DEBIASED GENERALIZED CALIBRATION MOMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern neural networks can be accurate but poorly calibrated. We revisit train-time multiclass calibration through a generalized moment view that unifies binning, mean-matching, and kernel objectives as fixed witness function choices on the probability simplex, and we ask whether the witness itself can be learned. We propose *Debiased Generalized Calibration Moments* (D-GCM), a lightweight min-max regularizer that trains a small critic on predicted probabilities to expose miscalibration patterns. The key challenge is empirical estimation: the naive squared batch moment is a biased V-statistic whose diagonal self-interaction term biases the objective for any fixed witness and becomes an optimization shortcut for a learned critic, rewarding sample-wise residual energy even when the target calibration moment is zero. D-GCM removes this shortcut with the corresponding U-statistic over distinct sample pairs. We theoretically show that rich witness families characterize strong multiclass calibration, while restricted families control only the miscalibration components they can approximate. Empirically, the debiased learned-critic objective achieves the best calibration among train-time calibration regularizers on four image and text benchmarks while maintaining competitive prediction accuracy, adds modest computational overhead, and remains complementary to post-hoc temperature scaling.

1 INTRODUCTION

Modern neural networks can achieve high accuracy while producing over-confident and unreliable class probabilities. This matters whenever the prediction score is used as a probability forecast, including abstention, selective prediction, ranking, and downstream decision-making under uncertainty (Guo et al., 2017; Mukhoti et al., 2020). Although cross-entropy is a proper scoring rule in population, minimizing it with over-parameterized deep networks does not reliably produce calibrated posteriors in practice (Guo et al., 2017; Liu et al., 2022). This has motivated post-hoc recalibration, indirect train-time confidence shaping, and explicit train-time objectives that penalize calibration violations (Guo et al., 2017; Mukhoti et al., 2020; Karandikar et al., 2021; Kumar et al., 2018; Hebbalaguppe et al., 2022; Liu et al., 2022; Lin et al., 2025; Widmann et al., 2019; Marx et al., 2023).

A common design choice in train-time calibration is to decide *in advance* how miscalibration will be observed. Soft-binned ECE losses use hand-designed binning functions (Karandikar et al., 2021), MDCA matches low-order classwise averages (Hebbalaguppe et al., 2022), and MMCE, KCE, and distribution-matching methods use a fixed RKHS or MMD geometry on the prediction space (Kumar et al., 2018; Widmann et al., 2019; Marx et al., 2023). These approaches can work well, but they all fix the *witness*, the test function of the predicted probability vector against which the label residual is checked, before seeing the model's actual calibration errors (Figure 1 (a)). A fixed witness spreads its capacity uniformly across a basis, a set of bins, or a kernel geometry. Under the same compute budget, a *learnable* witness can focus on the regions of the probability simplex where the current predictor is miscalibrated.

Strong multiclass calibration naturally suggests such a witness-based view. Once the model reports a probability vector, no systematic component of the label residual, the difference between the one-hot label and the reported vector, should remain predictable from that same vector. This makes calibration a conditional-moment problem on the probability simplex: every witness on the reported probability vector defines a calibration moment, and many familiar train-time losses differ mainly in which witness family they fix in advance. Prior fixed-kernel methods already use this moment view with

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

a pre-specified witness family. Our key point is that the same view provides the right interface for *adaptive* witness learning.

We instantiate this adaptive witness as a neural critic on the probability vector. A naive implementation would maximize the squared mini-batch moment over this critic. This squared batch moment is a V-statistic: it averages residual–witness products over all pairs of samples, including each sample paired with itself. For a fixed witness, these diagonal self-interactions bias the objective through a term that depends on the predictor. For a learned critic, they also become an optimization shortcut: the critic can increase the objective by emphasizing examples with large residual energy (squared residual norm) even when no cross-sample calibration structure exists, and under perfect calibration the V-statistic still gives strictly positive expected reward whenever labels remain uncertain. The key issue is therefore how to debias the objective used to train the witness.

To this end, we propose *Debiased Generalized Calibration Moments* (D-GCM), a train-time regularizer applicable to standard classification losses. D-GCM uses a debiased U-statistic over distinct sample pairs and admits both a learned-critic instantiation, *D-GCM-Critic*, and a fixed-basis instantiation, *D-GCM-Basis*. The critic version is our main method: a small network on the probability simplex adaptively searches for miscalibration patterns while observing only the reported probability vector. This preserves the idea that calibration should be readable from the forecast itself, while the U-statistic removes the critic's self-interaction shortcut and leaves only cross-sample residual-witness alignment.

Contributions. We formulate train-time multiclass calibration through a *conditional-moment view* on the probability simplex. This view gives a common framework for several fixed-statistic calibration objectives and motivates replacing a pre-specified witness family with an adaptive critic. We identify a failure mode of naive learned-moment objectives: the diagonal of the V-statistic creates a self-interaction shortcut that can reward the critic even at perfect calibration. We remove this shortcut with an efficient U-statistic objective, yielding learned-critic and fixed-basis instantiations. We also provide theory aligned with the method: rich witness families characterize strong calibration, while practical restricted families control only the miscalibration components they can approximate. Empirically, our D-GCM acts as an efficient plug-and-play regularizer, achieving state-of-the-art multiclass calibration while maintaining competitive accuracy, and it remains complementary to post-hoc temperature scaling.

2 Related Work

Post-hoc and indirect train-time calibration. Temperature scaling is a standard post-hoc baseline, often improving calibration by learning a single logit-rescaling parameter without changing the predicted class (Guo et al., 2017). A complementary train-time line improves calibration indirectly by reshaping confidence, margins, or gradients: focal loss (Mukhoti et al., 2020), margin-based label smoothing (Liu et al., 2022), BSCE-GRA (Lin et al., 2025), and frequency-aware gradient rectification (FGR), which projects the training gradient so that in-distribution calibration is preserved while training for robustness under distribution shift (Zhang et al., 2025). Meta-regularization takes one step toward learning: a meta-network outputs a sample-wise focal exponent and is trained against a fixed kernel-smoothed ECE surrogate (Wang & Golebiowski, 2023). Across this line, learning enters only through a rescaling parameter, a loss hyperparameter, or a gradient correction, and the calibration statistic, where one is used at all, is fixed in advance.

Explicit objectives with fixed witnesses. Other methods choose a calibration statistic in advance and optimize a differentiable surrogate. Soft Calibration smooths histogram bins to obtain trainable ECE-like losses (Karandikar et al., 2021), while MDCA matches average predicted confidence to empirical class frequency within a mini-batch (Hebbalaguppe et al., 2022). Function-class approaches are closest to ours: MMCE defines an RKHS moment objective for top-label confidence calibration (Kumar et al., 2018), Widmann et al. (2019) extend kernel calibration to strong multiclass calibration on the full simplex, and Marx et al. (2023) frame calibration as MMD-based distribution matching. D-GCM builds on this moment view and learns the witness itself as an adversarial critic. The diagonal term of the naive squared batch moment biases fixed-witness objectives and becomes an optimization shortcut for a learned witness, and our U-statistic objective removes it, which makes learned-witness calibration stable and well posed.

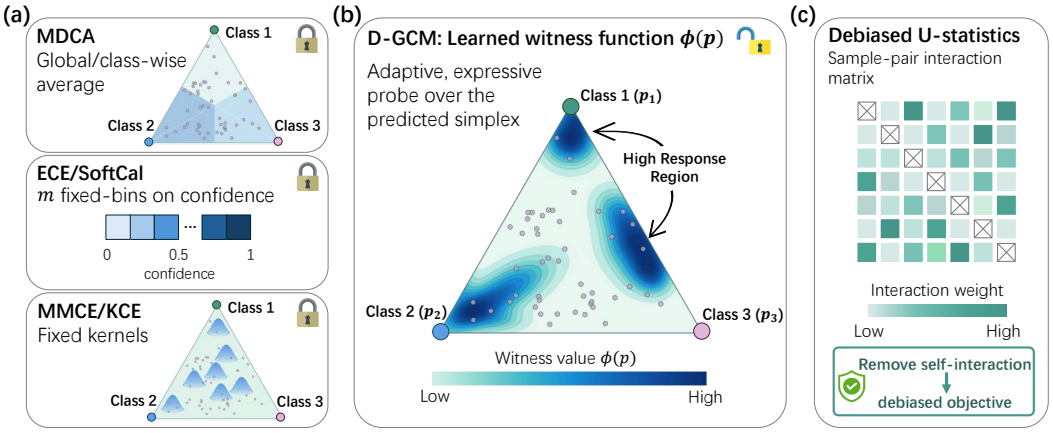


Figure 1: **Overview of Generalized Calibration Moments. (a)** Existing methods (e.g., MDCA (Hebbalaguppe et al., 2022), SoftCal (Karandikar et al., 2021), MMCE (Kumar et al., 2018)) act as predefined, fixed-statistic witnesses. **(b)** Our proposed D-GCM learns an adaptive witness (critic) on the predicted probability simplex to dynamically expose miscalibration. **(c)** The naive batch moment creates an optimization shortcut via diagonal self-interactions, which D-GCM eliminates by removing the diagonal, yielding a debiased U-statistic.

