Choosing Training-Time Calibration Objectives for Frozen Foundation-Model Features: A Linear-Probing Benchmark

Anonymous Authors[1]

Abstract

Calibration objectives for deep classifiers have historically been designed under end-to-end training. Foundation models, however, are increasingly used through frozen-feature adaptation, and full fine-tuning to recalibrate is often infeasible. Post-hoc temperature scaling is cheap but limited to a scalar transform. We ask whether *calibration-aware linear probing*—relearning only the head under a calibration objective—can occupy the middle ground. Across 15 dataset–model settings spanning CLIP, DINOv2, same-domain CNNs, and cross-domain CNN transfer, the answer is a clean representation-family split rather than a universal winning loss. CLIP gains, when present, come from a direct confidence–accuracy penalty. DINOv2 leaves little reliable headroom beyond temperature scaling. Same-domain CNNs favor confidence- and margin-sensitive reweighting, including a new diagnostic V-family introduced here. Calibration-aware probing therefore serves both as a lightweight recalibration tool and as a diagnostic that exposes how frozen representations encode confidence. Objective choice is part of evaluating uncertainty on frozen foundation-model features, not a minor implementation detail.

1. Introduction

Training-time calibration methods have been studied largely in end-to-end supervised regimes, typically from scratch on datasets such as CIFAR or ImageNet (Mukhoti et al., 2020; Kumar et al., 2018; Müller et al., 2019; Thulasidasan et al., 2019; Tao et al., 2023). In contrast, large pretrained vision backbones are increasingly used through frozen-feature adaptation: the representation is fixed, and only a lightweight downstream module is trained. This workflow is attractive because full fine-tuning is expensive, can be unnecessary, and may be impossible in closed-weight settings. Linear probing is also a strong practical baseline for adapting CLIP-like models (Huang et al., 2024). Post-hoc calibration methods such as temperature scaling remain applicable, since they require only validation logits. What is less clear is whether richer *training-time* calibration objectives still help when the only trainable component is the head.

This question is practically important because frozen-transfer performance is often summarized by top-1 accuracy, while many deployment decisions depend directly on confidence. Abstention, thresholding, ranking, routing, and human override all require confidence estimates whose scale is meaningful. In these settings, calibration is not an auxiliary diagnostic; it is part of the model's operational performance.

Existing methods occupy two extremes. *Post-hoc* methods are cheap and robust, but they apply only a low-dimensional transformation to an already-trained classifier (Guo et al., 2017; Kull et al., 2019). *Training-time* objectives are more expressive, but they ordinarily assume that the full network can be optimized end to end (Mukhoti et al., 2020; Kumar et al., 2018; Müller et al., 2019; Thulasidasan et al., 2019; Tao et al., 2023). Frozen-feature linear probing lies between these extremes: the backbone remains fixed, but the classifier can still be learned under a calibration-aware loss. The regime is simple, inexpensive once features are cached, and widely used, but it has not been systematically evaluated as a calibration method in its own right.

Recent work makes this gap sharper. Decoupling feature extraction from classifier training can improve calibration at low cost (Jordahn & Olmos, 2024), and calibratability analyses show that top layers can become a bottleneck through over-compression and over-training (Wang & Zhang, 2024). These results motivate head-level calibration, but they do not settle the frozen-foundation-model case. In that case, the backbone was not trained end to end on the downstream task, and the practitioner may only be able to train a supervised linear head.

We study this case on frozen CLIP (Radford et al., 2021) and

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

DINOv2 (Oquab et al., 2024) features, with same-domain and cross-domain CNNs as controlled non-foundation baselines. The central result is a *representation-family split*. CLIP benefits most, when it benefits, from a direct confidence–accuracy penalty. DINOv2 leaves little statistically reliable headroom beyond temperature scaling under our lightweight protocol. Same-domain CNN controls, especially CIFAR-100 with WRN-28-10, instead favor margin-sensitive reweighting. A single-objective summary would collapse these cases into one aggregate number and miss the structure.

The paper therefore supports two claims. **First**, calibration-aware linear probing is a practical middle ground between post-hoc scaling and end-to-end recalibration. **Second**, there is no universally preferred calibration objective across frozen representations: objective choice is itself part of the uncertainty evaluation. The goal is not to replace temperature scaling with one new loss, but to show that a small family of head-level calibration objectives reveals meaningful differences across feature families.

