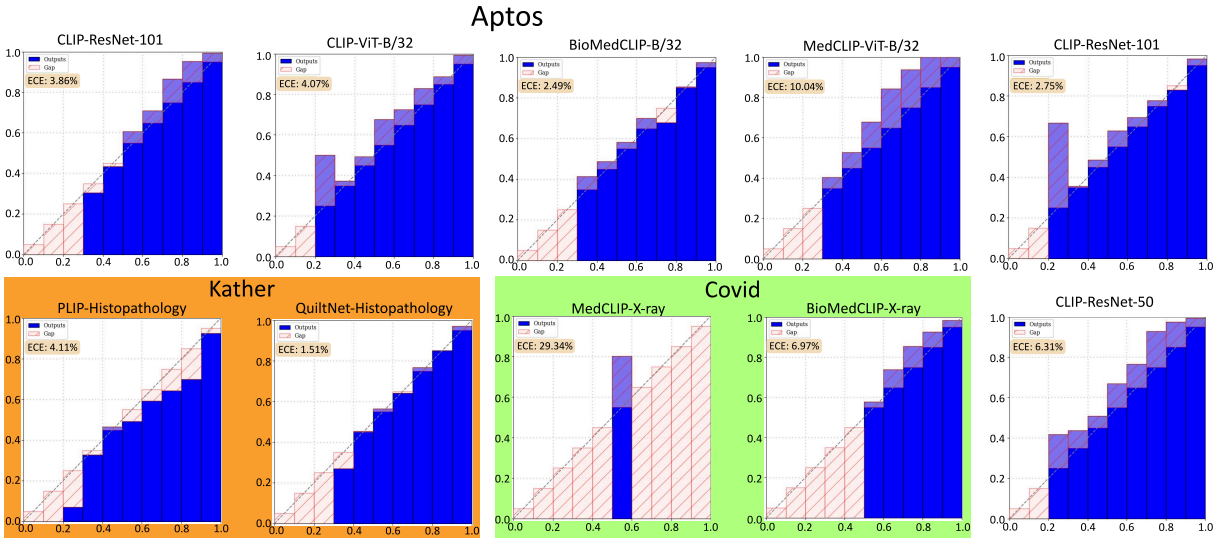


Figure A2: Reliability diagrams and Expected Calibration Error (ECE) analysis on medical datasets (Aptos, Kather, Covid). The diagrams present generic backbones (e.g., CLIP-ResNet-101, CLIP-ViT-B/32) and domain-specific models (e.g., PLIP, QuiltNet, MedCLIP, BioMedCLIP).

Held-out target	Hyperparameter-selection data	LODO-selected (α, β)	Base ECE	Default MCM ECE	LODO-MCM ECE
APTOS	EyePACS + Messidor + Messidor-2	(0.1, 0.01)	3.99	1.85	1.85
EyePACS	APTOS + Messidor + Messidor-2	(0.1, 0.01)	2.25	3.25	3.25
Messidor	APTOS + EyePACS + Messidor-2	(0.1, 0.01)	5.29	5.09	5.09
Messidor-2	APTOS + EyePACS + Messidor	(0.1, 0.01)	6.24	3.53	3.53
Average	–	–	4.44	3.43	3.43

Table A4: Leave-One-Dataset-Out (LODO) analysis of MCM hyperparameter selection across the four diabetic-retinopathy datasets. For each target dataset, the target is excluded entirely from hyperparameter selection, and (α, β) is selected using only the remaining three DR validation splits. Every fold independently selects $(\alpha, \beta) = (0.1, 0.01)$, and LODO-MCM reproduces the ECE obtained using the default MCM setting.

tive, rather than uniformly dominant, under domain shift.

J Scale of the Experimental Evaluation

To clarify the scale of MVC-Bench, we provide a detailed breakdown of the experimental evaluation used in the core calibration benchmark. Since our study evaluates calibration across multiple modalities, datasets, backbone families, calibration methods, and random seeds, the total number of experiments is obtained from the Cartesian product of these factors. We distinguish between three levels of counting: *experimental configurations*, *seed-level controlled runs*, and *metric-level evaluations*.

An *experimental configuration* denotes a unique combination of modality, dataset, backbone, and calibration method. A *seed-level controlled run* refers to one execution of an experimental configuration under a specific random seed. Since each configuration is repeated over three random seeds, the number of seed-level runs is three times the number of experimental configurations. Finally, a *metric-level evaluation* refers to the calibration metrics computed from each seed-level run. In the core calibration analysis, we evaluate each run using three binning-based calibration metrics: Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Adaptive Calibration Error (ACE).

J.1 In-Domain Evaluation

For the in-domain (ID) evaluation, we consider three medical imaging modalities: fundus imaging, chest X-ray, and histopathology. The fundus imaging setting includes eight backbones, four datasets, three random seeds, and eight calibration settings, including the vanilla Base method. This results in:

$$8 \times 4 \times 3 \times 8 = 768$$

seed-level controlled runs.

For chest X-ray evaluation, we use two backbones, two datasets, three random seeds, and eight calibration settings. This results in:

$$2 \times 2 \times 3 \times 8 = 96$$

Rank-transfer tests	Mean τ_b	Median τ_b	Non-positive τ_b	Negative τ_b	Positive τ_b	Significantly positive ($p < 0.05$)
30	-0.0598	-0.0476	17	17	13	0

Table A5: Aggregate statistical summary of ID-to-DS calibration-method rank transfer across 30 tests. Kendall's τ_b measures agreement between calibration-method rankings obtained on the in-domain source and corresponding external-domain datasets. Positive values indicate greater rank preservation, while negative values indicate rank reversals. No test exhibits significantly positive rank preservation at $p < 0.05$.

Setting	Baseline	Mean diff	Median diff	MCM wins /12	MCM losses /12	Wilcoxon p
ID	Base	0.8175	1.055	10	2	0.0881
ID	MDCA	0.7842	1.140	10	2	0.0881
ID	LS	2.4592	3.185	10	2	0.0549
ID	MBLS	0.8583	1.135	10	2	0.0881
ID	ECCV_ZS	7.9167	2.270	10	2	0.0212
ID	ECCV_Penalty	13.8900	15.840	11	1	0.0005
ID	TS	2.2217	2.915	10	2	0.0134
DS	Base	0.8175	0.985	8	4	0.2847
DS	MDCA	-0.5633	1.380	7	5	0.4548
DS	LS	-1.5650	0.605	8	4	0.4849
DS	MBLS	0.6475	1.775	8	4	0.2119
DS	ECCV_ZS	0.7033	1.480	7	5	0.4849
DS	ECCV_Penalty	1.1308	-0.050	6	6	0.4548
DS	TS	1.1100	2.795	9	3	0.1167

Table A6: Paired configuration-level comparison of MCM against existing calibration baselines across 12 matched backbone–modality settings. The difference is defined as $\Delta\mathrm{ECE} = \mathrm{ECE}_{\mathrm{baseline}} - \mathrm{ECE}_{\mathrm{MCM}}$, so positive values favor MCM. In ID settings, MCM obtains lower ECE in 10/12 configurations against most baselines and in 11/12 configurations against ECCV_Penalty. Under DS, MCM remains competitive but does not uniformly dominate the baselines.

seed-level controlled runs.

For histopathology, we use two backbones, three datasets, three random seeds, and eight calibration settings. This results in:

$$2 \times 3 \times 3 \times 8 = 144$$

seed-level controlled runs.

Therefore, the in-domain evaluation contains:

$$768 + 96 + 144 = 1008$$

seed-level controlled runs.

Equivalently, before accounting for random seeds, the ID evaluation contains:

$$(8 \times 4 \times 8) + (2 \times 2 \times 8) + (2 \times 3 \times 8) \\ = 256 + 32 + 48 = 336.$$

unique backbone–dataset–method configurations.

J.2 Domain-Shift Evaluation

For the domain shift (DS) evaluation, we evaluate calibration transfer from in-domain datasets to shifted datasets. In this setting, the zero-shot calibration setting is excluded, resulting in seven calibration methods. The fundus domain shift setting includes eight backbones, three shifted datasets, three random seeds, and seven calibration methods. This gives:

$$8 \times 3 \times 3 \times 7 = 504$$

seed-level controlled runs.

For chest X-ray domain shift evaluation, we use two backbones, one shifted dataset, three random seeds, and seven calibration methods. This gives:

$$2 \times 1 \times 3 \times 7 = 42$$

seed-level controlled runs.

For histopathology domain shift evaluation, we use two backbones, two shifted datasets, three random seeds, and seven calibration methods. This gives:

$$2 \times 2 \times 3 \times 7 = 84$$

seed-level controlled runs.

Therefore, the domain shift evaluation contains:

$$504 + 42 + 84 = 630$$

seed-level controlled runs.

Before accounting for random seeds, the DS evaluation contains:

$$(8 \times 3 \times 7) + (2 \times 1 \times 7) + (2 \times 2 \times 7) = 168 + 14 + 28 = 210$$

unique backbone–dataset–method configurations.

J.3 Total Scale of the Core Calibration Benchmark

Combining the ID and DS evaluations, the core MVC-Bench calibration benchmark contains:

$$336 + 210 = 546$$

unique backbone–dataset–method configurations.

Since each configuration is repeated over three random seeds, the total number of seed-level controlled runs is:

$$546 \times 3 = 1638.$$

Thus, the core calibration benchmark consists of 1,638 seed-level controlled runs. Since each run is evaluated using ECE, MCE, and ACE, the total number of binning-based calibration metric evaluations is:

$$1638 \times 3 = 4914.$$

In addition to these calibration metrics, accuracy is also computed for each seed-level run to analyze the relationship between predictive performance and calibration reliability. Therefore, the reported experimental scale supports both accuracy-based and calibration-based analysis across in-domain and domain shift settings.

J.4 Summary of Experimental Scale

Table A7 summarizes the experiment count across modalities and evaluation settings.

J.5 Scope of the Count

The above count refers only to the core calibration benchmark used to evaluate in-domain and domain shift calibration across backbones, datasets, calibration methods, and seeds. It does not include analyses such as prompt-tuning comparisons, Brier score evaluation, class-wise ECE analysis, or other supplementary diagnostic experiments. Therefore, the reported count of 1,638 seed-level controlled runs should be interpreted as a conservative estimate of the core experimental scale of MVC-Bench.