3 Generalized Calibration Moments

Setup and moment view. Let $P_\theta := p_\theta(X) \in \Delta^{K-1}$ denote the predicted class-probability vector, let $r_\theta(X,Y) := e_Y - P_\theta \in \mathbb{R}^K$ be the multiclass residual, and let μ_θ denote the distribution of P_θ on Δ^{K-1}. Strong multiclass calibration is the usual condition $\mathbb{E}[e_Y \mid P_\theta] = P_\theta$, equivalently

$$\mathbb{E}[r_\theta(X,Y) \mid P_\theta] = 0. \tag{1}$$

For a scalar witness (test) function $h : \Delta^{K-1} \to \mathbb{R}$, define

$$m_\theta(h) := \mathbb{E}[r_\theta(X,Y)\, h(P_\theta)] \in \mathbb{R}^K.$$

For a vector witness function $\phi = (\phi_1, \ldots, \phi_d) : \Delta^{K-1} \to \mathbb{R}^d$, define

$$M_\theta(\phi) := \mathbb{E}[r_\theta(X,Y)\phi(P_\theta)^\top] \in \mathbb{R}^{K\times d},$$

so that $\|M_\theta(\phi)\|_F^2 = \sum_{j=1}^d \|m_\theta(\phi_j)\|_2^2$. The scalar moments are convenient for theory, while the vector form is the quantity optimized by D-GCM. In words, a witness observes only the reported probability vector and tests whether the label residual still has a systematic component that the witness can predict.

Why learn the witness? Fixed witnesses are simple and analyzable, but they hard-code where the objective looks for miscalibration. Constant witnesses detect only global class-frequency mismatch, soft bins encode a hand-designed partition of the prediction space, and fixed kernels impose a predetermined notion of similarity on the simplex. Under the same compute budget, a learned critic allocates its limited capacity to the regions and directions where the current predictor exhibits systematic residual structure (Figure 1 (b)).

Population objective. Let Φ be a family of vector witnesses $\phi : \Delta^{K-1} \to \mathbb{R}^d$ satisfying the pointwise constraint

$$\sup_{q\in\Delta^{K-1}} \|\phi(q)\|_2 \le 1.$$

We define the worst-case calibration moment violation

$$\mathcal{C}_\Phi(\theta) := \sup_{\phi\in\Phi} \|M_\theta(\phi)\|_F^2, \tag{2}$$

where $M_\theta(\phi) = \mathbb{E}[r_\theta(X,Y)\phi(P_\theta)^\top]$. A strongly calibrated predictor makes the residual $r_\theta(X,Y)$ unpredictable from P_θ, so $\mathcal{C}_\Phi(\theta) = 0$ exactly when every witness in Φ sees zero moment violation. We therefore train with

$$\min_\theta \; \mathcal{L}_{\mathrm{CLS}}(\theta) + \lambda\, \mathcal{C}_\Phi(\theta), \tag{3}$$

Table 1: Unification of calibration methods under the generalized calibration moments framework. D-GCM shifts the paradigm from fixed-statistic witnesses to learned critics, and this requires a debiased U-statistic to resolve the resulting finite-batch optimization shortcut.

Method Family	Witness Choice (ϕ)	Fixed / Learned	Estimator
MDCA	Constant / classwise moments	Fixed	Batch moment / V-type
ECE / Soft Calibration	Hard / soft bin memberships	Fixed	Binned batch moment
MMCE	RKHS kernel / feature map	Fixed	V- or U-statistic
KCE	Vector-valued RKHS on full simplex	Fixed	Pairwise V- or U-statistic
D-GCM-Basis (Ours)	Basis on probability simplex	Fixed	**Debiased U-statistic**
D-GCM-Critic (Ours)	Small neural witness network	**Learned**	**Debiased U-statistic**

where $\mathcal{L}_{\mathrm{CLS}}$ is the classification loss (e.g., cross-entropy). The unit-ball constraint prevents arbitrary rescaling of the witness and puts all candidates on a common scale. The witness dimension d is a user-chosen capacity parameter, and a small bottleneck such as $d = 32$ suffices in practice.

Two instantiations. We consider two concrete choices for Φ. In our primary method, ***D-GCM-Critic***, Φ is a family of neural witnesses ϕ_ψ parameterized by weights ψ, trained adversarially with the classifier to adaptively discover miscalibration. As an ablation and non-adversarial baseline, we also define ***D-GCM-Basis***, where Φ is a fixed bounded basis on the simplex.

A unifying view of prior fixed objectives. Choosing Φ appropriately recovers many prior train-time calibration losses. Constant witnesses yield MDCA-style mean matching (Hebbalaguppe et al., 2022), hard or soft confidence bins yield ECE- or soft-binning-style moment losses (Karandikar et al., 2021), fixed RKHS features recover MMCE (Kumar et al., 2018), and vector-valued kernels recover KCE-type objectives (Widmann et al., 2019). These objectives share one moment recipe once the residual projection, outer aggregation, and finite-sample estimator are specified, and this shared recipe is the interface for our adaptive witness. Figure 1 (a) and Table 1 summarize the unification, and Appendix B.3 makes each correspondence explicit.

4 Train-Time Optimization via U-Statistics

The central technical issue is the mini-batch estimator of Equation (2). For fixed witnesses, biased and unbiased pairwise estimators are both standard choices. With a learned critic, the estimator changes the optimization problem itself.

Naive squared batch moment. Given a mini-batch $\{(x_i, y_i)\}_{i=1}^B$, let $p_i = p_\theta(x_i)$, $r_i = e_{y_i} - p_i$, and $\phi_i = \phi(p_i)$. A natural surrogate is

$$\widehat{\mathcal{C}}_V(\theta, \phi) := \left\| \frac{1}{B} \sum_{i=1}^{B} r_i \phi_i^\top \right\|_F^2 . \tag{4}$$

Expanding the square yields

$$\widehat{\mathcal{C}}_V(\theta, \phi) = \frac{1}{B^2} \sum_{i \neq j} (r_i^\top r_j)(\phi_i^\top \phi_j) + \frac{1}{B^2} \sum_{i=1}^{B} \|r_i\|_2^2 \, \|\phi_i\|_2^2 . \tag{5}$$

The first off-diagonal term couples distinct examples and estimates the desired squared calibration moment. The second diagonal term is a self-interaction (Figure 1 (c)). In expectation,

$$\mathbb{E}\Big[\widehat{\mathcal{C}}_V(\theta, \phi)\Big] = \|\mathbb{E}[Z]\|_2^2 + \frac{1}{B} \operatorname{tr}\big(\operatorname{Var}(Z)\big), \quad Z = \operatorname{vec}\big(r_\theta(X, Y)\phi(P_\theta)^\top\big), \tag{6}$$

so the diagonal contributes a nonnegative $\mathcal{O}(1/B)$ bias that depends on the predictor through the residual energy and pulls training away from the calibration target even for a fixed witness. For a learned critic, it also becomes an optimization shortcut. Under the constraint $\|\phi(q)\|_2 \leq 1$, the critic can increase $\widehat{\mathcal{C}}_V$ by saturating its norm on regions of the simplex with large conditional residual

energy $\mathbb{E}[\|R\|_2^2 \mid P]$, even if no cross-sample residual-witness alignment remains. In other words, the critic can act like a hard-example miner instead of finding calibration structure, and Proposition 1 makes this precise.

Debiased U-statistic. We therefore replace (4) by the debiased U-statistic

$$\widehat{\mathcal{C}}_U(\theta, \phi) := \frac{1}{B(B-1)} \sum_{i \neq j} (r_i^\top r_j)(\phi_i^\top \phi_j). \tag{7}$$

For a fixed *sample-wise* witness, one evaluated independently on each example, Proposition 3 (Section 5) shows that $\widehat{\mathcal{C}}_U$ is unbiased for $\|M_\theta(\phi)\|_F^2$. Its role here is more than debiasing in the usual evaluation sense: it removes the critic's self-interaction shortcut and forces the inner maximization to uncover cross-sample calibration structure.

The estimator also admits an efficient implementation without forming a $B \times B$ interaction matrix. Let $S_B := \sum_{i=1}^B r_i \phi_i^\top \in \mathbb{R}^{K \times d}$. Then

$$\widehat{\mathcal{C}}_U(\theta, \phi) = \frac{\|S_B\|_F^2 - \sum_{i=1}^B \|r_i\|_2^2 \|\phi_i\|_2^2}{B(B-1)}. \tag{8}$$

Thus the U-statistic is computed by taking the total residual-witness energy and subtracting the self-matches. Although $\widehat{\mathcal{C}}_U$ can be negative on an individual mini-batch, it is lower-bounded by $-2/(B-1)$ (Appendix B.2), so we leave it unclamped to preserve its debiasing properties.

Proposition 1 (Expected diagonal shortcut of the V-statistic)**.** *Let $R := r_\theta(X, Y)$ and $P_\theta := p_\theta(X)$. For any fixed sample-wise witness ϕ satisfying $\mathbb{E}\big[\|R\|_2^2 \|\phi(P_\theta)\|_2^2\big] < \infty$, the naive estimator satisfies*

$$\mathbb{E}\Big[\widehat{\mathcal{C}}_V(\theta, \phi)\Big] = \left(1 - \frac{1}{B}\right) \|M_\theta(\phi)\|_F^2 + \frac{1}{B}\, \mathbb{E}\big[\|R\|_2^2 \|\, \phi(P_\theta)\|_2^2\big]\,.$$

Consequently, if p_θ is strongly calibrated, then under the constraint $\sup_q \|\phi(q)\|_2 \leq 1$,

$$\sup_\phi \mathbb{E}\Big[\widehat{\mathcal{C}}_V(\theta, \phi)\Big] = \frac{1}{B}\, \mathbb{E}\big[\|R\|_2^2\big]\,, \quad \sup_\phi \mathbb{E}\Big[\widehat{\mathcal{C}}_U(\theta, \phi)\Big] = 0.$$

Takeaway. Even at perfect calibration, the V-statistic gives a learned witness strictly positive expected reward whenever the task retains label uncertainty. The U-statistic removes this self-interaction reward, and the shortcut is present for every fixed witness before any critic optimization takes place.