We make four contributions.

1. We conduct a matched evaluation of calibration-aware linear probing on frozen features, spanning CLIP, DINOv2, and controlled CNN baselines over 15 dataset–model settings.

2. We introduce V1–V4, a compact diagnostic family of confidence- and margin-sensitive probe objectives. The family separates confidence-only behavior from margin-aware behavior under a fixed representation, rather than serving as a single universal replacement loss.

3. We use a conservative evaluation protocol: macro-selection across seeds, paired tests where per-seed outputs are available, explicit accuracy guards, and measured compute cost.

4. We show that the preferred objective depends strongly on the representation family: CLIP, DINOv2, and same-domain CNNs occupy qualitatively different calibration regimes.

2. Related Work

2.1. Post-hoc and Training-Time Calibration

Modern neural classifiers are often miscalibrated under standard cross-entropy training, with confidence typically exceeding accuracy as models become more over-parameterized (Guo et al., 2017). Post-hoc methods address this by transforming the outputs of a trained classifier. Temperature scaling is the standard low-cost baseline (Guo et al., 2017); Dirichlet and matrix-scaling extensions increase flexibility while remaining post hoc (Kull et al., 2019). These methods are natural baselines in frozen-feature settings because they require only logits and a validation set, but their expressive power is limited by construction.

Training-time methods modify the objective or training procedure itself, including Focal Loss (Mukhoti et al., 2020), MMCE-style penalties (Kumar et al., 2018), Label Smoothing (Müller et al., 2019), Mixup (Thulasidasan et al., 2019), and Dual Focal Loss (Tao et al., 2023). They are more expressive than post-hoc scaling, but they have mostly been studied when all network parameters are trainable. Our problem sits between these two families: the representation is fixed, but the classifier is still learned under a calibration-aware objective.

2.2. Head-Level Calibration and Calibratability

A growing literature argues that calibration is not solely a property of the learned representation. Jordahn & Olmos (2024) show that decoupling feature extraction from classifier training can improve calibration while preserving accuracy. Wang & Zhang (2024) analyze *calibratability* across depth and identify a weak-classifier hypothesis: a head that is not over-trained can preserve more calibratable behavior. Classifier-centered methods such as BalCAL (Ni et al., 2025) continue this line by treating the classifier as a primary site of confidence distortion.

Our work shares this head-level perspective, but differs in scope and objective. We do not study a two-stage same-task training pipeline, and we do not propose a single replacement classifier. Instead, we evaluate multiple calibration-aware losses under the minimal frozen-feature regime that is common for foundation-model adaptation. This choice is less expressive than full fine-tuning, but it makes the role of objective choice and representation family easier to isolate.

2.3. Linear Probing and Foundation-Model Calibration

Linear probing has long been used to evaluate learned representations, and recent work shows that it remains a strong practical baseline for adapting large pretrained vision models (Huang et al., 2024). On the uncertainty side, calibration of vision-language models has been studied in zero-shot inference (LeVine et al., 2023; Tu et al., 2024), prompt tuning (Wang et al., 2025), and test-time prompt tuning (Yoon et al., 2024; Sharifdeen et al., 2025). These works show that CLIP-like models can be miscalibrated, that temperature scaling is a strong baseline, and that lightweight adaptation can materially alter confidence.

Our setting is different from zero-shot calibration, prompt tuning, adapter calibration, and test-time adaptation. We study *supervised frozen-feature linear probing*: the vi-

sual backbone is fixed, adaptation occurs only through a lightweight classifier, and calibration is evaluated across multiple objective families under one matched protocol. This isolates the role of the representation and the head more cleanly than methods that also modify prompts, textual features, or test-time dynamics.

3. Calibration-Aware Linear Probing on Frozen Features

Let $f_\theta(x) \in \mathbb{R}^d$ be a frozen feature extractor and $g_{W,b}(x) = Wf_\theta(x) + b$ a linear classifier over C classes. We optimize only the probe parameters:

$$\min_{W,b} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(Wf_\theta(x_i) + b,\, y_i)\,, \tag{1}$$

where $\mathcal{L}$ is a calibration-aware objective. Because the backbone is fixed, differences across objectives can be attributed to the head and loss rather than to representation drift. Once frozen features are cached, probe training is lightweight relative to full-network recalibration.