K Other Metrics Analysis on Research Questions

To provide a more comprehensive calibration assessment, we further evaluate our research questions using binning-based metrics, including Adaptive Calibration Error (ACE) (Nixon et al., 2019) and Maximum Calibration Error (MCE) (Naeini et al., 2015), in addition to ECE.

K.1 Robustness: How do domain shifts affect calibration

In Fig.A3 and Fig.A4, the ACE and MCE domain-shift analyses show that calibration transfer from in-domain to shifted datasets is metric-dependent, but domain shift still introduces clear instability. For ACE, the retinopathy setting shows several positive Kendall rank correlations, especially for CLIP-DR-ViT-B/16 and CLIP-DR-ViT-B/32, suggesting that adaptive-bin calibration rankings are partly preserved across APTOS, EyePACS, and Messidor-2. Histopathology also shows moderate consistency for both QuiltNet and PLIP, while the X-ray shift reveals a strong contrast: MedCLIP remains positively correlated, whereas BioMedCLIP shows negative transfer under COVID shift. For MCE, retinopathy again shows stronger rank preservation for several CLIP-based backbones, particularly on Messidor-2, indicating that worst-case calibration behaviour is sometimes stable across related fundus datasets. However, MCE exposes weaker or negative transfer in histopathology, especially for PLIP, and again shows poor transfer for BioMedCLIP under X-ray shift. Overall, compared with the ECE-based main analysis, ACE and MCE suggest that some calibration rankings can remain stable under closer domain shifts, but this stability is not universal across modalities, backbones, or metrics. This reinforces the takeaway that calibration should be evaluated under domain shift using multiple metrics, since average, adaptive-bin, and worst-case calibration errors may reveal different failure patterns. An aggregate statistical analysis over all 30 ID to DS rank transfer tests is provided in Sec. H it shows a median Kendall $\tau_b = -0.0476$, with 17/30 negative correlations and no significantly positive rank-preservation case at $p < 0.05$.

Setting	Modality	Backbones	Datasets	Methods	Seed-level Runs
ID	Fundus	8	4	8	768
ID	Chest X-ray	2	2	8	96
ID	Histopathology	2	3	8	144
DS	Fundus	8	3	7	504
DS	Chest X-ray	2	1	7	42
DS	Histopathology	2	2	7	84
		Total seed-level controlled runs			**1638**

Table A7: Breakdown of the core MVC-Bench experimental scale. ID denotes in-domain evaluation and DS denotes domain-shift evaluation. The zero-shot calibration setting is excluded from the DS count, resulting in seven calibration methods for DS evaluation.

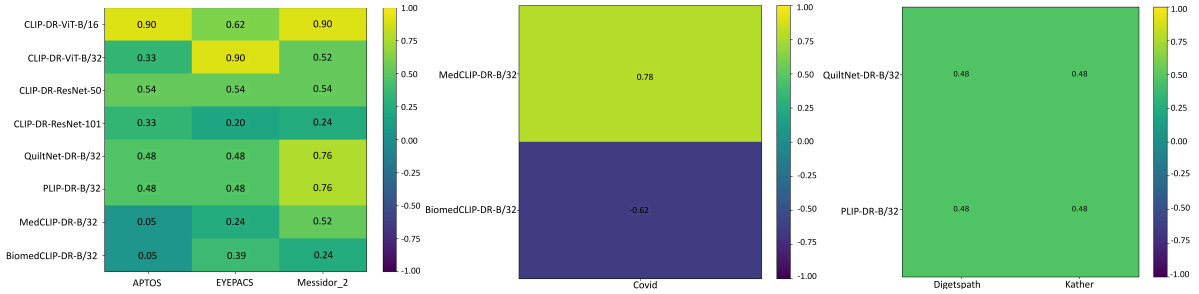


Figure A3: Kendall rank correlation (τ) between ACE-based method rankings on the In-domain(ID) source and each Domain shift(DS) dataset is shown for every backbone (rows = backbones, columns = DS sets). Cells are colored by τ (-1 to 1) and annotated with values: higher/positive τ means rankings (and thus relative calibration quality) are preserved under shift, while near-zero/negative τ indicates rank reversals of the calibration under shift.

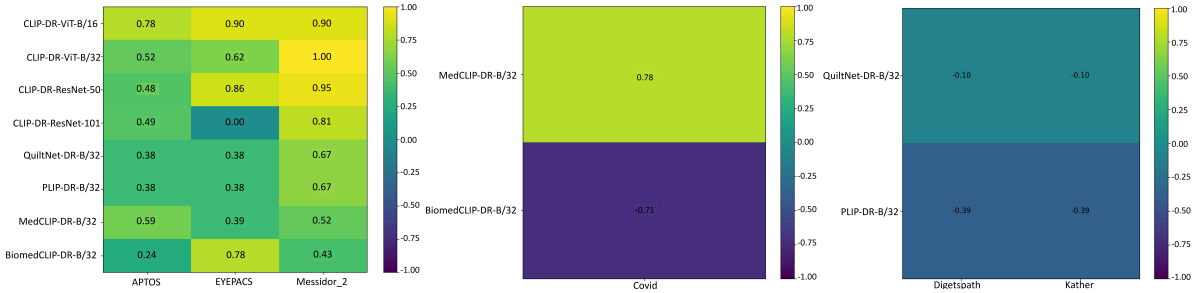


Figure A4: Kendall rank correlation (τ) between MCE-based method rankings on the In-domain(ID) source and each Domain shift(DS) dataset is shown for every backbone (rows = backbones, columns = DS sets). Cells are colored by τ (-1 to 1) and annotated with values: higher/positive τ means rankings (and thus relative calibration quality) are preserved under shift, while near-zero/negative τ indicates rank reversals of the calibration under shift.

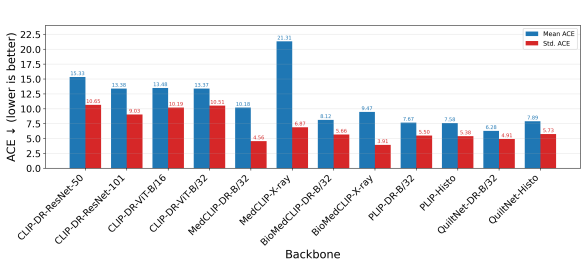


Figure A5: Comparison of mean and Std. of ACE for various VLM-based backbones across different medical modality inputs

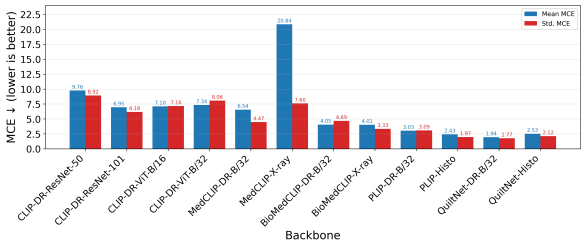


Figure A6: Comparison of mean and Std. of MCE for various VLM-based backbones across different medical modality inputs

K.2 Robustness: Which VLM-based backbones are better calibrated

The ACE and MCE analyses from Fig.A5 and Fig.A6 further confirm that backbone choice

strongly affects calibration, consistent with the ECE-based finding in the main paper that calibration varies substantially across VLM and Medical-VLM backbones. Under MCE, the best worst-case calibration behaviour is observed for QuiltNet-DR-B/32, PLIP-Histo, PLIP-DR-B/32, and QuiltNet-Histo, all showing low mean MCE values below approximately 3%, while BioMedCLIP-X-ray and BioMedCLIP-DR-B/32 also remain relatively stable. In contrast, MedCLIP-X-ray shows the poorest calibration, with a very high mean MCE of 20.84%, indicating severe worst-bin calibration failures. A similar pattern is visible in ACE, where QuiltNet, PLIP, and BioMedCLIP variants generally achieve lower adaptive calibration error than generic CLIP backbones, while MedCLIP-X-ray again reports the highest ACE of 21.31%. Generic CLIP variants show relatively high ACE values, especially CLIP-DR-ResNet-50, CLIP-DR-ResNet-101, and ViT-based CLIP models, suggesting that natural-image pretraining is less reliable for calibrated confidence in medical settings. Overall, ACE and MCE support the main ECE observation that domain-specific or medically aligned backbones are generally better calibrated than generic CLIP backbones, but they also reveal that some models, particularly MedCLIP-X-ray, can suffer from severe average and worst-case calibration instability.

K.3 Effectiveness: How does calibration vary with accuracy for different calibration methods?

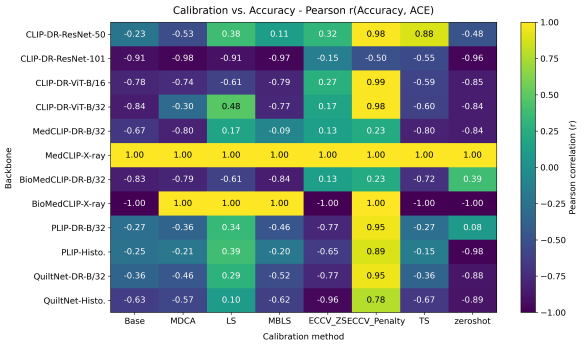


Figure A7: Pearson correlation matrix between accuracy and ACE evaluated across diverse calibration methods and backbones.

Fig.A7 and Fig.A8 show that ACE and MCE findings are Consistent with ECE-based findings. The ACE and MCE Pearson correlation analyses show that the relationship between accuracy and calibration is not universal, but strongly depends on the calibration method and backbone. In Fig.4,

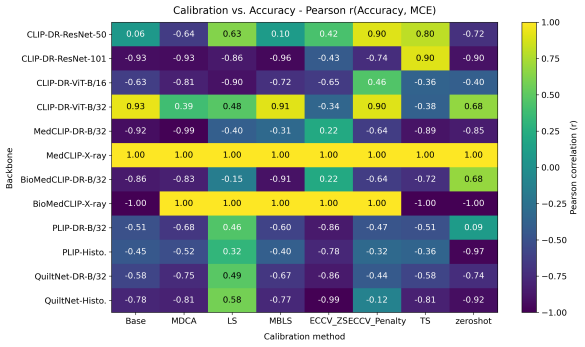


Figure A8: Pearson correlation matrix between accuracy and MCE evaluated across diverse calibration methods and backbones.

the ECE-based analysis showed that some methods, such as Base, MDCA, MBLS, and TS, often produced negative correlations, meaning that higher accuracy was generally associated with lower calibration error, while ECCV_Penalty frequently showed positive correlations, indicating an accuracy–calibration trade-off. The ACE results follow a broadly similar pattern, where Base, MDCA, MBLS, and zeroshot mostly show negative correlations across many backbones, suggesting that accuracy improvement often aligns with better adaptive-bin calibration, whereas ECCV_Penalty shows strong positive correlations for several CLIP-based backbones, meaning that higher accuracy can increase ACE and therefore worsen calibration. The MCE results also support the same conclusion, but with complementary observations in ECCV_Penalty across some of the backbones, because the difference arises where ACE summarizes calibration behaviour across bins, while MCE is dominated by the single worst bin. Therefore, ACE and MCE reveal complementary aspects of calibration, ACE captures broader calibration degradation, while MCE highlights whether the most severe confidence error is amplified or reduced. Overall, compared with the ECE-based findings, the ACE and MCE analyses confirm the same key takeaway that improving accuracy does not necessarily guarantee better calibration.