Learned critic. In D-GCM-Critic, we parameterize the witness function by a small network $h_\psi : \Delta^{K-1} \to \mathbb{R}^d$ and project its output to the unit ball,

$$\phi_\psi(q) = \frac{h_\psi(q)}{\max\{1, \|h_\psi(q)\|_2\}}. \tag{9}$$

The critic observes only the probability vector, not the original input or hidden features. This keeps the method aligned with the calibration target itself: the critic must detect patterns that are visible from the predicted probabilities. To preserve the exact U-statistic interpretation, the critic is a sample-wise network without batch-dependent layers.

The resulting training problem is

$$\min_\theta \max_\psi \; \mathcal{L}_{\mathrm{CLS}}(\theta) + \lambda\, \widehat{\mathcal{C}}_U(\theta, \phi_\psi). \tag{10}$$

We alternate one critic ascent step and one model descent step. In the critic step, $p_\theta(x)$ is detached so that only critic parameters ψ are updated. In the model step, ψ is frozen but the computation $\phi_\psi(p_\theta(x))$ remains in the graph, so gradients flow through both the residual r_θ and the current witness response. This is the correct derivative of the current saddle objective with ψ held fixed, and it coincides with the Danskin/envelope gradient for the outer problem when the critic is at (or closely tracks) its inner maximizer. The extra cost is dominated by forming $S_B = R^\top \Phi$ with $R \in \mathbb{R}^{B \times K}$ and $\Phi \in \mathbb{R}^{B \times d}$, which costs $\mathcal{O}(BKd)$ and is modest because d is user-chosen and can be kept small. Figure 2 gives a minimal PyTorch (Paszke et al., 2019) implementation of both steps.

```
# p: probs [B, K], y: labels [B], cls_loss: classification loss (e.g., cross-entropy)
B, K = p.shape
R = F.one_hot(y, num_classes=K).float() - p
phi_raw = critic(p)
Phi = phi_raw / phi_raw.norm(dim=1, keepdim=True).clamp_min(1.0)
S = R.T @ Phi
rsq = R.square().sum(dim=1)
psq = Phi.square().sum(dim=1)
diag = (rsq * psq).sum()
u_stat = (S.square().sum() - diag) / (B * (B - 1))
loss = cls_loss + lam * u_stat
```

Figure 2: Minimal PyTorch implementation of D-GCM-Critic. We use `critic(p.detach())` and maximize `u_stat` in the critic step. In the model step, use `critic(p)` and minimize `cls_loss + lam * u_stat`.

Fixed-basis variant. For D-GCM-Basis, the critic network is replaced by a predefined bounded feature map ϕ_0. Training then uses the same U-statistic in Equation (8) without the inner maximization step, which cleanly isolates the gain of debiasing from the adaptive witness learning.

5 Theoretical Study

The scope of the guarantees is set by the richness of the witness family: a rich family characterizes strong calibration exactly, and a restricted family controls the miscalibration components it can represent. The scalar moments $m_\theta(h)$ introduced in Section 3 are convenient for this analysis, while the vector-witness objective of D-GCM simply bundles finitely many such scalar witnesses. All proofs are deferred to Appendix B.

Theorem 1 (Conditional moments characterize strong calibration)**.** *Fix θ and let $\mathcal{H} \subset L^2(\mu_\theta)$ be a witness family whose linear span is dense in $L^2(\mu_\theta)$. Then p_θ is strongly calibrated if and only if $m_\theta(h) = 0$ for every $h \in \mathcal{H}$.*

Takeaway. Zero moments indicate strong calibration globally only when the witness family is rich enough to approximate arbitrary square-integrable functions of the predicted probability vector. A sufficient distribution-free condition is that the span of $\mathcal{H}$ be dense in $C(\Delta^{K-1})$, since continuous functions are dense in $L^2(\mu_\theta)$ on the compact simplex.

Proposition 2 (Restricted-family control)**.** *Let*

$$g_\theta(P_\theta) := \mathbb{E}[r_\theta(X, Y) \mid P_\theta] \in L^2(\mu_\theta; \mathbb{R}^K)$$

be the conditional calibration residual, and let $\mathcal{H}_m \subset L^2(\mu_\theta)$ be any witness class. If $m_\theta(h) = 0$ for every $h \in \mathcal{H}_m$, then for each coordinate k and every $h \in \mathcal{H}_m$,

$$\|g_{\theta,k}\|_{L^2(\mu_\theta)} \le \|g_{\theta,k} - h\|_{L^2(\mu_\theta)}. \tag{11}$$

In particular, if for each k there exists $h_k \in \mathcal{H}_m$ such that $\|g_{\theta,k} - h_k\|_{L^2(\mu_\theta)} \le \varepsilon_k$, then

$$\|g_\theta\|_{L^2(\mu_\theta;\mathbb{R}^K)} \le \Big(\sum_{k=1}^{K} \varepsilon_k^2\Big)^{1/2}.$$

Takeaway. A limited witness family controls the components of miscalibration that lie near the function class it can represent, and exact certification requires the rich families of Theorem 1.

Proposition 3 (Unbiased empirical objective for a fixed witness)**.** *Assume that $\{(X_i, Y_i)\}_{i=1}^B$ are i.i.d., that $p_\theta(x)$ and $\phi(p_\theta(x))$ are sample-wise functions of each example and that*

$$\mathbb{E}\big[\|r_\theta(X, Y)\|_2^2 \|\phi(P_\theta)\|_2^2\big] < \infty.$$

Then

$$\mathbb{E}\Big[\widehat{\mathcal{C}}_U(\theta, \phi)\Big] = \|M_\theta(\phi)\|_F^2. \tag{12}$$

In contrast, the naive estimator $\widehat{\mathcal{C}}_V$ in (4) *satisfies the bias formula in* (6)*.*

Takeaway. For a fixed sample-wise witness, the U-statistic matches the intended population objective exactly, whereas the naive squared batch mean systematically adds a nonnegative $\mathcal{O}(1/B)$ diagonal penalty.

Table 2: Main results on four benchmarks following the MDCA protocol. Acc and ECE are in percent (%) and SCE is in units of 10^{-3}. Metrics are means over 3 random seeds (standard deviations in Appendix C.5). Best results among calibration regularizers are bolded and second-best are underlined.

	CIFAR-10 (R56)			CIFAR-100 (R56)			Tiny-ImageNet (R34)			20 Newsgroups		
Method	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓
Naive CE/NLL	94.23	3.61	7.76	75.10	9.50	2.93	65.37	10.82	2.41	67.93	14.92	6.36
Focal Loss (Lin et al., 2017)	94.58	1.21	4.02	75.53	1.97	1.78	66.58	2.33	1.35	69.20	10.83	4.62
DCA (Liang et al., 2020)	93.31	3.12	7.91	63.68	9.29	2.57	63.20	13.17	2.04	62.93	11.35	3.07
MMCE (Kumar et al., 2018)	91.98	3.40	8.63	71.34	8.61	2.18	58.85	5.86	1.85	**67.90**	10.46	2.84
FLSD (Mukhoti et al., 2020)	92.65	3.29	6.89	71.49	1.90	1.91	63.14	2.18	1.33	67.06	5.37	3.30
MDCA (Hebbalaguppe et al., 2022)	92.92	0.72	3.30	72.07	0.72	1.77	62.77	1.73	1.16	63.28	10.08	1.91
SoftCal (Karandikar et al., 2021)	91.30	1.66	5.22	71.22	1.57	2.46	61.08	2.57	1.68	61.04	8.17	3.65
MbLS (Liu et al., 2022)	93.34	1.24	4.06	70.86	2.45	2.90	63.96	1.64	1.37	63.89	5.40	3.88
ESD (Yoon et al., 2023)	92.88	1.65	5.84	67.40	4.39	2.61	60.52	3.12	2.40	65.77	7.68	2.35
BSCE-GRA (Lin et al., 2025)	93.41	0.73	3.79	73.01	1.36	1.88	64.33	1.45	1.58	66.08	5.51	1.46
CALS (Liu et al., 2023)	92.17	1.73	4.88	70.35	4.38	2.39	64.18	1.54	1.96	64.84	4.32	2.15
DECE (Bohdal et al., 2021)	93.26	1.09	5.15	72.14	3.07	1.96	62.79	1.77	1.32	63.87	4.13	2.07
FGR (Zhang et al., 2025)	93.37	0.84	4.37	72.47	1.08	1.94	63.67	1.33	1.46	–	–	–
D-GCM-Basis	93.07	0.70	3.41	72.16	0.76	1.80	63.06	1.70	1.28	65.52	4.65	1.89
D-GCM-Critic	**93.52**	**0.48**	**3.27**	**73.10**	**0.69**	**1.63**	**64.53**	**1.08**	**1.04**	66.21	**3.90**	**1.11**

6 EXPERIMENTS

We use empirical evaluation to answer three questions: (1) Does adaptive witness learning improve over fixed-statistic calibration? (2) Does the debiased U-statistic successfully prevent the expected failure mode of the learned critic? (3) Can D-GCM act as an efficient, plug-and-play regularizer?