We study two probe variants. **LP-1L** uses final-layer features only. **LP-2L** concatenates an intermediate representation with the final representation, increasing the feature dimension while keeping the probe linear; it is not a two-layer MLP head. For same-domain CNN settings, pretrained classifier weights are reused when dimensions match, so the CE probe corresponds closely to the original classifier and calibration-aware probes measure incremental gains from objective choice alone. For transfer and foundation-model settings, the probe is initialized from class prototypes computed from frozen training features, yielding a stable starting point across backbones with very different feature scales. The initialization protocol is shared across feature families so that observed differences are not artifacts of initialization.

3.1. Baseline Objectives

Our evaluation includes a standard cross-entropy (CE) probe, CE followed by temperature scaling fitted on validation logits (CE+TS), and several calibration-aware training-time baselines: Focal Loss, Label Smoothing, Mixup, a modified dual focal-loss variant (MDFL), and a samplewise confidence–correctness penalty that we denote BCalErr. We use the name *BCalErr* deliberately to avoid conflating this implementation with kernel MMCE. For a mini-batch B, the objective penalizes the discrepancy between predicted confidence and the current correctness indicator:

$$\mathcal{L}_{\text{BCalErr}} = \text{CE}(z, y) + \lambda \cdot \frac{1}{|B|} \sum_{i \in B} \left| \max_c p_c^{(i)} - \mathbf{1}[\hat{y}_i = y_i] \right|, \tag{2}$$

where $p = \text{softmax}(z)$ and $\hat{y} = \arg\max_c p_c$. BCalErr penalizes overconfident errors and underconfident correct predictions, acting as a simple mini-batch surrogate for the confidence–accuracy gap. In implementation, gradients flow through the confidence term while the correctness indicator is treated as a fixed batch target. Because it operates on scalar confidence rather than on between-class margins, it provides a natural contrast to the confidence- and margin-sensitive family introduced next.

3.2. A New Confidence- and Margin-Sensitive Objective Family

To test whether confidence and decision-boundary structure matter for calibration under frozen transfer, we introduce a family of confidence- and margin-sensitive objectives, denoted V1–V4. Let

$$p = \text{softmax}(z/\tau), \qquad m = p_y - \max_{c \neq y} p_c, \tag{3}$$

where p_y is the probability of the true class, m is the margin to the strongest competitor, and τ is used only for weight computation (the logits in the CE term itself are unscaled).

We evaluate four reweighting rules:

$$\mathcal{L}_{\text{V1}} = [\text{ReLU}(1 - m)]^\gamma \cdot \text{CE}(z, y), \tag{4}$$

$$\mathcal{L}_{\text{V2}} = (1 - p_y)^\gamma \cdot \text{CE}(z, y), \tag{5}$$

$$\mathcal{L}_{\text{V3}} = \sigma(-s\,(p_y - c))^\gamma \cdot \text{CE}(z, y), \tag{6}$$

$$\mathcal{L}_{\text{V4}} = [\sigma(-s\,(p_y - c)) \cdot \sigma(-s\,m)]^{\gamma/2} \cdot \text{CE}(z, y). \tag{7}$$

The four variants form a small ablation grid over what the weight depends on (confidence vs. margin) and how it depends on it (sharp polynomial vs. smooth sigmoid). V1 is margin-only with a hinge-like polynomial weight. V2 is confidence-only with a polynomial weight on the true-class deficit $1 - p_y$. V3 is also confidence-only but uses a smooth sigmoid weight, providing a sharp/smooth ablation of V2. V4 combines confidence and margin signals through a product of two sigmoids. The shared diagnostic question is whether calibration improves by shaping confidence alone (V2, V3), by attending to runner-up competition alone (V1), or by combining both (V4). Our implementation of V2 used an equivalent but more verbose form (Section B); the reduced form shown above describes its actual behavior.