K.4 Effectiveness: Which calibration methods show consistent calibration performance

Fig. A9 illustrates ACE results across backbones. In the ID setting, **Ours** achieves the highest N_{best} ($= 8$), indicating strong consistency, while other methods show limited improvement and higher variability. Under DS, performance deteriorates

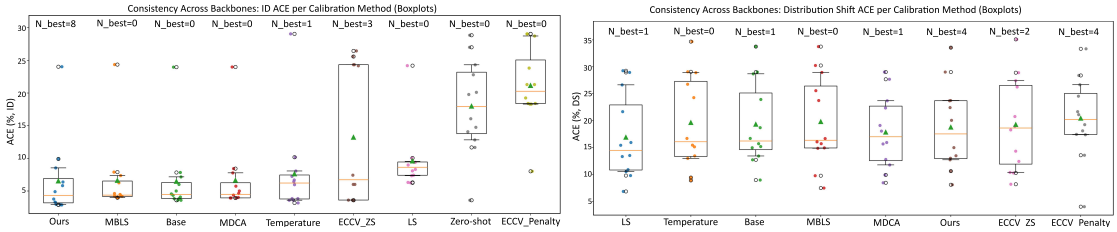


Figure A9: ACE (%) across backbones for each calibration method under ID (top) and DS (bottom) settings. Boxplots show the interquartile range (IQR), whiskers indicate the 10–90% range, and individual backbone results are overlaid as colored markers. N_{best} denotes the number of backbones for which a method achieves the lowest ACE. Lower medians and tighter distributions indicate more consistent calibration across architectures.

across all methods, though **Ours** and ECCV_ZS obtain the highest N_{best} counts.

Fig. A10 shows MCE results, where **Ours** leads in ID ($N_{\text{best}} = 8$) and DS ($N_{\text{best}} = 5$), followed by ECCV_ZS.

Overall, the results are illustrating consistent pattern across all metrics for all the methods for both ID and DS. Furthermore, our method demonstrates superior ID calibration consistency across metrics and maintains competitive DS performance compared to existing methods.

K.5 Effectiveness: How do different prompt-tuning methods interact with calibration?

Fig.A11 shows that ACE and MCE have similar trend as how ECE based observation for how different prompt-tuning families behave on calibration. HiCroPL provides the most stable calibration, achieving the lowest ACE 3.53% and MCE 1.06%, followed by CoOp and KgCoOp, which also maintain relatively low average and worst-case calibration errors.

K.6 Stability: How does hard-prompt initialization affect calibration?

In Fig.A12 and Fig.A13, the ACE and MCE analyses show that calibration is clearly sensitive to hard-prompt initialization. For MCE, early templates such as P1 and P2 are the most stable, with low standard deviations of 0.80 and 1.02, and low mean MCE values of 1.45 and 1.87, respectively. A similar trend is observed for ACE, where P1 and P2 also show the lowest variation, while later templates become increasingly unstable. In particular, P6 and P7 show much larger variations in both metrics, with P7 producing the highest MCE variation 3.70 and ACE variation 4.28, suggesting that poorly aligned or noisy prompt initializations can strongly destabilize calibration across datasets. Overall, both ACE and MCE support the ECE-based finding that prompt initialization is not a minor design choice, it directly affects both average calibration behaviour and worst-case confidence reliability.

K.7 Stability: How do different seed initializations affect calibration performance?

In Fig.A14 and Fig.A15, the ACE and MCE analyses show that calibration stability differs substantially across backbones when evaluated beyond ECE. Most backbones exhibit relatively low MCE variability, such as CLIP-DR-ResNet-50 0.22, BioMedCLIP-X-ray 0.28, CLIP-DR-ViT-B/32 0.28, PLIP-DR-B/32 0.38, and QuiltNet-DR-B/32 0.38, indicating that their worst-case calibration error remains comparatively stable. A similar trend is observed in ACE, where BioMedCLIP-X-ray 0.86 and CLIP-DR-ViT-B/32 0.78 show strong stability, suggesting that their adaptive-bin calibration behaviour is less sensitive. However, MedCLIP-based models show much higher variability across all the metrics, especially MedCLIP-X-ray, which records the largest ACE and MCE variability 5.37, followed by MedCLIP-DR-B/32 with ACE 2.68 and MCE 2.67. This indicates that some backbones may appear acceptable under average calibration metrics but remain unstable when calibration is evaluated across datasets or initialization conditions. Overall, ACE and MCE support the ECE-based finding that calibration performance is sensitive to seed initialization and backbone choice, with domain-aligned models such as BioMedCLIP, PLIP, and CLIP-DR-ViT-B/32 generally showing more stable behaviour across ECE, ACE, and MCE than MedCLIP, while worst-case calibration failures are most visible through all the metrics.

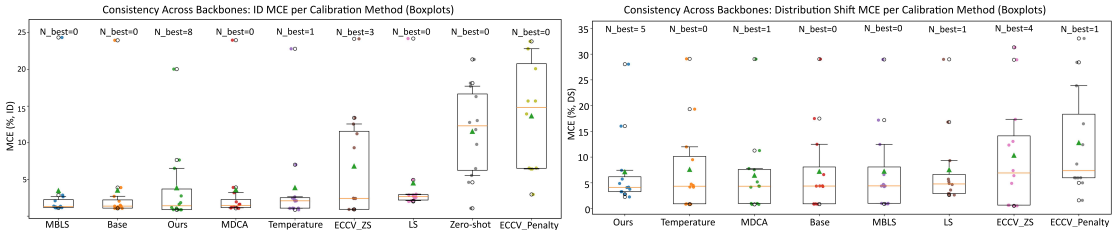


Figure A10: MCE (%) across backbones for each calibration method under ID (top) and DS (bottom) settings. Boxplots show the interquartile range (IQR), whiskers indicate the 10–90% range, and individual backbone results are overlaid as colored markers. N_{best} denotes the number of backbones for which a method achieves the lowest MCE. Lower medians and tighter distributions indicate more consistent calibration across architectures.

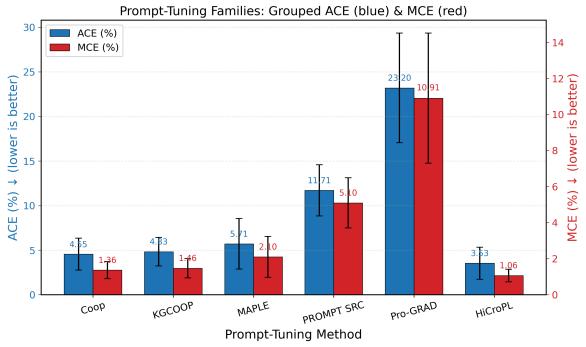


Figure A11: Comparison of ACE and MCE for different prompt-tuning methods.

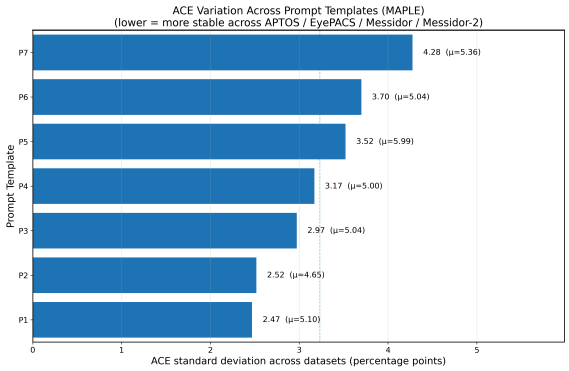


Figure A12: Comparison of Std. and mean of ACE (μ) across datasets for seven different hard-prompt templates (P1-P7). Templates P1-P4 have low and stable ACE across datasets, while P7 has a high variance. This highlights that the prompting method is sensitive to hard-prompt initialization.

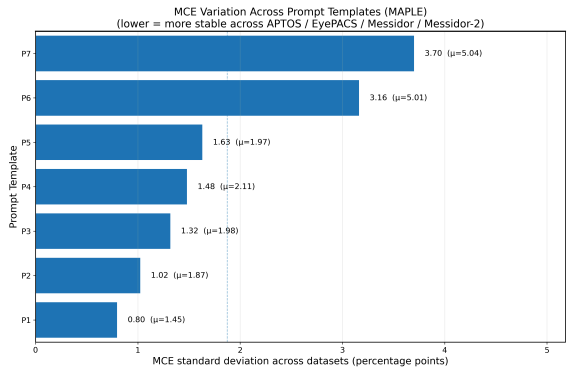


Figure A13: Comparison of Std. and mean of MCE (μ) across datasets for seven different hard-prompt templates (P1-P7). Templates P1-P4 have low and stable MCE across datasets, while P7 has a high variance. This highlights that the prompting method is sensitive to hard-prompt initialization.

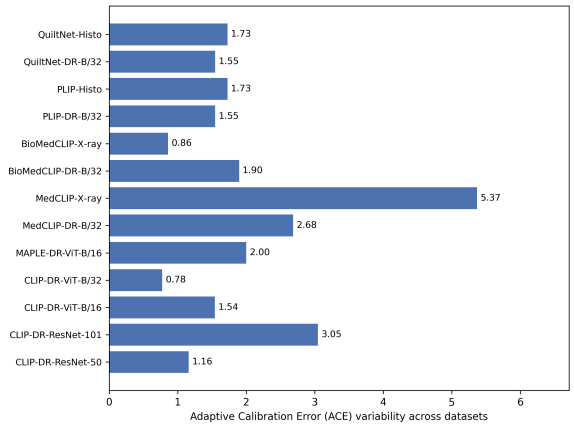


Figure A14: Comparison of seed variability, i.e., Std. of ACE, for various backbones across three random seeds.