6.1 EXPERIMENTAL SETTINGS

Datasets and Architectures. We strictly follow the experimental protocol of the MDCA baseline (Hebbalaguppe et al., 2022) to isolate the effect of the calibration objective. We evaluate on three image classification tasks, CIFAR-10 and CIFAR-100 (ResNet-56) (Krizhevsky et al., 2009) and Tiny-ImageNet (ResNet-34) (Deng et al., 2009), and one text classification task, 20 Newsgroups (Global-Pool CNN) (Mitchell, 1997). Data is split into $90\%/10\%$ train/validation subsets, with the validation split used for model selection and for methods requiring a calibration signal.

Evaluation Metrics. We report top-1 accuracy (Acc), Expected Calibration Error (ECE) (Posocco & Bonnefoy, 2021; Guo et al., 2017), and Static Calibration Error (SCE) (Nixon et al., 2019). ECE evaluates the confidence-accuracy gap on the top predicted class, whereas SCE averages the calibration error over all class probabilities and is therefore more aligned with multiclass calibration. Following MDCA, both metrics use $M = 15$ bins. Acc and ECE are reported in percent and SCE in units of 10^{-3}, since SCE averages the per-class errors over all K classes and is an order of magnitude smaller. All metrics are averaged over 3 random seeds, and full details are given in Appendix C.

Baselines. We compare against three groups of calibration methods. Indirect train-time methods shape the confidence distribution through adjusted margins or gradients: FLSD (Mukhoti et al., 2020), Margin-based Label Smoothing (MbLS) (Liu et al., 2022), CALS (Liu et al., 2023), BSCE-GRA (Lin et al., 2025), and FGR (Zhang et al., 2025), whose frequency filtering is defined on image inputs, so it is evaluated on the three image benchmarks. Explicit fixed-witness regularizers penalize predefined calibration errors on the simplex: DCA (Liang et al., 2020), MMCE (Kumar et al., 2018), SoftCal (Karandikar et al., 2021), MDCA (Hebbalaguppe et al., 2022), and ESD (Yoon et al., 2023). DECE (Bohdal et al., 2021) optimizes a differentiable ECE surrogate on the validation set by meta-learning. For a fair comparison with MDCA (which finds focal loss to be the strongest base loss), D-GCM-Critic and D-GCM-Basis are applied on top of Focal Loss in the main benchmark.

6.2 MAIN RESULTS

Table 2 shows that D-GCM-Critic establishes a new state-of-the-art across all four datasets, consistently achieving the lowest ECE and SCE. The most informative comparison is to MDCA, the closest train-time multiclass baseline. Relative to MDCA, D-GCM-Critic improves ECE/SCE from

Table 3: Ablation on the estimator. All settings use the same backbone, witness dimension, optimizer, batch size, and regularization weight. Acc and ECE in %, SCE in 10^{-3}. The U-statistic brings the largest gain when the witness is learned, and the gap is smaller for fixed witnesses.

Witness family	Method	CIFAR-10			Tiny-ImageNet		
		Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓
Learned Critic	V-GCM-Critic	93.11	0.66	3.38	64.61	1.46	1.32
Learned Critic	D-GCM-Critic	93.52	0.48	3.27	64.53	1.08	1.04
Fixed Basis	V-GCM-Basis	92.94	0.84	3.50	63.27	1.83	1.35
Fixed Basis	D-GCM-Basis	93.07	0.70	3.41	63.06	1.70	1.28

Table 4: Plug-and-play effect of D-GCM-Critic on different base classification losses. For each base loss, we report ECE (%) and SCE (10^{-3}) before and after adding D-GCM. Lower is better.

	CIFAR-10		CIFAR-100		Tiny-ImageNet		20 Newsgroups	
Base loss	ECE	SCE	ECE	SCE	ECE	SCE	ECE	SCE
Cross-Entropy	3.61	7.76	9.50	2.93	10.82	2.41	14.92	6.36
+D-GCM-Critic	2.64	6.15	8.61	2.43	8.87	1.82	10.14	2.99
Label Smoothing (Szegedy et al., 2016)	4.98	11.63	9.15	2.37	6.42	1.57	4.19	8.86
+D-GCM-Critic	2.37	5.77	7.33	2.06	5.89	1.51	3.72	6.00
Focal Loss (Lin et al., 2017)	1.21	4.02	1.97	1.78	2.33	1.35	10.83	4.62
+D-GCM-Critic	0.48	3.27	0.69	1.63	1.08	1.04	3.90	1.11

$0.72/3.30$ to $0.48/3.27$ on CIFAR-10, from $0.72/1.77$ to $0.69/1.63$ on CIFAR-100, from $1.73/1.16$ to $1.08/1.04$ on Tiny-ImageNet, and from $10.08/1.91$ to $3.90/1.11$ on 20 Newsgroups, and it is also more accurate than MDCA by $0.60/1.03/1.76/2.93$ points on the four datasets. Train-time regularizers trade a portion of accuracy for calibration, and D-GCM-Critic attains the best accuracy among all competing calibration methods on the three vision datasets (93.52% on CIFAR-10, 73.10% on CIFAR-100, and 64.53% on Tiny-ImageNet) and top-3 accuracy on 20 Newsgroups (66.21%). It also improves on the two recent gradient-based baselines, BSCE-GRA and FGR, on every reported metric. This smaller accuracy cost is consistent with two design choices: once the calibration moment is zero, the debiased objective exerts no diagonal pressure on residual variance (Proposition 1), and the learned witness penalizes only residual structure that remains predictable from the probability vector, whereas MDCA's global classwise correction acts on every example alike.

The calibration advantage holds beyond the $M = 15$ evaluation grid and without any post-hoc step. Re-evaluating the saved CIFAR-10 predictions (Appendix C.6), D-GCM-Critic has lower ECE and SCE than Focal Loss and MDCA for $M \in \{10, 15, 30, 50\}$, lower ECE under equal-mass binning, and a lower NLL than MDCA (0.207 vs. 0.218). Without temperature scaling it already reaches a lower ECE (0.48) than Focal Loss or MDCA after temperature scaling (1.13 and 0.60), and temperature scaling on top of D-GCM-Critic gives the best entry overall (0.37), so the two are complementary.

The non-adversarial D-GCM-Basis is itself a competitive regularizer that often matches or exceeds prior fixed-witness objectives, and the consistent gap to D-GCM-Critic supports our central hypothesis: a learned witness that searches the simplex for miscalibration patterns uses its capacity more effectively than a pre-specified witness family.

6.3 Ablation Study: The Necessity of the Debiased Estimator

Table 3 tests the central claim of Proposition 1 by changing only the estimator: V-GCM rows use the naive V-statistic (4) and D-GCM rows the U-statistic (7). The U-statistic yields the largest gains when the witness is learned: ECE/SCE improve from $0.66/3.38$ to $0.48/3.27$ on CIFAR-10 and from $1.46/1.32$ to $1.08/1.04$ on Tiny-ImageNet, whereas for the *fixed basis* the same change gives smaller improvements (Tiny-ImageNet ECE $1.83 \to 1.70$). This asymmetry matches the theory.

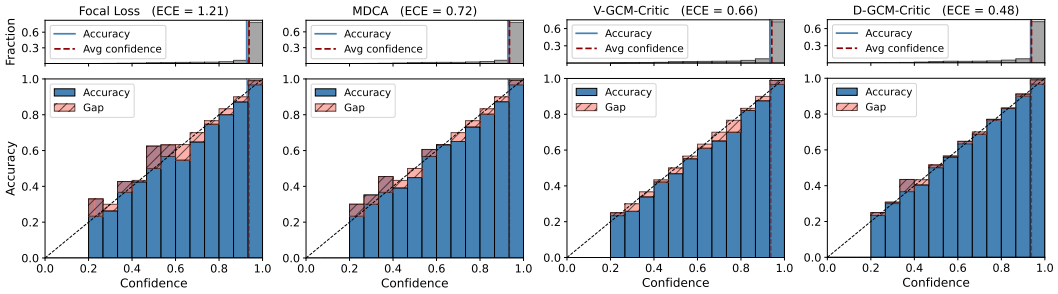


Figure 3: Reliability diagrams on CIFAR-10 (ResNet-56), computed from the test predictions of one seed of Tables 2 and 3. Blue bars are per-bin accuracy, red bars the gap to perfect calibration, and the top histograms show the confidence distribution with average accuracy (blue solid line) and confidence (red dashed line).

For a fixed witness the diagonal adds a predictor-dependent bias, while for a *learned critic* it also rewards saturating the witness norm on individual high-residual samples, which turns the critic into a hard-example miner and leaves the cross-sample calibration structure undiscovered. Removing the self-interactions restores the intended target and produces the calibration gains.

Figure 3 shows the same effect on reliability diagrams. V-GCM-Critic retains visible per-bin gaps between confidence and accuracy because its critic never enforces cross-sample alignment, whereas D-GCM-Critic aligns the two in nearly every bin. Debiasing is therefore a requirement for adversarial witness learning, and it is what makes the learned critic target calibration structure.

6.4 Plug-and-Play Compatibility and Overhead

Because D-GCM operates solely on the predicted probability simplex, it is agnostic to the base classification objective. Table 4 shows that adding D-GCM lowers ECE and SCE in all 24 dataset–loss–metric cells across Cross-Entropy, Label Smoothing, and Focal Loss. For example, on CIFAR-100, adding D-GCM to Focal Loss reduces ECE from 1.97 to 0.69 and SCE from 1.78 to 1.63.

Table 5: Training overhead on CIFAR-100 with ResNet-56. Relative time and peak memory are measured on the same hardware and with the same batch size of 128.