The V-family is best understood as a *diagnostic* family rather than as a single universal replacement loss. Its role in this paper is to test whether confidence- or margin-sensitive reweighting is the right inductive bias for a frozen representation. In the main experiments, V3 and V4 use the fixed defaults $s = 5$ and $c = 0.5$; adaptive or representation-specific choices of these shape parameters are left to future work.

4. Evaluation Protocol

Our evaluation contains 15 dataset–model settings spanning three regimes:

- **Same-domain CNN** (4 settings): CIFAR-10 and CIFAR-100 with ResNet-110 and WRN-28-10 features.
- **Cross-domain CNN transfer** (3 settings): ImageNet-pretrained ResNet-50 features on STL-10, Skin Cancer, and TinyImageNet.
- **Foundation models** (8 settings): frozen CLIP ViT-B/16 and DINOv2 ViT-B/14 features on CIFAR-10, CIFAR-100, STL-10, and TinyImageNet.

At the top level, nine loss functions, two probe variants, and five random seeds yield $15 \times 9 \times 2 \times 5 = 1{,}350$ method–variant–seed cells; the hyperparameter sweeps in Section A expand these cells into a larger number of training jobs.

Training protocol. All probes use the same recipe: Adam, weight decay 10^{-4}, 25 epochs, batch size 2048, early-stopping patience 5, and ECE computed with 15 equal-width confidence bins. Seeds are $\{42, 123, 456, 789, 1024\}$. All methods share the learning-rate grid $\{10^{-4}, 3\times10^{-4}, 5\times10^{-4}, 10^{-3}, 3\times10^{-3}, 5\times10^{-3}\}$. Method-specific grids are reported in Section A. Because these grids are not perfectly matched in size across methods, we treat search-budget inequality as a limitation rather than normalize it away.

Metrics and claim scope. ECE is our primary selection and reporting metric because the paper focuses on confidence calibration under fixed features; confidence is the maximum softmax probability and ECE uses 15 equal-width bins. We also track accuracy and filter degenerate runs below twice random chance, but we do not claim that ECE alone captures every uncertainty property. The same-domain CNN comparisons against temperature scaling are aggregate because per-seed TS outputs are unavailable; foundation-model comparisons against TS are paired where reported.

Configuration selection. Hyperparameters are chosen by *macro-averaged validation ECE across seeds*: for each method and setting, we group runs by configuration, average validation ECE over the five seeds, select the single best configuration, and report the mean and standard deviation of its test ECE and accuracy across those same seeds. This prevents per-seed cherry-picking; the accuracy guard described above filters degenerate configurations before selection.

Statistical testing. The "best proposed method" in each setting is defined as the lowest-ECE choice among $\{\text{V1}, \text{V2}, \text{V3}, \text{V4}\}$ across LP-1L and LP-2L under the macro-selection protocol. We compare this method against BCalErr on all 15 settings and against temperature scaling on the eight foundation-model settings for which per-seed TS outputs are available. We report paired two-sided t-tests, Wilcoxon signed-rank p-values when applicable, and wins/ties/losses across seeds. Reported p-values are exploratory, post-selection summaries and are not adjusted for multiplicity.

Compute analysis. A compute study measures wall-clock time and peak memory with cached features to quantify the practical cost of calibration-aware probing.

5. Results

5.1. Foundation Models: Same Protocol, Different Regimes

When the best learned probe is summarized per setting, the broader calibration-aware probe family achieves lower mean ECE than paired temperature scaling on six of the eight foundation-model settings (Figure 1, Table 1). The two exceptions are CIFAR-10/CLIP and TinyImageNet/CLIP, where temperature scaling remains lower. Thus, frozen-feature recalibration through a training-time objective is often more expressive than a scalar post-hoc temperature, but it is not uniformly better. The identity of the strongest objective is itself representation-dependent.

Within probe objectives, CLIP consistently favors BCalErr. Across the four CLIP settings, the strongest calibration-aware probe objective is BCalErr. The strongest example is STL-10 with CLIP features: the CE probe starts at 19.60% ECE, temperature scaling reduces this to 6.29%, and LP-2L with BCalErr reaches 0.69% (Table 1). CIFAR-100/CLIP shows the same qualitative pattern. At the same time, learned probing does *not* uniformly dominate temperature scaling on CLIP: on CIFAR-10/CLIP and TinyImageNet/CLIP, TS is as good as or better than any learned objective in this summary. The defensible conclusion is therefore specific: when a learned objective helps on CLIP, the gain is explained by direct confidence–accuracy correction, not by margin-sensitive reweighting.