L Brier Score Analyses

Brier Score (BS) (Brier, 1950) is a bin-free proper scoring rule that measures the mean squared difference between the predicted probability distribution and the true one-hot encoded class label. Unlike ECE-based metrics, it does not depend on confidence bins and therefore provides a complementary assessment of probabilistic prediction quality. For a multi-class classification problem, it is defined as:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}(p_{ic} - y_{ic})^2\,,$$

where N is the number of samples, C is the number of classes, p_{ic} is the predicted probability assigned to class c for sample i, and $y_{ic} \in \{0, 1\}$ is the corresponding one-hot ground-truth label. A lower

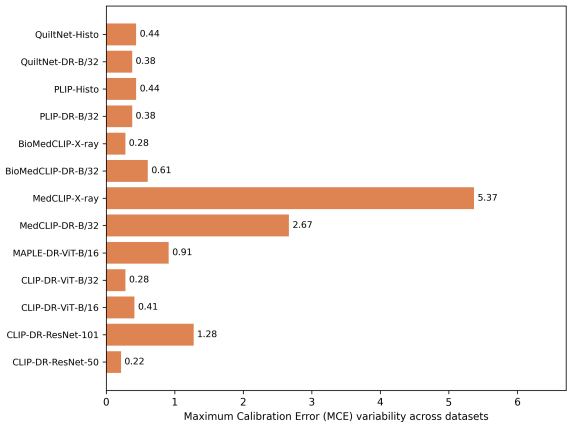


Figure A15: Comparison of seed variability, i.e., Std. of MCE, for various backbones across three random seeds.

Brier Score indicates better probabilistic prediction quality.

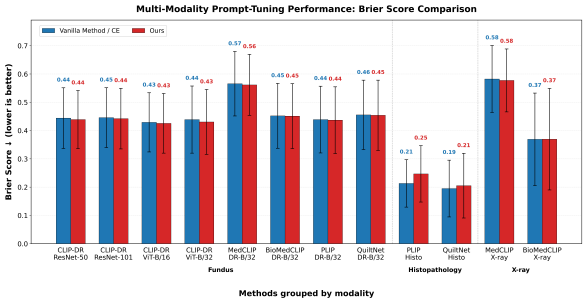


Figure A16: Brier Score Analyses across Backbones and Modalities.

In Fig.A16, the Brier score analysis shows that integrating our loss function generally preserves or slightly improves probabilistic prediction quality compared with the vanilla CE baseline across most backbones, especially in the fundus modality. For fundus models, our method gives small but consistent reductions in Brier score for several backbones, such as CLIP-DR-ResNet-101, CLIP-DR-ViT-B/32, MedCLIP-DR-B/32, and QuiltNet-DR-B/32, while maintaining similar performance for CLIP-DR-ResNet-50, CLIP-DR-ViT-B/16, BioMedCLIP-DR-B/32, and PLIP-DR-B/32. This suggests that the proposed loss improves calibration-related probability quality without substantially harming predictive behaviour. In the X-ray modality, the Brier scores remain almost unchanged for both MedCLIP-X-ray and BioMedCLIP-X-ray, indicating that our loss preserves the baseline probabilistic performance but does not produce a strong improvement. However, in histopathology, the Brier score slightly increases for PLIP-Histo and QuiltNet-Histo, suggesting that the benefit of the proposed loss is less consistent in this modality. Overall, the Brier score results support the main calibration findings by showing that our loss is mostly stable across modalities, with the strongest benefit in fundus datasets, while also highlighting that modality-specific behaviour should be considered when applying calibration-oriented training objectives.

M Weighted Subset ECE Analyses

Weighted Subset ECE (WSECE) (Obadinma et al., 2021). Weighted Subset ECE measures calibration separately for each class and then combines the class-level calibration errors using the number of samples in each class as weights. Therefore, classes with more test samples contribute more to the final score, while smaller classes still remain explicitly evaluated. This metric is useful for imbalanced datasets because it shows whether the model is well calibrated across different classes, rather than only reporting one overall ECE value.

For class c, let ECE_c be the ECE computed using only the samples that belong to class c. If N_c is the number of samples in class c, and N is the total number of samples, WSECE is defined as

$$\mathrm{WSECE} = \sum_{c=1}^{K} \frac{N_c}{N} \mathrm{ECE}_c,$$

where K is the number of classes. A lower WSECE indicates better class-wise calibration.

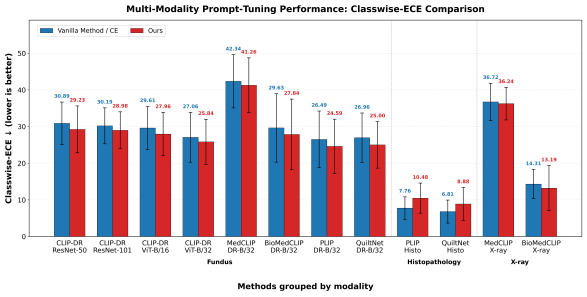


Figure A17: Weighted Subset ECE Analyses across Backbones and Modalities.

In Fig.A17, the WSECE analysis shows that integrating our loss function generally improves class-wise calibration compared with the vanilla CE baseline, especially in the fundus and X-ray modalities. In fundus datasets, class imbalance is high, for example, the "No DR" category represents up to 74% of the EyePACS dataset. Our method consistently reduces WSECE across all

backbones, including CLIP-DR-ResNet-50, CLIP-DR-ResNet-101, CLIP-DR-ViT-B/16, CLIP-DR-ViT-B/32, MedCLIP-DR-B/32, BioMedCLIP-DR-B/32, PLIP-DR-B/32, and QuiltNet-DR-B/32, indicating that the proposed loss helps align confidence more reliably across class-specific distributions. In X-ray, the improvement is smaller but still visible, particularly for BioMedCLIP-X-ray, where WSECE decreases from 14.31 to 13.19, while MedCLIP-X-ray shows only a slight reduction. However, in histopathology, our method increases WSECE for both PLIP-Histo and QuiltNet-Histo, suggesting that the proposed loss is less effective for class-wise calibration in this modality, possibly due to stronger visual heterogeneity or class-level ambiguity. Overall, the WSECE results support the broader calibration findings by showing that our loss improves class-wise calibration in most fundus and X-ray settings, while also highlighting that calibration gains are modality-dependent and should be evaluated separately across class-imbalanced medical datasets.

N Results

We evaluate three medical imaging modalities, diabetic retinopathy (APTOS (APTOS, 2019), EyePACS (Kaggle, 2015), Messidor, Messidor-2 (Decencière et al., 2014b)), histopathology (DigestPath (Da et al., 2022), Kather (Kather et al., 2019), PaNuke (Gamper et al., 2019)), and chest X-ray (COVIDX (Rahman et al., 2021), RSNA18 (Stein et al., 2018)) across 12 modality-specific configurations (CLIP-ResNet/ViT (Radford et al., 2021), MedCLIP (Wang et al., 2022), BioMedCLIP (Zhang et al., 2023), PLIP (Huang et al., 2023a), QuiltNet (Ikezogwo et al., 2023a)). We report accuracy and calibration with Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Adaptive Calibration Error (ACE). All results are averaged across three seeds. Complete results are provided in Tab. A8 and Tab. A9.

N.1 In-domain (ID) performance

Diabetic retinopathy (DR) MCM improves calibration in CLIP backbones:

- CLIP-DR-ResNet-50 (avg over 4 datasets): Acc. (67.87 → 67.71), ECE 4.40 → 2.96, MCE 1.49 → 1.11, ACE 5.00 → 3.26
- CLIP-DR-ResNet-101: Acc. 68.29 → 68.19, ECE 5.82 → 4.82, MCE 2.04 → 1.62, ACE 6.01 → 4.90
- CLIP-DR-ViT-B/16: Acc. 69.18 → 69.05, ECE 4.44 → 3.43, MCE 1.37 → 1.18, ACE 4.56 → 3.77
- CLIP-DR-ViT-B/32: Acc. 68.13 → 67.79, ECE 3.89 → 2.77, MCE 1.35 → 0.90, ACE 4.35 → 2.89

Medical vision-language models show the same trend, e.g., MedCLIP-DR-B/32 ECE 6.69 → 4.94 without compromising accuracy.

Chest X-ray

- MedCLIP-X-ray (avg over 2 datasets): Acc. 65.77 → 65.83, ECE 23.97 → 20.03, MCE/ACE also drop.
- BioMedCLIP-X-ray: Acc. 73.56 → 73.70, ECE 7.78 → 5.71.

Histopathology Calibration gains are not consistent:

- PLIP-Histo (avg over 3 datasets): Acc. (85.43 → 85.30) but ECE increases (7.00 → 9.92).
- QuiltNet-Histo (avg): Acc. (86.73 → 86.87) with higher ECE (6.21 → 8.55).

N.2 Domain-shift (out-domain) performance

We compare domain-shifted (DS) variants.

- CLIP-DR-ViT-B/16-DS: Acc. 32.90 → 33.36, ECE 33.69 → 20.08
- CLIP-DR-ViT-B/32-DS: Acc. 43.58 → 46.59, ECE 15.83 → 12.44.
- MedCLIP-DR-B/32-DS: Acc. 52.69 → 59.83, ECE 23.95 → 23.54.
- BioMedCLIP-X-ray-DS: Acc. 74.56 → 78.84, ECE 9.00 → 8.02.

Histology backbones can degrade under shift:

- PLIP-DR-B/32-DS: Acc. 55.40 → 47.16, ECE 12.79 → 13.56.
- QuiltNet-DR-B/32-DS: Acc. 40.08 → 27.92, ECE 15.42 → 23.72.

MCM provides consistent calibration gains on DR and CXR in ID and remains competitive in DS.