Method	Extra Params	Time/Epoch	Peak VRAM
Focal Loss	0 K	23.4 s	3.154 GB
MDCA	0 K	23.6 s	3.154 GB
MMCE	0 K	24.0 s	3.154 GB
D-GCM-Basis	0 K	24.1 s	3.154 GB
D-GCM-Critic	12.7 K	24.3 s	3.154 GB

This adaptive capacity comes at negligible overhead. Table 5 shows that the critic (d=32) adds only 12.7K parameters ($\sim$ 1M in ResNet-56) and 0.3–0.9 s per epoch over MMCE, MDCA, and Focal Loss, with no increase in peak VRAM. Despite the min–max formulation, training is stable: after the first learning-rate drop on CIFAR-10, the validation ECE of D-GCM-Critic follows the lowest curve, stays flat through the second drop, and shows none of the late rebound toward over-confidence seen for Focal Loss and, mildly, MDCA (Appendix C.7). The critic hyperparameters are easy to set: d=32 captures nearly all the gain over d=16 (d=64 adds about 0.005 ECE points), and validation ECE varies by only about 0.03 points over critic learning rates from 10^{-4} to 10^{-2} (Appendix C.8).

7 Conclusion

We formulate train-time multiclass calibration as a generalized conditional-moment problem, shifting the focus from predefined statistics to adaptive witness learning on the probability simplex. Our analysis identifies a critical failure mode of the naive V-statistic: its diagonal term gives the learned critic a finite-batch shortcut, rewarding sample-wise residual energy rather than cross-sample miscalibration. D-GCM removes this shortcut with an efficient U-statistic over distinct sample pairs. Experiments validate the necessity of this debiasing and show that D-GCM serves as an efficient plug-and-play regularizer, achieving state-of-the-art multiclass calibration while maintaining competitive prediction accuracy. Appendix A discusses the limitations of D-GCM.

AI USE STATEMENT

A generative AI was used for language editing, LaTeX formatting, checking the consistency of tables, and organizing the codebase uploaded for review. All research ideas, theoretical results, experiments, and analyses were developed and carried out by the authors, who take full responsibility for the content.

REPRODUCIBILITY STATEMENT

Appendix C lists all datasets, splits, evaluation metrics with their exact definitions, baselines, and hyperparameters (Tables 6–11), Figure 2 gives a complete implementation of the D-GCM objective, and all proofs are in Appendix B. Our implementation of D-GCM-Critic and its V-statistic variant is included in the supplementary material, and the full codebase will be released upon acceptance.

REFERENCES

Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Meta-calibration: Learning of model calibration using differentiable expected calibration error. *arXiv preprint arXiv:2106.09613*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16060–16069. IEEE, 2022.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs. Soft calibration objectives for neural networks. *Advances in neural information processing systems*, 34:29768–29779, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

Gongbo Liang, Yu Zhang, Xiaoqin Wang, and Nathan Jacobs. Improved trainable calibration method for neural networks on medical imaging classification. *arXiv preprint arXiv:2009.04057*, 2020.

Jinxu Lin, Linwei Tao, Minjing Dong, and Chang Xu. Uncertainty weighted gradients for model calibration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15497–15507. IEEE, 2025.

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

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. 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*, pp. 80–88, 2022.

Bingyuan Liu, Jérôme Rony, Adrian Galdran, Jose Dolz, and Ismail Ben Ayed. Class adaptive network calibration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16070–16079, 2023.

Charlie Marx, Sofian Zalouk, and Stefano Ermon. Calibration by distribution matching: Trainable kernel calibration metrics. *Advances in Neural Information Processing Systems*, 36:25910–25928, 2023.

Tom Mitchell. Twenty Newsgroups. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C5C323.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, 33:15288–15299, 2020.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning. *arXiv preprint arXiv:1904.01685*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Nicolas Posocco and Antoine Bonnefoy. Estimating expected calibration errors. In *International conference on artificial neural networks*, pp. 139–150. Springer, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Cheng Wang and Jacek Golebiowski. Towards unbiased calibration using meta-regularization. *arXiv preprint arXiv:2303.15057*, 2023.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. *Advances in neural information processing systems*, 32, 2019.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.

Hee Suk Yoon, Joshua Tian Jin Tee, Eunseop Yoon, Sunjae Yoon, Gwangsu Kim, Yingzhen Li, and Chang D Yoo. Esd: Expected squared difference as a tuning-free trainable calibration measure. *arXiv preprint arXiv:2303.02472*, 2023.

Yilin Zhang, Cai Xu, You Wu, Ziyu Guan, and Wei Zhao. Target-agnostic calibration under distribution shift with frequency-aware gradient rectification. *arXiv preprint arXiv:2508.19830*, 2025.

Appendix

A Limitations

Finite critic capacity. A restricted neural witness controls only the miscalibration components it can represent (Proposition 2). Low ECE and SCE are therefore evidence of calibration, and exact certification requires the rich witness families of Theorem 1. Propositions 1 and 3 hold for a fixed sample-wise witness, so optimizing the critic on the same mini-batch adds the usual gap between the empirical and the population inner maximization. The training-stability comparison in Appendix C.7 is an empirical observation and not a convergence guarantee for the min–max problem.

Weak cross-sample signal. Small batches, rare classes, and sparsely populated simplex regions provide fewer informative distinct pairs and can increase the variance of the regularizer gradient. Individual mini-batch values of the U-statistic may be negative, although they are bounded below by $-2/(B-1)$ (Appendix B.2).

Sample-wise assumption. Exact unbiasedness requires both the predictor and the witness to be sample-wise functions of each example. BatchNorm (Ioffe & Szegedy, 2015) in training mode violates this condition unless its statistics are frozen when the regularizer is evaluated, and sample-wise normalization such as LayerNorm or GroupNorm avoids it (Appendix B.9).

Accuracy and scale. D-GCM attains the best accuracy among the compared calibration regularizers on the three vision datasets, but it still costs 1.06–2.99 accuracy points relative to the Focal Loss reference, and a large λ_{DGCM} can induce under-confidence. Evaluation with very large class counts K (the overhead scales as $\mathcal{O}(BKd)$), on tabular data, and in foundation-model fine-tuning is outside the current scope and remains future work.

B Additional Proofs and Discussion

B.1 Proof of Proposition 1

Proof. Let

$$Z := \mathrm{vec}\big(R\,\phi(P_\theta)^\top\big),$$

where $R = r_\theta(X, Y)$. From the derivation in the proof of Proposition 3,

$$\mathbb{E}\Big[\widehat{\mathcal{C}}_V(\theta,\phi)\Big] = \frac{B-1}{B}\,\|\mathbb{E}[Z]\|_2^2 + \frac{1}{B}\,\mathbb{E}\big[\|Z\|_2^2\big]\,.$$

Since

$$\|\mathbb{E}[Z]\|_2^2 = \|M_\theta(\phi)\|_F^2 \qquad \text{and} \qquad \|Z\|_2^2 = \|R\|_2^2\,\|\phi(P_\theta)\|_2^2,$$

we obtain

$$\mathbb{E}\Big[\widehat{\mathcal{C}}_V(\theta,\phi)\Big] = \left(1-\frac{1}{B}\right)\|M_\theta(\phi)\|_F^2 + \frac{1}{B}\,\mathbb{E}\big[\|R\|_2^2\|\phi(P_\theta)\|_2^2\big]\,.$$

Now assume that p_θ is strongly calibrated. Then $M_\theta(\phi) = 0$ for every fixed sample-wise witness ϕ, because

$$M_\theta(\phi) = \mathbb{E}\big[\mathbb{E}[R \mid P_\theta]\phi(P_\theta)^\top\big] = 0.$$

Therefore

$$\mathbb{E}\Big[\widehat{\mathcal{C}}_V(\theta,\phi)\Big] = \frac{1}{B}\,\mathbb{E}\big[\|R\|_2^2\|\phi(P_\theta)\|_2^2\big] \le \frac{1}{B}\,\mathbb{E}\big[\|R\|_2^2\big]\,,$$

where the inequality uses $\sup_q \|\phi(q)\|_2 \le 1$. Equality is attained by any constant witness $\phi(q) \equiv u$ with $\|u\|_2 = 1$. Therefore

$$\sup_{\phi:\,\sup_q \|\phi(q)\|_2 \le 1} \mathbb{E}\Big[\widehat{\mathcal{C}}_V(\theta,\phi)\Big] = \frac{1}{B}\,\mathbb{E}\big[\|R\|_2^2\big]\,.$$

This quantity is strictly positive whenever $R \not\equiv 0$ almost surely.