DINOv2 leaves little room beyond temperature scaling. DINOv2 behaves differently. Temperature scaling is already strong, paired differences are small, and the best V-method is usually close to the best overall probe. Mean improvements exist, especially on TinyImageNet, but they are modest and not the dominant story. Under this protocol, DINOv2 therefore looks closer to a calibration-saturated regime than to one that needs aggressive head-level recalibration.

Taken together, these results support the paper's main empirical claim. Calibration-aware probing is useful on frozen

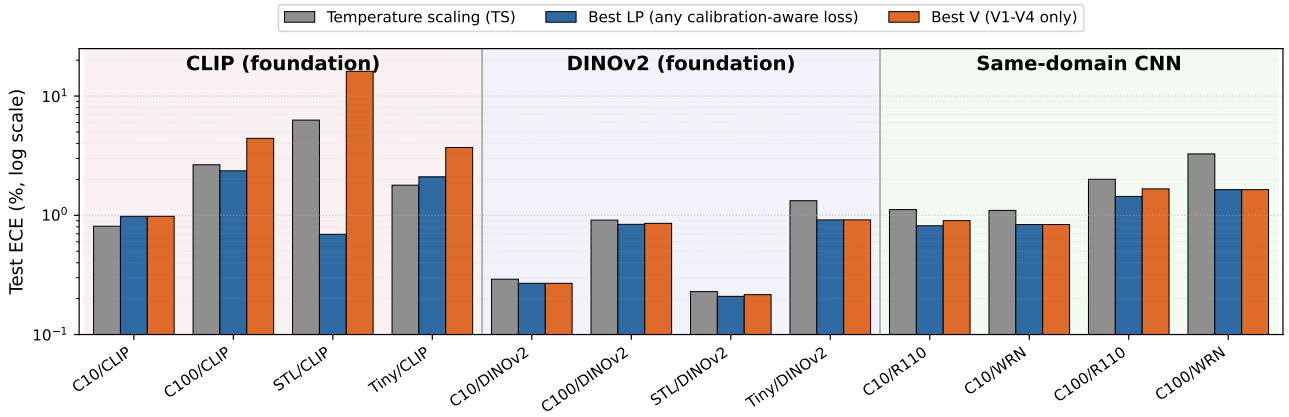

Figure 1. Test ECE (%, log scale) across 12 primary frozen-feature settings, grouped by representation family. TS denotes temperature scaling applied to the CE probe; Best LP is a post-hoc summary of the lowest-ECE learned probe among the evaluated calibration-aware objectives; Best V restricts this summary to V1–V4. The plot illustrates the representation-family split: CLIP gains, when present, come from BCalErr rather than the V-family; DINOv2 remains close to TS; and the same-domain CNN controls, especially WRN-28-10, are more favorable to the V-family, including its margin-sensitive variants.

Table 1. Foundation-model summary (test ECE %, lower is better). Best LP summarizes the lowest-ECE learned probe among calibration-aware objectives and V1–V4; Best V restricts the summary to V1–V4. These columns are descriptive benchmark summaries, not deployable selectors.

Setting	TS	Best LP	Best V
C10 / CLIP	**0.810**	0.981	0.981
C10 / DINOv2	0.291	**0.269**	**0.269**
C100 / CLIP	2.653	**2.360**	4.422
C100 / DINOv2	0.911	**0.840**	0.856
STL-10 / CLIP	6.286	**0.693**	16.125
STL-10 / DINOv2	0.229	**0.209**	0.216
Tiny / CLIP	**1.789**	2.103	3.705
Tiny / DINOv2	1.325	**0.915**	**0.915**

foundation-model features, but the preferred objective depends strongly on the representation family. CLIP and DINOv2 do not share the same calibration regime, and a single default objective does not describe both.

5.2. Paired Tests Clarify the Role of the V-Family

Paired tests sharpen the interpretation (Table 2). Against BCalErr across all 15 settings, the best proposed V-method records **3 significant wins, 3 significant losses, and 9 non-significant**. The V-family and BCalErr therefore serve different parts of the evaluation; neither dominates the other.