Backbone	Dataset	Metric	Base	MDCA	LS	MBLS	ECCV_ZS	ECCV_Penalty	TS	zeroshot	Ours (MCM)
CLIP-DR-ResNet-50	APTOS	Acc.	76.69	76.19	76.64	76.82	75.34	66.72	78.59	8.62	76.59
		ECE	6.31	5.11	9.87	7.11	40.82	28.60	3.44	26.47	1.99
		MCE	1.49	1.37	2.37	2.01	30.20	25.01	25.01	18.84	0.48
		ACE	6.26	5.15	9.81	7.00	39.31	33.31	33.31	26.47	2.65
	EyePACS	Acc.	74.48	74.06	74.02	74.12	74.40	73.34	74.57	7.39	74.09
		ECE	2.16	2.06	3.58	2.00	12.89	38.52	7.22	20.41	3.60
		MCE	1.66	0.79	1.99	0.67	3.06	26.81	26.81	13.70	1.84
		ACE	3.24	2.39	3.84	2.22	13.08	38.52	38.52	20.41	3.60
	Messidor	Acc.	56.12	54.44	57.19	57.06	75.34	44.70	55.69	20.58	56.20
		ECE	4.88	4.43	5.06	4.48	39.31	14.65	6.42	18.83	4.06
		MCE	1.66	1.68	1.97	1.53	13.62	13.96	13.96	9.80	1.52
		ACE	5.79	5.79	5.36	4.97	24.93	14.70	14.70	18.83	4.37
	Messidor-2	Acc.	64.20	63.45	63.95	64.07	63.28	58.31	63.97	13.46	63.95
		ECE	4.26	3.84	5.97	3.06	20.00	29.68	6.05	14.76	2.19
		MCE	1.13	1.17	1.71	0.86	6.73	25.42	25.42	9.77	0.58
		ACE	4.69	4.10	6.24	3.66	20.00	29.68	29.68	14.76	2.41
	Average	Acc.	67.87	67.04	67.95	68.02	72.09	60.77	68.21	12.51	67.71
		ECE	4.40	3.86	6.12	4.16	28.26	27.86	5.78	20.12	2.96
		MCE	1.49	1.25	2.01	1.27	13.40	22.80	22.80	13.03	1.11
		ACE	5.00	4.36	6.31	4.46	24.33	29.05	29.05	20.12	3.26
CLIP-DR-ResNet-101	APTOS	Acc.	75.75	75.12	75.02	75.78	74.81	57.18	75.75	8.03	75.43
		ECE	3.86	3.80	7.18	3.76	38.90	24.53	5.49	33.49	3.40
		MCE	1.12	1.12	1.64	1.11	18.21	14.78	3.30	19.20	1.04
		ACE	3.95	4.11	7.32	3.77	38.90	25.64	5.77	33.46	3.54
	EyePACS	Acc.	73.92	73.93	73.88	73.95	74.14	73.47	73.93	2.59	73.95
		ECE	2.89	3.08	4.81	2.98	9.15	34.81	7.96	32.94	2.92
		MCE	0.84	0.93	2.83	0.87	1.98	2.85	2.85	22.17	0.86
		ACE	3.10	3.14	5.15	3.06	9.16	7.86	7.86	32.94	2.92
	Messidor	Acc.	59.47	57.67	59.31	59.50	58.21	45.50	59.47	20.67	59.08
		ECE	10.67	9.09	10.90	10.67	29.36	16.91	6.87	11.31	8.97
		MCE	4.10	3.87	4.14	4.04	19.00	16.49	2.02	11.30	3.35
		ACE	10.95	9.03	10.85	10.92	29.36	16.99	7.58	11.31	9.13
	Messidor-2	Acc.	64.03	63.68	64.45	64.11	63.59	58.31	64.01	2.01	64.28
		ECE	5.86	6.58	8.84	7.01	24.95	26.53	6.31	37.70	4.00
		MCE	2.11	1.87	2.91	2.37	11.04	21.58	1.69	18.16	1.24
		ACE	6.03	6.68	9.01	7.25	24.95	26.53	7.76	37.70	4.01
	Average	Acc.	68.29	67.60	68.17	68.34	67.69	58.62	68.29	8.33	68.19
		ECE	5.82	5.64	7.93	6.11	25.59	25.70	6.66	28.86	4.82
		MCE	2.04	1.95	2.88	2.10	12.56	13.93	2.47	17.71	1.62
		ACE	6.01	5.74	8.08	6.25	25.59	19.26	7.24	28.85	4.90
CLIP-DR-ViT-B/16	APTOS	Acc.	77.90	77.45	77.84	77.94	76.14	56.78	77.90	7.95	77.50
		ECE	3.99	4.87	7.88	4.09	38.96	24.87	4.44	29.78	1.85
		MCE	1.46	1.46	1.81	1.38	12.54	13.95	1.86	17.20	0.61
		ACE	3.99	4.84	7.90	4.08	38.96	26.23	4.74	29.78	2.03
	EyePACS	Acc.	74.64	74.65	74.53	74.64	74.85	72.71	74.64	3.21	74.53
		ECE	2.25	2.31	4.36	2.25	19.06	41.84	7.53	24.38	3.25
		MCE	0.68	0.78	2.23	0.72	5.45	21.84	3.24	19.22	1.08
		ACE	2.30	2.29	4.29	2.33	18.80	41.84	7.47	24.38	3.25
	Messidor	Acc.	59.50	57.17	59.61	59.53	58.31	43.00	59.50	21.00	59.72
		ECE	5.29	6.02	9.15	5.31	26.87	14.80	6.95	14.72	5.09
		MCE	1.78	2.17	2.53	1.82	17.43	12.69	2.91	14.55	1.73
		ACE	5.69	6.55	9.18	5.81	26.87	14.91	7.08	14.72	5.89
	Messidor-2	Acc.	64.66	64.51	64.96	64.62	64.95	58.26	64.66	2.18	64.43
		ECE	6.24	6.10	8.24	6.13	21.25	32.13	7.36	28.41	3.53
		MCE	1.54	2.05	2.31	1.55	9.51	31.87	2.51	14.18	1.28
		ACE	6.24	6.33	8.37	6.26	21.17	31.94	7.46	28.41	3.91
	Average	Acc.	69.18	68.45	69.24	69.18	68.56	57.69	69.18	8.59	69.05
		ECE	4.44	4.83	7.41	4.45	26.54	28.41	6.57	24.32	3.43
		MCE	1.37	1.62	2.22	1.37	11.23	20.09	2.63	16.29	1.18
		ACE	4.56	5.00	7.44	4.62	26.45	28.73	6.69	24.32	3.77

O Declaration of LLM usage

During the preparation of this manuscript, ChatGPT was used only for language polishing, grammar checking, and improving sentence clarity and readability. The tool was not used to generate experimental results, design the methodology, conduct analysis, create citations, or formulate the scientific claims of the paper. All technical content,

Backbone	Dataset	Metric	Base	MDCA	LS	MBLS	ECCV_ZS	ECCV_Penalty	TS	zeroshot	Ours (MCM)
CLIP-DR-ViT-B/32	APTOS	Acc.	78.28	77.98	78.53	78.26	75.34	66.72	78.59	7.86	77.52
		ECE	4.07	4.91	8.97	4.15	39.31	28.60	3.44	31.08	2.53
		MCE	1.52	1.74	1.78	1.51	13.62	25.01	0.78	17.28	0.81
		ACE	4.00	4.89	8.93	4.10	39.31	33.31	3.54	31.08	2.53
	EyePACS	Acc.	74.58	74.49	74.52	74.55	74.40	73.34	74.57	8.86	74.54
		ECE	3.29	3.05	6.24	3.06	12.89	38.52	7.22	23.22	2.39
		MCE	1.66	1.55	3.26	1.52	3.06	26.81	2.99	18.52	0.72
		ACE	3.24	2.97	6.22	3.04	13.08	38.52	7.20	24.91	2.41
	Messidor	Acc.	55.47	53.22	55.45	55.39	56.17	44.70	55.69	20.58	55.47
		ECE	3.11	3.82	5.01	3.07	24.93	14.65	6.42	21.65	3.58
		MCE	0.91	1.61	1.62	0.93	13.88	13.96	2.32	18.74	1.20
		ACE	5.15	4.56	6.21	5.04	24.93	14.70	7.11	21.65	3.96
	Messidor-2	Acc.	64.20	63.90	63.51	64.10	63.28	58.31	63.97	7.63	63.61
		ECE	5.09	6.11	8.04	5.06	20.00	29.68	6.05	27.48	2.57
		MCE	1.31	1.53	1.81	1.39	6.73	25.42	2.13	17.94	0.85
		ACE	4.99	6.09	8.03	4.90	20.00	29.68	6.05	30.38	2.66
	Average	Acc.	68.13	67.40	68.00	68.08	67.30	60.77	68.21	11.23	67.79
		ECE	3.89	4.47	7.07	3.84	24.28	27.86	5.78	25.86	2.77
		MCE	1.35	1.61	2.12	1.34	9.32	22.80	2.06	18.12	0.90
		ACE	4.35	4.63	7.35	4.27	24.33	29.05	5.98	27.01	2.89
MedCLIP-DR-B/32	APTOS	Acc.	62.27	62.14	63.16	63.96	62.27	49.43	62.27	11.25	61.73
		ECE	10.04	9.99	14.14	10.68	9.98	21.19	7.48	8.82	7.64
		MCE	3.27	3.16	4.46	3.22	3.16	16.74	4.29	8.82	2.80
		ACE	10.04	10.00	14.13	10.68	9.98	22.19	6.77	8.82	7.63
	EyePACS	Acc.	73.66	73.66	73.66	73.66	73.65	73.66	73.66	3.54	73.66
		ECE	2.84	2.09	7.65	2.89	3.40	20.87	7.85	16.56	3.17
		MCE	1.36	1.12	4.10	1.36	1.85	14.62	4.74	16.56	3.46
		ACE	3.54	3.47	7.64	3.66	3.67	20.87	7.84	16.00	5.18
	Messidor	Acc.	45.03	45.05	46.14	45.39	45.17	44.33	45.03	14.33	45.27
		ECE	9.42	9.19	6.16	5.04	3.25	19.09	12.55	10.72	1.93
		MCE	8.37	7.54	4.82	2.27	1.47	19.09	11.27	10.72	0.99
		ACE	9.41	10.01	7.10	5.28	3.37	19.09	13.72	10.72	3.71
	Messidor-2	Acc.	58.31	58.31	58.31	58.31	58.31	57.78	58.31	8.94	58.31
		ECE	4.45	6.96	10.34	9.75	5.90	22.96	12.41	11.13	7.03
		MCE	2.61	3.82	6.41	3.73	3.17	12.28	7.77	11.13	3.78
		ACE	5.62	7.60	11.31	9.81	6.98	23.00	12.41	11.13	9.03
	Average	Acc.	59.82	59.79	60.32	60.33	59.85	56.30	59.82	9.52	59.74
		ECE	6.69	7.06	9.57	7.09	5.63	21.03	10.07	11.81	4.94
		MCE	3.90	3.91	4.95	2.65	2.41	15.68	7.02	11.81	2.76
		ACE	7.15	7.77	10.04	7.36	6.00	21.29	10.19	11.67	6.39
MedCLIP-X-ray	COVIDX	Acc.	79.45	79.46	79.56	79.69	79.53	79.02	79.40	78.77	79.50
		ECE	29.34	29.34	29.45	29.58	29.41	28.91	29.30	28.67	24.39
		MCE	29.34	29.34	29.45	29.58	29.41	28.91	1.26	28.67	24.39
		ACE	29.34	29.34	29.45	29.58	29.41	28.91	4.98	28.67	29.39
	RSNA18	Acc.	52.08	52.11	52.43	52.60	52.40	52.21	52.02	47.60	52.15
		ECE	18.60	18.63	18.95	19.11	18.92	18.73	18.56	14.05	15.67
		MCE	18.60	18.63	18.95	19.11	18.92	18.73	0.49	14.05	15.67
		ACE	18.60	18.63	18.95	19.11	18.92	18.76	1.38	16.94	18.67
	Average	Acc.	65.77	65.78	66.00	66.15	65.96	65.62	65.71	63.19	65.83
		ECE	23.97	23.99	24.20	24.35	24.17	23.82	23.93	21.36	20.03
		MCE	23.97	23.99	24.20	24.35	24.17	23.82	0.88	21.36	20.03
		ACE	23.97	23.99	24.20	24.35	24.17	23.82	3.18	22.81	24.03