For the U-statistic, Proposition 3 yields

$$\mathbb{E}\Big[\widehat{\mathcal{C}}_U(\theta,\phi)\Big] = \|M_\theta(\phi)\|_F^2 = 0$$

for every fixed sample-wise witness ϕ. Taking the supremum over the same witness class gives

$$\sup_{\phi:\,\sup_q \|\phi(q)\|_2 \le 1} \mathbb{E}\Big[\widehat{\mathcal{C}}_U(\theta,\phi)\Big] = 0.$$

□

B.2 Lower bound for the mini-batch U-statistic

Proof. Let $z_i = \mathrm{vec}(r_i\phi_i^\top)$. Since $\|r_i\|_2 \le \sqrt{2}$ and $\|\phi_i\|_2 \le 1$, we have $\|z_i\|_2^2 \le 2$. Moreover,

$$\widehat{\mathcal{C}}_U(\theta,\phi) = \frac{\left\|\sum_{i=1}^B z_i\right\|_2^2 - \sum_{i=1}^B \|z_i\|_2^2}{B(B-1)} \ge -\frac{\sum_{i=1}^B \|z_i\|_2^2}{B(B-1)} \ge -\frac{2B}{B(B-1)} = -\frac{2}{B-1}.$$

□

B.3 Connections to MDCA, ECE/Soft Calibration, and MMCE

The comparison separates four ingredients: the residual summary, the witness family, the outer aggregation that maps moments to a scalar, and the finite-sample estimator. For this subsection only, we therefore allow a residual summary

$$\rho_\theta(X,Y) \in \mathbb{R}^s$$

that is any measurable function of (P_θ, Y), and define the induced moment matrix

$$M_\theta^\rho(\phi) := \mathbb{E}\big[\rho_\theta(X,Y)\,\phi(P_\theta)^\top\big] \in \mathbb{R}^{s\times d}.$$

The main objective in Section 3 corresponds to the full multiclass residual

$$\rho_\theta^{\mathrm{full}}(X,Y) = r_\theta(X,Y) = e_Y - P_\theta \in \mathbb{R}^K.$$

Many classical calibration objectives instead use a scalar projected residual. For top-label confidence calibration, define

$$C(P_\theta) := \max_{k\in[K]} P_{\theta,k}, \qquad \hat{k}(P_\theta) := \arg\max_{k\in[K]} P_{\theta,k},$$

and

$$\delta_\theta(X,Y) := e_{\hat{k}(P_\theta)}^\top r_\theta(X,Y) = \mathbf{1}\{Y = \hat{k}(P_\theta)\} - C(P_\theta) \in \mathbb{R}.$$

Although ECE-, soft-binning-, and MMCE-style objectives are usually written on the one-dimensional confidence variable $C(P_\theta)$, they still fit the same simplex-based template because $C(\cdot)$ and $\hat{k}(\cdot)$ are deterministic functions of the predicted probability vector P_θ. The key restrictions are that these methods project the full residual r_θ to the scalar top-label residual δ_θ and restrict witnesses to depend on P_θ only through $C(P_\theta)$ (or similarly coarse classwise summaries).

MDCA. Take the full residual $\rho_\theta^{\mathrm{full}}$ and the constant scalar witness $h(q) \equiv 1$. The resulting empirical moment is

$$\widehat{m}_B^{\mathrm{full}}(1) = \frac{1}{B}\sum_{i=1}^B (e_{y_i} - p_i) \in \mathbb{R}^K.$$

The MDCA loss of Hebbalaguppe et al. (2022) is exactly

$$\mathcal{L}_{\mathrm{MDCA}} = \frac{1}{K}\sum_{k=1}^K \left|\frac{1}{B}\sum_{i=1}^B (e_{y_i,k} - p_{i,k})\right| = \frac{1}{K}\left\|\widehat{m}_B^{\mathrm{full}}(1)\right\|_1.$$

Therefore MDCA is the constant-witness, classwise mean-matching instance of the generalized moment viewpoint. It uses the full multiclass residual, but only a single witness $h \equiv 1$, so it controls global classwise averages rather than arbitrary functions of the probability vector.

Hard-binned ECE and soft-binned ECE. Let $C_i := \max_k p_{i,k}$, $\hat{k}_i := \arg\max_k p_{i,k}$, and

$$\delta_i := \mathbf{1}\{y_i = \hat{k}_i\} - C_i.$$

For hard confidence bins $I_1, \dots, I_M \subset [0,1]$, define

$$h_b(q) := \mathbf{1}\{C(q) \in I_b\}.$$

Then the empirical ECE can be written as

$$\widehat{\mathrm{ECE}} = \sum_{b=1}^{M} \left| \frac{1}{B} \sum_{i=1}^{B} \delta_i \, h_b(p_i) \right|.$$

Indeed, if $B_b := \{i : C_i \in I_b\}$, then, with the convention that empty bins contribute zero,

$$\left| \frac{1}{B} \sum_{i=1}^{B} \delta_i \, h_b(p_i) \right| = \frac{|B_b|}{B} \left| \frac{1}{|B_b|} \sum_{i \in B_b} \big(\mathbf{1}\{y_i = \hat{k}_i\} - C_i \big) \right|,$$

which is exactly the b-th empirical ECE contribution.

Soft Calibration (Karandikar et al., 2021) replaces the hard bin indicator by a smooth bin-membership function $u_b^{M,T}(C(q))$. For the SB-ECE objective with $p = 1$,

$$\text{SB-ECE}_{\text{bin},1}(M,T) = \sum_{b=1}^{M} \left| \frac{1}{B} \sum_{i=1}^{B} u_b^{M,T}(C_i) \, \delta_i \right|.$$

This follows from the identities

$$S_b = \sum_{i=1}^{B} u_b^{M,T}(C_i), \qquad A_b = \frac{1}{S_b} \sum_{i=1}^{B} u_b^{M,T}(C_i) \mathbf{1}\{y_i = \hat{k}_i\}, \qquad \bar{C}_b = \frac{1}{S_b} \sum_{i=1}^{B} u_b^{M,T}(C_i) C_i,$$

assuming $S_b > 0$. When $S_b = 0$, the corresponding bin contribution is defined as zero. Thus, we have

$$\frac{S_b}{B} \, |A_b - \bar{C}_b| = \left| \frac{1}{B} \sum_{i=1}^{B} u_b^{M,T}(C_i) \big(\mathbf{1}\{y_i = \hat{k}_i\} - C_i \big) \right|.$$

Thus hard ECE and SB-ECE are fixed-witness confidence-calibration objectives on the scalar top-label residual δ_θ. Similar classwise variants follow from class-indexed coordinate summaries of q. The S-AvUC objective of Karandikar et al. (2021) uses a different accuracy-versus-uncertainty construction and lies outside this template.

MMCE. Let $k : [0,1] \times [0,1] \to \mathbb{R}$ be a kernel with RKHS $\mathcal{H}$ and feature map φ. Using the same scalar top-label residual δ_θ, the population MMCE of Kumar et al. (2018) can be written as

$$\mathrm{MMCE} = \| \mathbb{E}[\delta_\theta(X,Y) \, \varphi(C(P_\theta))] \|_{\mathcal{H}}.$$

Its empirical squared form is

$$\mathrm{MMCE}_B^2 = \frac{1}{B^2} \sum_{i,j=1}^{B} \delta_i \delta_j \, k(C_i, C_j).$$

Therefore MMCE is the fixed-RKHS witness instance of the same moment principle, specialized to top-label confidence calibration. The weighted training variant changes pair weights but not this basic fixed-kernel interpretation.

Within our framework, the corresponding debiased estimator is obtained by removing the diagonal:

$$\widehat{\mathcal{C}}_U^{\mathrm{MMCE}} = \frac{1}{B(B-1)} \sum_{i \neq j} \delta_i \delta_j \, k(C_i, C_j),$$

which is the off-diagonal U-statistic analogue of the squared MMCE moment. This changes only the finite-sample estimator, not the witness class itself.

KCE / SKCE. Let $k : \Delta^{K-1} \times \Delta^{K-1} \to \mathbb{R}^{K\times K}$ be a matrix-valued kernel with associated RKHS $\mathcal{H}_k$ of vector-valued witnesses $f : \Delta^{K-1} \to \mathbb{R}^K$. The kernel calibration error of Widmann et al. (2019) can be written as

$$\mathrm{KCE}[k, g] = \sup_{\|f\|_{\mathcal{H}_k}\le 1} \mathbb{E}\big[(e_Y - g(X))^\top f(g(X))\big] ,$$

which is exactly a fixed full-multiclass witness instance of our generalized moment viewpoint. Using the kernel trick, its squared form becomes

$$\mathrm{KCE}[k, g]^2 = \mathbb{E}\big[(e_Y - g(X))^\top k(g(X), g(X'))(e_{Y'} - g(X'))\big] .$$

The corresponding finite-sample estimators of squared KCE (SKCE) include both the biased all-pairs estimator and unbiased estimators over distinct pairs. Therefore KCE/SKCE matches our framework at the level of the residual, witness class, and pairwise estimation template, but differs from D-GCM-Critic in that the witness is fixed by a pre-specified RKHS rather than learned jointly with the classifier.

Remark on terminology. The identities above are exact at the level of the moment-building recipe after specifying the residual summary, witness family, outer aggregation, and finite-sample estimator. For example, MDCA and ECE/SB-ECE use absolute L_1-type aggregation over classes or bins, MMCE uses the RKHS norm induced by its kernel, and D-GCM uses the squared Frobenius norm together with the debiased U-statistic. We therefore describe MDCA, hard/soft-binned ECE, and MMCE as restricted instances of the generalized moment template underlying D-GCM, and the differences are substantive: (1) MDCA uses only the constant witness, (2) ECE/soft-binned ECE/MMCE in their standard forms project to top-label confidence calibration and restrict witnesses to depend on P_θ only through $C(P_\theta)$, and (3) D-GCM further differs by using the debiased U-statistic (7) and, in the critic version, a learned witness family. The term *debiased* in D-GCM refers to removing the diagonal self-interaction of the V-statistic in (5). This differs from the *unbiased* calibration of Wang & Golebiowski (2023), where the term refers to a kernel-smoothed ECE surrogate that is free of binning bias, and where the learned component outputs a sample-wise focal exponent while the calibration statistic itself stays fixed.