Against temperature scaling on the eight foundation-model settings, the best proposed V-method records **0 significant wins, 4 significant losses, and 4 ties**. The four losses are exactly the CLIP settings; the four ties are exactly the DINOv2 settings. This result prevents overclaiming: the broader probe family can improve over TS on foundation

Table 2. Paired significance summary. Comparisons use paired two-sided t-tests across seeds. Values are post-selection summaries; p-values are exploratory and not adjusted for multiplicity.

Comparison	Wins	Losses	Ties
Best V vs BCalErr (15)	3	3	9
Best V vs TS (8 foundation)	0	4	4

features, but those gains are driven primarily by BCalErr on CLIP, not by the proposed V-family.

The V-family's role is therefore diagnostic. It does not establish a universal replacement loss. Instead, it helps reveal that objective choice is representation-dependent: the same margin-sensitive alternatives that are weak on CLIP become competitive with the best probe objective on the WRN-28-10 same-domain settings, where they tie the best aggregate probe.

5.3. CNN Controls Reveal a Margin-Sensitive Regime

The same-domain CNN settings provide a useful contrast for the middle-ground claim (Table 3). At the aggregate level, the broader calibration-aware probe family improves over temperature scaling on all four same-domain CIFAR settings. Because per-seed TS outputs are unavailable for these settings, this comparison is aggregate rather than paired; we therefore use it as evidence for practical value rather than as a paired significance claim.

The internal ranking differs sharply from CLIP. On WRN-28-10 with both CIFAR-10 and CIFAR-100, the proposed confidence- and margin-sensitive family ties the best probe objective in aggregate, matching Best LP exactly. On CIFAR-10 with WRN-28-10, temperature scaling slightly

Table 3. Same-domain CNN controls (test ECE %, lower is better). Comparisons to TS are aggregate because paired per-seed TS outputs are unavailable for these four settings.

Setting	TS	Best LP	Best V
C10 / R110	1.117	**0.817**	0.902
C10 / WRN	1.099	**0.837**	**0.837**
C100 / R110	2.004	**1.439**	1.666
C100 / WRN	3.272	**1.641**	**1.641**

worsens ECE while a V-family objective still improves it. On the ResNet-110 settings, non-V learned probes lead, with the V-family close behind. These cases are consistent with a regime in which runner-up competition remains informative and margin-sensitive variants can exploit it more often than they can on CLIP.

The three cross-domain CNN transfer settings are more heterogeneous and provide weaker evidence. Relative to BCalErr, none of the best proposed V-methods is significantly better on STL-10 / ResNet-50, Skin / ResNet-50, or TinyImageNet / ResNet-50. We therefore treat the transfer CNN settings as supporting evidence for the broad viability of calibration-aware probing rather than as the main source of claims.

The value of the CNN controls is precisely this contrast. The objective family that explains the strongest CLIP gains is not the one that ties best on same-domain WRN-28-10. This cross-regime inversion is what makes the main result scientifically informative rather than merely negative for the V-family.

5.4. Compute Cost

Measured wall-clock time on an RTX 4090 with cached features shows that all probe methods train in roughly 9–10 seconds per seed. Temperature scaling adds approximately 0.4 seconds on top of CE, and peak GPU memory is 32–52 MB across methods. These measurements support the practical claim that frozen-feature probing is cheap compared with end-to-end recalibration, with the explicit caveat that feature extraction is excluded from this accounting. When features are cached, calibration-aware probing is effectively a per-dataset operation rather than a per-experiment one.

6. Discussion

The central contribution of this paper is an empirical regularity: calibration under frozen probing exhibits a clear *representation-family structure*. CLIP favors direct confidence–accuracy correction when learned recalibration helps. Same-domain CNN controls, especially WRN-28-10, are more favorable to the V-family and its margin-sensitive variants. DINOv2 leaves little statistically reliable room beyond temperature scaling. Three qualitatively different regimes emerge from a single matched protocol.

This result has immediate implications for evaluation. A reasonable default recipe for frozen-feature calibration is not to choose one loss and report it once. It is to evaluate a small family of head-level baselines: temperature scaling, a direct confidence–accuracy penalty such as BCalErr, and at least one confidence- or margin-sensitive reweighting objective. Which method wins is itself informative about the representation.