interpretations, figures, tables, and references were reviewed, verified, and finalized by the authors.

Backbone	Dataset	Metric	Base	MDCA	LS	MBLS	ECCV_ZS	ECCV_Penalty	TS	zeroshot	Ours (MCM)
BioMedCLIP-DR-B/32	APTOS	Acc.	77.25	77.02	76.88	76.92	62.27	49.43	77.25	54.59	76.71
		ECE	2.49	2.27	5.62	2.47	9.98	21.19	2.88	18.68	3.14
		MCE	0.80	0.67	1.33	0.77	3.16	16.74	0.59	9.01	1.23
		ACE	2.65	2.51	5.60	2.60	9.98	22.19	2.45	18.98	3.40
	EyePACS	Acc.	74.13	74.13	74.07	74.13	73.65	73.66	74.13	70.51	3.14
		ECE	1.48	1.67	72.21	1.50	3.40	20.87	7.86	16.19	1.96
		MCE	0.56	0.55	2.61	0.56	1.85	14.62	2.81	9.31	0.01
		ACE	1.83	1.85	3.62	1.83	3.67	20.87	7.73	16.16	2.01
	Messidor	Acc.	55.53	54.97	55.31	55.42	45.17	44.33	55.53	45.67	55.36
		ECE	5.13	4.25	5.98	4.62	3.25	19.09	9.04	8.90	2.92
		MCE	1.73	1.38	1.83	1.57	1.47	19.09	3.45	3.43	1.07
		ACE	5.36	5.12	6.73	5.38	3.37	19.09	8.84	9.22	2.97
	Messidor-2	Acc.	62.96	63.13	63.19	63.21	58.31	57.78	62.96	57.57	62.92
		ECE	6.43	6.19	8.88	6.14	5.90	22.96	6.59	7.37	3.26
		MCE	2.01	1.68	2.62	1.59	3.17	12.28	2.24	4.15	1.19
		ACE	6.46	6.59	9.09	6.29	6.98	23.00	6.77	6.94	3.47
	Average	Acc.	67.47	67.31	67.36	67.42	59.85	56.30	67.47	57.09	66.80
		ECE	3.88	3.60	6.03	3.68	5.63	21.03	6.59	12.79	2.82
		MCE	1.28	1.07	2.10	1.12	2.41	15.68	2.27	6.48	0.88
		ACE	4.07	4.02	6.26	4.02	6.00	21.29	6.45	12.83	2.96
BioMedCLIP-X-ray	COVIDX	Acc.	83.32	78.27	82.83	83.39	82.28	81.46	83.36	84.37	83.64
		ECE	6.97	8.84	11.67	8.65	6.72	8.95	6.99	10.70	3.50
		MCE	2.42	3.85	4.56	3.34	2.51	3.60	0.90	8.88	1.24
		ACE	6.97	8.93	11.69	8.64	6.72	9.03	7.09	10.07	3.66
	RSNA18	Acc.	63.80	63.97	63.36	63.58	63.61	61.67	63.81	49.71	63.76
		ECE	8.60	7.31	4.78	7.39	8.15	6.72	8.66	29.49	7.92
		MCE	2.98	2.56	1.44	2.38	2.39	2.33	3.30	16.67	2.44
		ACE	8.69	7.93	4.93	7.17	8.14	7.01	9.04	29.49	7.98
	Average	Acc.	73.56	71.12	73.09	73.48	72.94	71.57	73.59	67.04	73.70
		ECE	7.78	8.08	8.23	8.02	7.44	7.84	7.82	20.10	5.71
		MCE	2.70	3.21	3.00	2.86	2.45	2.97	2.10	12.78	1.84
		ACE	7.83	8.43	8.31	7.91	7.43	8.02	8.06	19.78	5.82
PLIP-DR-B/32	APTOS	Acc.	78.66	78.76	78.03	78.67	77.52	74.79	78.67	78.77	78.46
		ECE	4.98	5.26	12.19	5.14	3.60	21.00	1.93	11.12	0.96
		MCE	1.26	1.24	2.42	1.30	0.76	4.24	1.26	6.85	0.24
		ACE	4.98	5.24	12.19	5.05	3.52	21.00	4.98	10.71	0.99
	EyePACS	Acc.	74.43	74.44	74.28	74.48	74.33	74.02	74.43	57.41	74.45
		ECE	1.44	1.16	7.39	1.45	1.31	19.37	7.56	29.22	3.70
		MCE	0.49	0.39	3.38	0.52	0.33	7.69	0.49	21.04	1.21
		ACE	1.38	1.15	7.35	1.43	1.32	19.35	1.38	29.22	3.70
	Messidor	Acc.	55.64	55.83	55.47	55.58	56.03	54.36	55.64	45.58	55.25
		ECE	4.19	4.26	8.99	5.14	5.70	15.00	6.36	2.80	4.05
		MCE	1.52	1.74	2.39	1.64	1.67	7.74	1.52	2.62	1.05
		ACE	5.11	5.41	9.03	5.45	5.92	14.86	5.11	4.39	5.48
	Messidor-2	Acc.	63.46	63.42	63.74	63.64	63.28	62.27	63.46	47.25	63.70
		ECE	2.92	3.67	8.99	4.31	3.52	18.01	6.24	14.69	2.28
		MCE	1.01	1.35	2.14	1.22	0.93	6.07	1.01	8.61	0.97
		ACE	2.91	3.95	9.03	4.78	3.66	18.01	2.91	14.69	3.04
	Average	Acc.	68.05	68.12	67.88	68.09	67.79	66.36	68.05	57.25	67.97
		ECE	3.38	3.59	9.39	4.01	3.53	18.34	5.52	14.46	2.75
		MCE	1.07	1.18	2.58	1.17	0.92	6.43	1.07	9.78	0.87
		ACE	3.60	3.94	9.40	4.18	3.61	18.30	3.60	14.75	3.30