B.4 Fixed-basis specialization

A non-adversarial variant, ***D-GCM-Basis***, is obtained by choosing a fixed basis $b = (b_1, \dots, b_m) : \Delta^{K-1} \to \mathbb{R}^m$ and replacing ϕ_ψ by b in (8). This removes the inner maximization and directly penalizes the chosen calibration moments

$$M_\theta(b) = \mathbb{E}\big[r_\theta(X, Y)b(P_\theta)^\top\big] .$$

Useful choices include low-degree polynomials, Bernstein-style bases on the simplex, and smooth soft-bin features, where the latter recover binned or soft-binned moment-matching objectives as restricted-family special cases. In all cases we normalize or clip the basis so that $\sup_q \|b(q)\|_2 \le 1$, matching the constraint used in the main objective. This variant serves as an ablation and a non-adversarial baseline, since the main contribution of D-GCM is adaptive witness learning.

B.5 Proof of Theorem 1

Proof. Let $P = P_\theta$ and $R = r_\theta(X, Y)$. Since Y is finite and $P \in \Delta^{K-1}$, each coordinate of R is bounded and therefore belongs to L^2.

Assume first that p_θ is strongly calibrated. Then $\mathbb{E}[R \mid P] = 0$ almost surely, and for any $h \in \mathcal{H}$,

$$m_\theta(h) = \mathbb{E}[R\, h(P)] = \mathbb{E}[\mathbb{E}[R \mid P]h(P)] = 0.$$

This proves the forward implication.

For the converse, define

$$g(P) := \mathbb{E}[R \mid P] \in L^2(\mu_\theta; \mathbb{R}^K).$$

For each class coordinate k and each $h \in \mathcal{H}$,

$$\mathbb{E}[g_k(P)h(P)] = \mathbb{E}[R_k h(P)] = 0.$$

Therefore g_k is orthogonal to every function in $\mathcal{H}$, and therefore to the linear span of $\mathcal{H}$. By density, g_k is orthogonal to all of $L^2(\mu_\theta)$. Taking $h = g_k$ gives

$$\mathbb{E}[g_k(P)^2] = 0,$$

so $g_k(P) = 0$ μ_θ-almost surely. Since this holds for every k, we obtain

$$g(P) = \mathbb{E}[R \mid P] = 0 \quad \text{a.s.}$$

which is exactly (1). This proves the converse. □

B.6 Proof of Proposition 2

Proof. Let $P = P_\theta$ and $g_k(P) = \mathbb{E}[r_{\theta,k}(X,Y) \mid P]$. For any $h \in \mathcal{H}_m$,

$$\langle g_k, h \rangle_{L^2(\mu_\theta)} = \mathbb{E}[g_k(P)h(P)] = \mathbb{E}[r_{\theta,k}(X,Y)h(P)] = 0.$$

Thus g_k is orthogonal to every witness in $\mathcal{H}_m$. Therefore, for any $h \in \mathcal{H}_m$,

$$\begin{aligned}\|g_k - h\|^2_{L^2(\mu_\theta)} &= \|g_k\|^2_{L^2(\mu_\theta)} + \|h\|^2_{L^2(\mu_\theta)} - 2\langle g_k, h\rangle_{L^2(\mu_\theta)} \\ &= \|g_k\|^2_{L^2(\mu_\theta)} + \|h\|^2_{L^2(\mu_\theta)} \geq \|g_k\|^2_{L^2(\mu_\theta)},\end{aligned}$$

which proves (11).

Now suppose that for each k there exists $h_k \in \mathcal{H}_m$ with

$$\|g_k - h_k\|_{L^2(\mu_\theta)} \leq \varepsilon_k.$$

Applying (11) with $h = h_k$ gives

$$\|g_k\|_{L^2(\mu_\theta)} \leq \varepsilon_k.$$

Summing over coordinates yields

$$\|g_\theta\|^2_{L^2(\mu_\theta;\mathbb{R}^K)} = \sum_{k=1}^K \|g_k\|^2_{L^2(\mu_\theta)} \leq \sum_{k=1}^K \varepsilon_k^2,$$

and taking square roots concludes the proof. □

B.7 Proof of Proposition 3

Proof. Let

$$Z_i := \mathrm{vec}(r_i \phi_i^\top) \in \mathbb{R}^{Kd},$$

where $r_i = e_{y_i} - p_\theta(x_i)$ and $\phi_i = \phi(p_\theta(x_i))$. Under the sample-wise assumption, the random vectors $(Z_i)_{i=1}^B$ are i.i.d. Then

$$\widehat{\mathcal{C}}_U(\theta,\phi) = \frac{1}{B(B-1)} \sum_{i \neq j} Z_i^\top Z_j.$$

Taking expectations and using independence for $i \neq j$,

$$\mathbb{E}\Big[\widehat{\mathcal{C}}_U(\theta,\phi)\Big] = \mathbb{E}[Z_1^\top Z_2] = \mathbb{E}[Z_1]^\top \mathbb{E}[Z_2] = \|\mathbb{E}[Z]\|_2^2.$$

Finally,

$$\|\mathbb{E}[Z]\|_2^2 = \big\|\mathrm{vec}\big(\mathbb{E}[r_\theta(X,Y)\phi(P_\theta)^\top]\big)\big\|_2^2 = \|M_\theta(\phi)\|_F^2,$$

which proves (12).

For the naive V-statistic,

$$\widehat{\mathcal{C}}_V(\theta,\phi) = \left\| \frac{1}{B} \sum_{i=1}^B Z_i \right\|_2^2 = \frac{1}{B^2} \sum_{i=1}^B \|Z_i\|_2^2 + \frac{1}{B^2} \sum_{i \neq j} Z_i^\top Z_j.$$

Therefore

$$\mathbb{E}\Big[\widehat{\mathcal{C}}_V(\theta,\phi)\Big] = \frac{1}{B}\mathbb{E}\big[\|Z\|_2^2\big] + \frac{B-1}{B}\|\mathbb{E}[Z]\|_2^2 = \|\mathbb{E}[Z]\|_2^2 + \frac{1}{B}\,\mathrm{tr}\big(\mathrm{Var}(Z)\big),$$

which is (6). □

B.8 Derivation of the efficient form (8)

Proof. Let $A_i = r_i \phi_i^\top \in \mathbb{R}^{K \times d}$. Then

$$\left\| \sum_{i=1}^{B} A_i \right\|_F^2 = \sum_{i=1}^{B} \sum_{j=1}^{B} \langle A_i, A_j \rangle_F = \sum_{i=1}^{B} \sum_{j=1}^{B} (r_i^\top r_j)(\phi_i^\top \phi_j).$$

The diagonal contribution is

$$\sum_{i=1}^{B} \|A_i\|_F^2 = \sum_{i=1}^{B} \|r_i\|_2^2 \|\phi_i\|_2^2.$$

Subtracting the diagonal terms leaves exactly

$$\sum_{i \neq j} (r_i^\top r_j)(\phi_i^\top \phi_j).$$

Dividing by $B(B-1)$ yields (8). □

B.9 Practical note on normalization layers

Proposition 3 is exact when the predictor and witness are sample-wise functions. If the backbone uses batch normalization in training mode, each $p_\theta(x_i)$ depends weakly on the full mini-batch, so the estimator is no longer exactly unbiased. In practice, this can be handled by using architectures with sample-wise normalizations (e.g., LayerNorm (Xu et al., 2019) or GroupNorm (Wu & He, 2018)), or by evaluating the regularizer with frozen normalization statistics.

C Experiment Details

C.1 Reproducibility

We include our implementation of D-GCM-Critic and its V-statistic variant in the supplementary material to support reproducibility during review. Detailed hyperparameter settings, architectural choices, training protocols, and other implementation details are provided in Appendix C.4. If the paper is accepted, we will release the complete codebase for all experiments and datasets.

C.2 Datasets and Metrics

Datasets. We follow the experimental protocol of MDCA (Hebbalaguppe et al., 2022) so that differences in performance can be attributed to the calibration objective rather than to the data split, backbone, or optimization recipe. We evaluate on four classification benchmarks: CIFAR-10, CIFAR-100, Tiny-ImageNet, and 20 Newsgroups. For datasets with a predefined training split, we reserve a validation subset for model selection and keep the test set untouched. The validation set is also used by baselines that require a held-out calibration signal. The dataset statistics and splits are summarized in Table 6.

Table 6: Datasets used in our experiments. For CIFAR-10/100 we use the standard train/test split and reserve 10% of the original training set as validation, following MDCA. For Tiny-ImageNet, following the MDCA/FLSD protocol, we reserve 50 images per class from the training set for validation and use the official validation set as the test set.

Dataset	Classes	Input	Train	Val.	Test / evaluation split
CIFAR-10 (Krizhevsky et al., 2009)	10	32×32 RGB	45,000	5,000	10,000
CIFAR-100 (Krizhevsky et al., 2009)	100	32×32 RGB	45,000	5,000	10,000
Tiny-ImageNet (Deng et al., 2009)	200	64×64 RGB	90,000	10,000	10,000
20 Newsgroups (Mitchell, 1997)	20	Text	90% training	10% training	Official test split

Evaluation metrics. Let $\mathcal{D}_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{n}$ be the test set with K classes. Let $p_i = p_\theta(x_i) \in \Delta^{K-1}$ denote the predicted probability vector, $\hat{y}_i = \arg\max_k p_{i,k}$ the predicted class, and $c_i = \max_k p_{i,k}$ the top-class confidence. We report top-1 accuracy, Expected Calibration Error (ECE), and Static Calibration Error (SCE). Following MDCA, ECE and SCE are computed with $M = 15$ equally spaced bins. Accuracy and ECE are multiplied by 100 and reported in percentage points, and SCE is multiplied by 1000 and reported in units of 10^{-3}, because its average over the K classes is an order of magnitude smaller than ECE.