The result also sharpens how this work relates to recent calibration studies of CLIP-like models. Recent prompt-tuning and test-time prompt-tuning papers show that lightweight adaptation can substantially distort confidence (Yoon et al., 2024; Wang et al., 2025; Sharifdeen et al., 2025). Our evaluation complements that literature by showing that even in the simpler supervised linear-probing regime, the appropriate corrective objective depends on the representation family. In other words, the calibration problem does not disappear when adaptation is reduced to a linear head; it becomes a question of which head-level objective matches the frozen features.

We do not offer a formal explanation for the observed regime split, but the pattern suggests hypotheses that future work could test directly. One possibility is that CLIP's residual miscalibration after probing is often dominated by scalar confidence mismatch, so a direct confidence–accuracy penalty is the right tool when TS is insufficient. Same-domain CNNs may retain more calibration-relevant margin information near decision boundaries, which makes margin-sensitive reweighting useful. DINOv2 may induce logits that are already mild enough that temperature scaling removes most residual error. These are hypotheses, not conclusions, but they provide theory-relevant targets for future work.

7. Limitations

Several limitations bound the strength of our claims.

1. Per-seed temperature-scaling outputs are unavailable for the four same-domain CNN settings, so CNN-versus-TS comparisons are aggregate rather than paired.
2. Search budgets are unequal across methods; although all methods share the learning-rate grid, method-specific grids differ in size.
3. Best-LP and Best-V summaries are selected after comparing multiple objectives, so they should be interpreted as descriptive post-selection evidence for regime structure rather than as deployable selection rules or confirmatory hypothesis tests.

4. The compute accounting excludes feature extraction; when backbones must be re-run rather than cached, absolute costs will be higher.

5. The evaluation is in-distribution only; corruption robustness, open-set behavior, and out-of-distribution calibration remain open.

6. The foundation-model set is limited to CLIP ViT-B/16 and DINOv2 ViT-B/14; broader backbone families may reveal additional regimes.

7. ECE depends on binning and is not a proper scoring rule. We use a fixed 15-bin protocol for comparability, but additional metrics such as NLL, Brier score, ACE, or classwise calibration would strengthen future evaluations.

8. V3 and V4 use fixed sigmoid shape parameters in the main experiments. Learning or selecting these parameters from representation statistics is a natural extension of the V-family.

These limitations do not overturn the main empirical regularity — that calibration objective choice is representation-dependent — but they do bound the strength of what we claim. The paper offers a structured evaluation and a set of comparative findings, not a complete theory or a formal guarantee.

8. Conclusion

We studied how training-time calibration objectives transfer to frozen foundation-model features through calibration-aware linear probing. Relearning the head is often stronger than post-hoc temperature scaling, but the central result is not a universal winning loss. It is that CLIP, DINOv2, and same-domain CNNs occupy different calibration regimes, and that the preferred objective changes with the representation.

For frozen-feature evaluation, the implication is straightforward: uncertainty should be evaluated with a small family of calibration objectives rather than reduced to one default correction. Calibration-aware probing is therefore useful both as a lightweight recalibration tool and as a diagnostic tool for understanding how frozen representations encode confidence and decision-boundary uncertainty.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Huang, Y., Shakeri, F., Dolz, J., Boudiaf, M., Bahig, H., and Ben Ayed, I. LP++: A surprisingly strong linear probe for few-shot CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23773–23782, 2024.

Jordahn, M. and Olmos, P. M. Decoupling feature extraction and classification layers for calibrated neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22530–22550. PMLR, 2024. URL https://proceedings.mlr.press/v235/jordahn24a.html.

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

Kumar, A., Sarawagi, S., and Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2805–2814. PMLR, 2018.

LeVine, W., Pikus, B., Raja, P., and Amat Gil, F. Enabling calibration in the zero-shot inference of large vision-language models. *arXiv preprint arXiv:2303.12748*, 2023.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. H. S., and Dokania, P. K. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? *Advances in Neural Information Processing Systems*, 32, 2019.