Backbone	Dataset	Metric	Base	MDCA	LS	MBLS	ECCV_ZS	ECCV_Penalty	TS	zeroshot	Ours (MCM)
PLIP-Histo.	DigestPath	Acc.	90.74	91.12	89.30	91.09	91.47	83.32	92.33	80.69	90.86
		ECE	4.92	3.62	4.07	4.62	4.43	14.67	3.87	6.23	6.80
		MCE	1.26	1.24	2.42	1.30	0.76	4.24	1.26	1.83	5.72
		ACE	4.98	5.24	12.19	5.05	3.52	21.00	4.98	6.23	6.79
	Kather	Acc.	87.00	86.51	86.71	85.90	87.33	87.69	87.21	57.74	87.24
		ECE	4.11	5.03	7.86	3.76	2.91	5.56	4.06	16.41	7.56
		MCE	0.49	0.39	3.38	0.52	0.33	7.69	0.49	5.34	5.31
		ACE	1.38	1.15	7.35	1.43	1.32	19.35	1.38	16.38	7.54
	PaNuke	Acc.	78.57	77.84	78.21	78.35	80.00	67.86	79.43	56.74	77.79
		ECE	11.98	9.75	2.95	10.18	10.91	4.83	11.21	18.82	15.39
		MCE	1.52	1.74	2.39	1.64	1.67	7.74	1.52	6.66	11.95
		ACE	5.11	5.41	9.03	5.45	5.92	14.86	5.11	19.64	15.39
	Average	Acc.	85.43	85.16	84.74	85.11	86.27	79.63	86.32	65.06	85.30
		ECE	7.00	6.13	5.69	6.19	6.08	8.35	6.38	13.82	9.92
		MCE	1.09	1.12	2.89	1.15	0.92	6.55	1.09	4.61	7.66
		ACE	3.83	3.93	8.98	3.98	3.59	18.40	3.83	14.08	9.91
QuiltNet-DR-B/32	APTOS	Acc.	76.97	76.12	77.00	76.72	77.52	73.04	76.97	10.03	75.81
		ECE	2.75	2.08	4.43	2.23	3.60	16.75	3.72	23.64	2.71
		MCE	1.26	1.24	2.42	1.30	0.76	4.24	1.26	1.26	0.84
		ACE	4.98	5.24	12.19	5.05	3.52	21.00	4.98	4.98	2.75
	EyePACS	Acc.	74.43	74.40	74.38	74.33	74.33	73.90	74.43	22.22	74.37
		ECE	3.15	3.41	4.73	3.46	1.31	19.92	7.52	11.37	1.21
		MCE	0.49	0.39	3.38	0.52	0.33	7.69	0.49	0.49	0.39
		ACE	1.38	1.15	7.35	1.43	1.32	19.35	1.38	1.38	1.25
	Messidor	Acc.	52.31	53.00	53.36	51.59	56.03	51.20	52.31	12.58	53.08
		ECE	4.01	5.76	4.66	4.37	5.70	13.31	6.21	26.70	2.56
		MCE	1.52	1.74	2.39	1.64	1.67	7.74	1.52	1.52	0.97
		ACE	5.11	5.41	9.03	5.45	5.92	14.86	5.11	5.11	3.89
	Messidor-2	Acc.	62.44	62.44	62.90	63.23	63.28	61.75	62.44	24.03	62.71
		ECE	3.89	3.88	5.13	3.79	3.52	15.55	8.73	7.85	3.10
		MCE	1.01	1.35	2.14	1.22	0.93	6.07	1.01	1.01	1.37
		ACE	2.91	3.95	9.03	4.78	3.66	18.01	2.91	2.91	3.57
	Average	Acc.	66.54	66.49	66.91	66.47	67.79	64.97	66.54	17.22	66.49
		ECE	3.45	3.78	4.74	3.46	3.53	16.38	6.54	17.39	2.40
		MCE	1.07	1.18	2.58	1.17	0.92	6.43	1.07	1.07	0.89
		ACE	3.60	3.94	9.40	4.18	3.61	18.30	3.60	3.60	2.86
QuiltNet-Histo.	DigestPath	Acc.	89.76	90.07	89.68	89.80	90.40	83.71	89.22	53.43	89.62
		ECE	5.27	4.09	2.42	5.35	4.42	13.58	5.31	19.85	7.81
		MCE	1.26	1.24	2.42	1.30	0.76	4.24	1.26	7.31	6.57
		ACE	4.98	5.24	12.19	5.05	3.52	21.00	4.98	20.23	7.81
	Kather	Acc.	91.56	90.75	90.61	90.49	93.23	91.40	91.84	60.32	91.63
		ECE	1.51	1.91	2.29	1.74	1.41	10.95	1.30	4.24	4.38
		MCE	0.49	0.39	3.38	0.52	0.33	7.69	0.49	1.20	2.87
		ACE	1.38	1.15	7.35	1.43	1.32	19.35	1.38	3.84	4.36
	PaNuke	Acc.	78.86	78.46	79.19	78.98	79.99	72.62	78.11	55.43	79.37
		ECE	11.86	10.44	7.98	11.04	10.43	8.95	11.74	24.15	13.45
		MCE	1.52	1.74	2.39	1.64	1.67	7.74	1.52	8.12	10.13
		ACE	5.11	5.41	9.03	5.45	5.92	14.86	5.11	24.15	13.42
	Average	Acc.	86.73	86.43	86.49	86.42	87.87	82.58	86.39	56.39	86.87
		ECE	6.21	5.48	4.23	6.04	5.42	11.16	6.12	16.08	8.55
		MCE	1.09	1.12	2.73	1.15	0.92	6.55	1.09	5.54	6.52
		ACE	3.83	3.93	9.52	3.98	3.59	18.40	3.83	16.07	8.53

Table A8: Accuracy and ECE of each calibration method in in-domain (ID) across 12 modality-specific configurations and 7 calibration methods covering three modalities: Diabetic retinopathy grading, histopathology, and chest-X-ray.

Backbone	Dataset	Metric	Base	MDCA	LS	MBLS	ECCV_ZS	ECCV_Penalty	TS	Ours (MCM)
CLIP-DR-ResNet-50-DS	APTOS	Acc.	43.75	48.01	50.54	43.80	20.71	43.83	44.10	43.68
		ECE	24.13	15.37	19.67	24.25	12.22	17.57	20.38	17.77
		MCE	6.07	4.58	8.21	6.05	7.43	17.26	6.88	7.23
		ACE	24.01	15.33	19.74	24.03	13.02	18.49	20.46	17.69
	EyePACS	Acc.	66.33	66.48	69.84	66.34	44.33	55.79	68.16	65.52
		ECE	10.94	13.97	10.83	10.95	25.52	31.43	10.07	7.49
		MCE	2.99	4.19	3.26	2.99	19.14	31.29	2.50	2.20
		ACE	10.71	13.68	10.89	10.70	25.37	31.89	10.01	7.50
	Messidor-2	Acc.	46.79	50.29	52.96	46.69	18.90	48.95	49.87	44.72
		ECE	10.15	10.18	8.41	9.98	18.57	23.22	9.52	7.89
		MCE	3.82	3.71	3.21	3.81	12.57	23.01	3.13	2.85
		ACE	10.17	9.32	8.03	10.03	18.48	23.22	9.89	8.62
	Average	Acc.	52.29	54.93	57.78	52.28	27.98	49.52	54.04	51.31
		ECE	15.07	13.17	12.97	15.06	18.77	24.07	13.32	10.26
		MCE	4.29	4.16	4.89	4.28	13.05	23.85	4.17	4.09
		ACE	14.96	12.78	10.89	14.92	18.96	24.53	13.45	10.61
CLIP-DR-ResNet-101-DS	APTOS	Acc.	35.38	49.90	36.88	35.45	51.13	18.08	34.20	41.95
		ECE	21.38	27.42	17.84	21.24	26.89	15.56	32.71	19.32
		MCE	6.25	7.12	4.96	6.32	23.43	11.09	10.77	4.86
		ACE	21.48	28.20	17.92	21.55	27.81	16.61	32.66	19.51
	EyePACS	Acc.	50.16	69.27	64.66	50.25	69.71	18.47	35.70	61.60
		ECE	16.53	23.59	15.13	16.60	43.77	11.73	26.14	13.52
		MCE	7.57	8.83	5.26	7.59	37.90	11.25	13.28	4.84
		ACE	15.98	22.81	15.35	16.07	43.77	18.68	26.02	13.17
	Messidor-2	Acc.	40.20	57.72	46.46	40.27	58.05	9.12	35.11	51.64
		ECE	13.07	17.10	12.52	13.32	33.72	16.72	22.07	12.54
		MCE	5.98	7.39	3.81	5.90	32.64	14.89	11.87	4.96
		ACE	12.60	16.04	12.98	12.36	33.72	22.46	21.61	12.22
	Average	Acc.	41.91	58.96	49.33	41.99	59.63	15.22	35.00	51.73
		ECE	16.99	22.70	15.16	17.05	34.79	14.67	26.97	15.12
		MCE	6.60	7.78	4.67	6.60	31.32	12.41	11.97	4.89
		ACE	16.69	22.35	15.41	16.66	35.10	19.25	26.76	14.97
CLIP-DR-ViT-B/16-DS	APTOS	Acc.	28.69	34.50	26.82	28.82	31.30	47.17	30.16	29.74
		ECE	41.70	22.61	30.89	41.80	8.17	23.21	29.15	30.48
		MCE	12.65	7.69	8.20	12.65	7.21	22.98	7.32	6.28
		ACE	41.56	22.96	31.06	41.64	11.50	23.36	29.98	30.48
	EyePACS	Acc.	40.95	45.32	45.77	41.02	53.78	68.11	26.71	41.91
		ECE	32.71	15.60	14.94	32.77	25.69	43.89	22.74	12.89
		MCE	12.93	6.28	7.22	12.94	17.74	43.84	11.20	6.79
		ACE	32.49	15.25	15.11	32.55	27.30	43.89	23.11	12.72
	Messidor-2	Acc.	29.07	34.81	29.80	29.07	30.47	57.72	25.21	28.42
		ECE	26.65	15.61	19.79	26.59	13.53	32.77	19.18	16.87
		MCE	11.87	8.64	12.51	11.81	12.08	32.30	9.99	9.22
		ACE	27.23	15.91	18.80	27.10	16.05	32.77	19.71	16.96
	Average	Acc.	32.90	38.21	34.13	32.97	38.52	57.67	27.36	33.36
		ECE	33.69	17.94	21.87	33.72	15.80	33.29	23.69	20.08
		MCE	12.48	7.54	9.31	12.47	12.34	33.04	9.50	7.43
		ACE	33.76	18.04	21.66	33.76	18.28	33.34	24.27	20.05

Backbone	Dataset	Metric	Base	MDCA	LS	MBLS	ECCV_ZS	ECCV_Penalty	TS	Ours (MCM)
CLIP-DR-ViT-B/32-DS	APTOS	Acc.	40.88	44.18	28.73	40.88	26.70	25.76	36.44	41.29
		ECE	23.27	22.18	16.39	23.27	17.19	17.13	20.60	23.10
		MCE	5.59	5.42	7.72	5.59	7.73	8.38	5.03	6.72
		ACE	23.12	22.64	16.42	23.12	16.73	21.06	20.60	23.01
	EyePACS	Acc.	51.50	58.45	49.68	51.50	37.91	22.99	52.39	59.97
		ECE	11.67	15.79	8.15	11.67	11.95	20.01	9.98	8.37
		MCE	4.40	7.86	3.82	4.40	5.29	10.52	2.99	2.52
		ACE	11.38	15.56	8.17	11.38	11.75	20.90	10.10	8.66
	Messidor-2	Acc.	38.36	38.28	26.76	38.36	22.48	27.93	35.51	38.50
		ECE	12.55	7.75	14.19	12.55	14.46	21.29	15.24	5.84
		MCE	3.17	2.25	5.62	3.17	6.20	7.08	6.23	0.59
		ACE	12.65	8.82	15.37	12.65	14.51	21.37	15.61	6.42
	Average	Acc.	43.58	46.97	35.06	43.58	29.03	25.56	41.45	46.59
		ECE	15.83	15.24	12.91	15.83	14.53	19.48	15.27	12.44
		MCE	4.39	5.18	5.72	4.39	6.41	8.66	4.75	3.28
		ACE	15.72	15.67	13.32	15.72	14.33	21.11	15.44	12.70
MedCLIP-DR-B/32-DS	APTOS	Acc.	47.79	40.03	48.56	49.10	48.91	34.97	49.05	48.75
		ECE	18.92	8.21	15.99	14.13	14.17	16.08	18.07	12.14
		MCE	16.60	6.35	10.29	12.22	10.50	16.08	9.34	8.42
		ACE	18.90	11.89	16.12	14.68	14.37	17.72	18.12	12.57
	EyePACS	Acc.	58.90	53.76	72.61	68.24	73.14	30.90	73.29	73.04
		ECE	30.47	27.10	39.25	9.73	40.88	18.75	41.64	37.27
		MCE	19.74	17.21	25.79	24.11	26.87	18.75	31.25	25.20
		ACE	30.47	28.81	39.25	38.30	40.88	18.98	41.72	37.27
	Messidor-2	Acc.	51.38	46.63	57.51	57.97	57.67	30.66	57.97	57.70
		ECE	22.44	15.69	24.59	23.73	23.64	14.57	26.46	21.20
		MCE	16.11	10.26	14.34	15.18	14.61	14.57	17.30	14.43
		ACE	22.44	16.61	24.55	23.73	23.49	15.37	26.92	21.20
	Average	Acc.	52.69	46.81	59.56	58.44	59.91	32.18	60.10	59.83
		ECE	23.95	17.00	26.61	15.86	26.23	16.47	28.72	23.54
		MCE	17.48	11.27	16.80	17.17	17.33	16.47	19.30	16.02
		ACE	23.94	19.10	26.64	25.57	26.25	17.36	28.92	23.68
MedCLIP-X-ray-DS	COVIDX	Acc.	79.15	79.15	79.08	79.06	79.03	78.50	79.16	79.15
		ECE	29.03	29.03	28.97	28.95	28.91	28.39	29.07	28.04
		MCE	29.03	29.03	28.97	28.95	28.91	28.39	29.07	28.04
		ACE	29.03	29.03	28.97	28.95	28.91	28.39	29.07	29.04
	Average	Acc.	79.15	79.15	79.08	79.06	79.03	78.50	79.16	79.15
		ECE	29.03	29.03	28.97	28.95	28.91	28.39	29.07	28.04
		MCE	29.03	29.03	28.97	28.95	28.91	28.39	29.07	28.04
		ACE	29.03	29.03	28.97	28.95	28.91	28.39	29.07	29.04

Backbone	Dataset	Metric	Base	MDCA	LS	MBLS	ECCV_ZS	ECCV_Penalty	TS	Ours (MCM)
BioMedCLIP-DR-B/32-DS	APTOS	Acc.	50.20	51.77	50.74	50.58	50.65	52.80	51.02	50.12
		ECE	23.70	26.15	20.65	21.85	30.02	15.19	25.28	22.19
		MCE	7.02	6.59	7.90	6.91	10.80	6.45	6.78	6.56
		ACE	24.07	26.86	20.75	21.92	30.26	15.76	26.01	22.35
	EyePACS	Acc.	68.82	69.50	68.33	68.24	69.08	70.98	69.19	68.58
		ECE	8.96	9.34	12.53	9.73	15.68	13.41	9.08	6.55
		MCE	2.26	2.37	3.13	2.38	4.25	3.76	4.17	1.67
		ACE	8.96	9.33	12.53	9.72	15.43	13.41	7.41	6.54
	Messidor-2	Acc.	59.35	59.46	58.64	58.73	57.61	59.50	59.56	59.07
		ECE	12.52	11.95	14.44	13.02	17.38	11.51	12.74	10.08
		MCE	3.79	4.11	4.83	3.92	7.12	4.82	2.54	3.07
		ACE	12.49	11.64	14.50	12.95	16.59	11.51	11.98	10.09
	Average	Acc.	59.45	60.24	59.24	59.19	59.11	61.10	59.92	59.26
		ECE	15.06	15.81	15.88	14.86	21.03	13.37	15.70	12.94
		MCE	4.36	4.35	5.29	4.40	7.39	5.01	4.50	3.77
		ACE	15.17	15.94	15.93	14.86	20.76	13.56	15.13	12.99
BioMedCLIP-X-ray-DS	COVIDX	Acc.	74.56	73.40	75.55	77.45	73.63	77.15	74.10	78.84
		ECE	9.00	9.79	6.82	7.31	10.38	3.83	8.70	8.02
		MCE	4.41	4.36	3.61	4.71	4.91	1.61	4.20	5.74
		ACE	8.96	9.91	6.83	7.47	10.27	3.99	8.87	8.08
	Average	Acc.	74.56	73.40	75.55	77.45	73.63	77.15	74.10	78.84
		ECE	9.00	9.79	6.82	7.31	10.38	3.83	8.70	8.02
		MCE	4.41	4.36	3.61	4.71	4.91	1.61	4.20	5.74
		ACE	8.96	9.91	6.83	7.47	10.27	3.99	8.87	8.08
PLIP-B/32-DS	APTOS	Acc.	39.01	41.09	41.26	28.77	39.45	37.15	39.01	35.19
		ECE	17.75	17.02	13.12	23.04	14.11	9.86	16.16	21.78
		MCE	1.26	1.24	2.42	1.30	0.76	4.24	1.26	4.92
		ACE	17.87	17.18	12.19	22.91	14.21	9.95	15.64	21.96
	EyePACS	Acc.	70.34	70.73	66.95	69.96	69.88	71.12	69.59	61.84
		ECE	10.96	9.69	16.86	13.50	8.88	25.38	13.09	10.69
		MCE	0.49	0.39	3.38	0.52	0.33	7.69	0.49	3.99
		ACE	10.38	8.15	17.35	12.43	8.32	24.35	12.38	10.50
	Messidor-2	Acc.	56.84	53.48	52.31	53.73	53.48	46.89	54.95	44.45
		ECE	9.65	9.97	12.94	13.07	9.66	21.19	11.10	8.20
		MCE	1.01	1.35	2.14	1.22	0.93	6.07	1.01	3.57
		ACE	9.91	9.95	11.03	12.78	8.56	20.01	10.91	7.93
	Average	Acc.	55.40	55.10	53.51	50.82	54.27	51.72	54.51	47.16
		ECE	12.79	12.23	14.30	16.54	10.88	18.81	13.45	13.56
		MCE	0.92	0.99	2.65	1.01	0.67	6.00	0.92	4.16
		ACE	12.72	11.76	13.52	16.04	10.36	18.10	12.97	13.47

Backbone	Dataset	Metric	Base	MDCA	LS	MBLS	ECCV_ZS	ECCV_Penalty	TS	Ours (MCM)
PLIP-Histo.-DS	DigestPath	Acc.	65.15	46.46	61.52	50.25	74.22	39.36	64.95	56.56
		ECE	24.36	39.39	15.87	35.59	16.72	33.58	24.06	33.77
		MCE	1.26	1.24	2.42	1.30	0.76	4.24	1.26	2.58
		ACE	24.98	35.24	12.19	35.05	15.52	33.00	24.98	33.77
	Kather	Acc.	40.51	26.24	27.38	26.89	56.49	21.15	51.50	51.20
		ECE	12.79	13.11	5.09	12.53	9.77	9.37	7.23	11.10
		MCE	0.49	0.39	3.38	0.52	0.33	7.69	0.49	1.95
		ACE	12.38	12.15	7.35	12.43	9.32	10.35	8.38	11.07
	Average	Acc.	52.83	36.35	44.45	38.57	65.35	30.26	58.23	53.88
		ECE	18.58	26.25	10.48	24.06	13.24	21.48	15.65	22.43
		MCE	0.88	0.81	2.90	0.91	0.54	5.96	0.88	2.26
		ACE	18.68	23.69	9.77	23.74	12.42	21.67	16.68	22.42
QuiltNet-DR-B/32-DS	APTOS	Acc.	32.36	45.50	42.62	39.80	47.18	48.22	42.51	27.86
		ECE	19.28	10.25	9.73	12.67	10.23	11.89	8.67	22.03
		MCE	1.26	1.24	2.42	1.30	0.76	4.24	1.26	2.42
		ACE	14.98	8.24	12.19	11.05	9.52	15.00	6.98	22.13
	EyePACS	Acc.	55.53	56.39	61.05	66.04	62.40	58.75	61.90	34.33
		ECE	8.80	11.51	9.27	11.78	8.74	18.87	10.81	23.30
		MCE	0.49	0.39	3.38	0.52	0.33	7.69	0.49	4.00
		ACE	7.38	10.15	9.35	9.43	8.32	19.35	11.38	23.28
	Messidor-2	Acc.	32.36	40.25	45.95	41.93	41.94	42.79	38.65	21.56
		ECE	18.19	7.85	11.12	8.99	7.09	19.38	10.15	25.82
		MCE	1.01	1.35	2.14	1.22	0.93	6.07	1.01	2.01
		ACE	17.91	6.95	10.03	8.78	6.66	18.01	9.91	25.77
	Average	Acc.	40.08	47.38	49.87	49.26	50.50	49.92	47.69	27.92
		ECE	15.42	9.87	10.04	11.15	8.69	16.71	9.87	23.72
		MCE	0.92	0.99	2.65	1.01	0.67	6.00	0.92	2.81
		ACE	13.42	8.44	10.52	9.75	8.16	17.45	9.42	23.73
QuiltNet-Histo.-DS	DigestPath	Acc.	47.40	39.66	42.96	36.97	43.63	31.02	31.43	49.01
		ECE	41.30	46.38	42.33	51.88	44.48	46.04	60.39	42.93
		MCE	1.26	1.24	2.42	1.30	0.76	4.24	1.26	1.11
		ACE	41.98	45.24	42.19	50.05	43.52	43.00	58.98	42.93
	Kather	Acc.	45.53	50.59	37.08	47.95	46.23	28.13	47.77	44.51
		ECE	15.02	9.54	15.59	10.38	13.41	7.46	12.94	24.27
		MCE	0.49	0.39	3.38	0.52	0.33	7.69	0.49	5.43
		ACE	15.38	10.15	16.35	10.43	11.32	10.35	10.38	24.19
	Average	Acc.	46.46	45.13	40.02	42.46	44.93	29.57	39.60	46.76
		ECE	28.15	27.96	28.96	31.13	28.94	26.75	36.66	33.60
		MCE	0.88	0.81	2.90	0.91	0.54	5.96	0.88	3.27
		ACE	28.68	27.69	29.27	30.24	27.42	26.67	34.68	33.56

Table A9: Accuracy and ECE of each calibration method in domain-shift (out-domain) across 12 modality-specific configurations and 7 calibration methods covering three modalities: Diabetic retinopathy grading, histopathology, and chest-X-ray.