Ni, J., Zhao, H., Gao, J., Guo, D., and Zha, H. Balancing two classifiers via a simplex etf structure for model calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30712–30721, June 2025.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Sharifdeen, A., Munir, M. A., Baliah, S., Khan, S., and Khan, M. H. O-TPT: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Tao, L., Dong, M., and Xu, C. Dual focal loss for calibration. In *International Conference on Machine Learning*, pp. 33833–33849. PMLR, 2023.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Tu, W., Deng, W., Campbell, D., Gould, S., and Gedeon, T. An empirical study into what matters for calibrating vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48791–48808. PMLR, 2024. URL https://proceedings.mlr.press/v235/tu24a.html.

Wang, D.-B. and Zhang, M.-L. Calibration bottleneck: Over-compressed representations are less calibratable. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52156–52170. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24cm.html.

Wang, S., Li, Y., and Wei, H. Understanding and mitigating miscalibration in prompt tuning for vision-language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63467–63489. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wang25bd.html.

Yoon, H. S., Yoon, E., Tee, J. T. J., Hasegawa-Johnson, M., Li, Y., and Yoo, C. D. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jzzEHTBFOT.

A. Hyperparameter Grids

All methods share the learning-rate grid $\{10^{-4}, 3\times10^{-4}, 5\times10^{-4}, 10^{-3}, 3\times10^{-3}, 5\times10^{-3}\}$. Method-specific grids used in the main experiments are listed in Table 4.

Table 4. Method-specific hyperparameter grids.

Method	Grid
V1	$\gamma \in \{1.5, 2, 3, 4, 5\}$, $\tau \in \{1.5, 2, 3, 4, 5\}$
V2	$\gamma \in \{1.5, 2, 3, 4, 5\}$, $\tau \in \{1.5, 2, 3, 4, 5\}$
V3	$\gamma \in \{0.5, 1, 1.5, 2, 3\}$, $\tau \in \{1, 2, 3, 5, 10\}$
V4	$\gamma \in \{0.5, 1, 1.5, 2, 3\}$, $\tau \in \{1, 2, 3, 5\}$
Focal Loss	$\gamma \in \{0.5, 1, 1.5, 2, 3, 4, 5, 7, 10\}$
Label Smoothing	$\epsilon \in \{0.01, 0.02, 0.03, 0.05, 0.07, 0.1, 0.15, 0.2\}$
Mixup	$\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0\}$
BCalErr	$\lambda \in \{0.1, 0.2, 0.5, 1, 2, 3, 5, 10\}$
MDFL	$\gamma \in \{1.5, 2, 2.5, 3, 3.5, 3.9, 4.5, 5, 6.1\}$

B. Implementation Note on V2

Our implementation of V2 used the equivalent verbose form

$$\big[\min\big(1-p_y,\ 1-\text{clamp}(m,0,1)\big)\big]^{\gamma}\cdot \text{CE}(z,y), \quad m = p_y - \max_{c\neq y} p_c.$$

Because $\max_{c\neq y} p_c \geq 0$, we have $m \leq p_y$, so:

- when $m > 0$: $1-m \geq 1-p_y$, hence $\min(1-p_y,\ 1-m) = 1-p_y$;
- when $m \leq 0$: $1-\text{clamp}(m,0,1) = 1 \geq 1-p_y$, hence the min is again $1-p_y$.

The displayed min rule therefore reduces to $(1-p_y)^{\gamma}\cdot \text{CE}(z,y)$ for all valid inputs. We document the implemented form here only for code correspondence; it has no behavioral effect on training. The main text uses the reduced form throughout.

C. Significance Outcomes

For the best proposed V-method selected per setting under the macro-selection protocol:

- **Significant wins over BCalErr**: CIFAR-10 / ResNet-110, CIFAR-10 / DINOv2, and TinyImageNet / DINOv2.
- **Significant losses to BCalErr**: CIFAR-100 / CLIP, STL-10 / CLIP, and TinyImageNet / CLIP.
- **Non-significant**: the remaining nine settings.
- **Against temperature scaling on the eight foundation settings**: zero significant wins, four significant losses (all CLIP settings: CIFAR-10 / CLIP, CIFAR-100 / CLIP, STL-10 / CLIP, TinyImageNet / CLIP), and four ties (all DINOv2 settings).