Top-1 accuracy is

$$\text{Acc} = \frac{100}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{y}_i = y_i\}. \tag{13}$$

ECE measures top-label confidence calibration (Guo et al., 2017). Let $I_m = \{i : c_i \in (\frac{m-1}{M}, \frac{m}{M}]\}$ be the set of examples whose top confidence falls in bin m. For a non-empty bin I_m, define

$$\text{acc}(I_m) = \frac{1}{|I_m|} \sum_{i \in I_m} \mathbf{1}\{\hat{y}_i = y_i\}, \qquad \text{conf}(I_m) = \frac{1}{|I_m|} \sum_{i \in I_m} c_i. \tag{14}$$

Then

$$\text{ECE} = 100 \sum_{m=1}^{M} \frac{|I_m|}{n} \left|\text{acc}(I_m) - \text{conf}(I_m)\right|, \tag{15}$$

where empty bins contribute zero.

SCE measures calibration over all class probabilities rather than only the top predicted class (Nixon et al., 2019). For class k and bin m, let $I_{m,k} = \{i : p_{i,k} \in (\frac{m-1}{M}, \frac{m}{M}]\}$. For a non-empty $I_{m,k}$, define

$$\text{acc}(I_{m,k}; k) = \frac{1}{|I_{m,k}|} \sum_{i \in I_{m,k}} \mathbf{1}\{y_i = k\}, \qquad \text{conf}(I_{m,k}; k) = \frac{1}{|I_{m,k}|} \sum_{i \in I_{m,k}} p_{i,k}. \tag{16}$$

The SCE is

$$\text{SCE} = \frac{1000}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} \frac{|I_{m,k}|}{n} \left|\text{acc}(I_{m,k}; k) - \text{conf}(I_{m,k}; k)\right|. \tag{17}$$

Compared with ECE, SCE is more sensitive to multiclass calibration because it evaluates every coordinate of the probability vector.

C.3 Compute Resources

All experiments were conducted on an Ubuntu 24.04 server equipped with six NVIDIA A100 GPUs (80 GB each) and 512 GB of system RAM.

Unless otherwise stated, the runtime measurements below correspond to single-GPU training on one A100 under our default implementation settings. For D-GCM-Critic, the typical training time is approximately 24.1 seconds per epoch on CIFAR-10, 24.3 seconds on CIFAR-100, and 185.7 seconds on Tiny-ImageNet. These numbers are rough measurements of computational cost.

C.4 Baseline and Hyperparameters

Model selection and repeated runs. For all train-time calibration methods, we follow the MDCA model-selection protocol: for each method and each hyperparameter setting, we train on the training split, select the checkpoint using the validation set and evaluate the selected model on the test set. All main results are averaged over three random seeds. Unless otherwise stated, no post-hoc temperature scaling or validation-time recalibration is applied to the reported train-time calibration results.

Baselines. Table 7 summarizes the baseline losses and hyperparameters. For the main comparison, our D-GCM variants are added to Focal Loss, since MDCA identifies Focal Loss as a strong base loss for train-time calibration. In the plug-and-play experiments, the same D-GCM regularizer is added to Cross-Entropy, Label Smoothing, and Focal Loss.

Table 7: Baseline losses and hyperparameters. We follow the baselines and their hyperparameters in the MDCA appendix.

Method	Training objective / description	Hyperparameters
Cross-Entropy / NLL	Standard negative log-likelihood, $\mathcal{L}_{\mathrm{CE}} = -\frac{1}{B}\sum_{i=1}^{B} \log p_{i,y_i}$.	None.
Label Smoothing (Szegedy et al., 2016)	$q_{i,y_i} = 1-\alpha$ and $q_{i,k} = \alpha/(K-1)$ for $k \neq y_i$, $\mathcal{L}_{\mathrm{LS}} = -\frac{1}{B}\sum_i \sum_k q_{i,k} \log p_{i,k}$.	$\alpha = 0.1$.
Focal Loss (Lin et al., 2017; Mukhoti et al., 2020)	$\mathcal{L}_{\mathrm{FL}} = -\frac{1}{B}\sum_{i=1}^{B}(1-p_{i,y_i})^{\gamma} \log p_{i,y_i}$.	$\gamma \in \{1,2,3\}$. $\gamma = 3$ is the default fixed-γ setting.
DCA (Liang et al., 2020)	Trains with $\mathcal{L}_{\mathrm{CE}} + \beta\,\mathcal{L}_{\mathrm{DCA}}$.	$\beta \in \{1,5,10,15,20,25\}$.
MMCE (Kumar et al., 2018)	Uses the weighted MMCE calibration regularizer with NLL.	$\lambda_{\mathrm{MMCE}} \in \{2,4\}$.
FLSD (Mukhoti et al., 2020)	Focal-loss-based calibration baseline under the MDCA protocol.	$\gamma = 3$.
MDCA (Hebbalaguppe et al., 2022)	$\mathcal{L}_{\mathrm{MDCA}} = \frac{1}{K}\sum_{k=1}^{K}\left\lvert \frac{1}{B}\sum_{i=1}^{B} p_{i,k} - \frac{1}{B}\sum_{i=1}^{B}\mathbf{1}\{y_i = k\}\right\rvert$. The full loss is $\mathcal{L}_{\mathrm{base}} + \beta\mathcal{L}_{\mathrm{MDCA}}$.	$\beta \in \{0.25, 0.5, \ldots, 50\}$. MDCA reports best calibration near $\beta = 1$.
SoftCal(Karandikar et al., 2021), MbLS(Liu et al., 2022), ESD(Yoon et al., 2023), BSCE-GRA(Lin et al., 2025), CALS(Liu et al., 2023), DECE(Bohdal et al., 2021), FGR(Zhang et al., 2025)	Evaluated under the same dataset splits, backbone, training budgets, and validation protocol. FGR is run on the three image benchmarks only (Appendix C.5).	We follow the settings in the official implementation.
D-GCM-Basis	Our fixed-witness variant. The training loss is $\mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{DGCM}}\widehat{\mathcal{C}}_U(\theta, \phi_0)$, where ϕ_0 is a bounded fixed basis and $\widehat{\mathcal{C}}_U$ is the debiased U-statistic.	$\lambda_{\mathrm{DGCM}} \in [10^{-2}, 10]$ with default 1; witness dimension $d \in \{16, 32, 64\}$ with default 32.
D-GCM-Critic	Our learned-witness variant. We alternate one critic ascent step and one model descent step using $\min_\theta \max_\psi \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{DGCM}}\widehat{\mathcal{C}}_U(\theta, \phi_\psi)$.	$\lambda_{\mathrm{DGCM}} \in [10^{-2}, 10]$ with default 1; critic learning rate $\eta_\psi \in [10^{-5}, 10^{-2}]$ with default 10^{-3}; critic dimension $d \in \{16, 32, 64\}$ with default 32.

Dataset-specific training hyperparameters. Tables 8–11 give the dataset-specific optimization settings. Apart from the D-GCM-specific hyperparameters, all settings follow the corresponding MDCA protocol.

Table 8: Hyperparameters for CIFAR-10.

Component	Value
Backbone	ResNet-56
Train / validation / test split	45,000 / 5,000 / 10,000
Base loss in main benchmark	Focal Loss, $\gamma \in \{1,2,3\}$, default $\gamma = 3$

Table 8 (continued)

Component	Value
Optimizer	SGD
Initial model learning rate	0.1
Learning-rate schedule	Multiply by 0.1 at epochs 80 and 120
Epochs	160
Batch size	128
Momentum	0.9
Weight decay	5×10^{-4}
Data augmentation	Random crop and horizontal flip
Calibration bins	$M = 15$
D-GCM regularizer weight	$\lambda_{\mathrm{DGCM}} \in [10^{-2}, 10]$, default 1
D-GCM critic learning rate	$\eta_\psi \in [10^{-5}, 10^{-2}]$, default 10^{-3}
D-GCM critic / witness dimension	$d \in \{16, 32, 64\}$, default 32
Critic-to-model update ratio	[1,. . . ,8], default 1

Table 9: Hyperparameters for CIFAR-100.

Component	Value
Backbone	ResNet-56
Train / validation / test split	45,000 / 5,000 / 10,000
Base loss in main benchmark	Focal Loss, $\gamma \in \{1, 2, 3\}$, default $\gamma = 3$
Optimizer	SGD
Initial model learning rate	0.1
Learning-rate schedule	Multiply by 0.1 at epochs 100 and 150
Epochs	200
Batch size	128
Momentum	0.9
Weight decay	5×10^{-4}
Data augmentation	Random crop and horizontal flip
Calibration bins	$M = 15$
D-GCM regularizer weight	$\lambda_{\mathrm{DGCM}} \in [10^{-2}, 10]$, default 1
D-GCM critic learning rate	$\eta_\psi \in [10^{-5}, 10^{-2}]$, default 10^{-3}
D-GCM critic / witness dimension	$d \in \{16, 32, 64\}$, default 32
Critic-to-model update ratio	[1,. . . ,8], default 1

Table 10: Hyperparameters for Tiny-ImageNet.

Component	Value
Backbone	ResNet-34
Train / validation / test split	90,000 / 10,000 / 10,000
Validation construction	Reserve 50 training images per class
Test split	Official Tiny-ImageNet validation set
Base loss in main benchmark	Focal Loss, $\gamma \in \{1, 2, 3\}$, default $\gamma = 3$
Optimizer	SGD, following the MDCA/FLSD protocol
Initial model learning rate	0.1
Learning-rate schedule	Multiply by 0.1 at epochs 40 and 60
Epochs	100
Batch size	64
Model-selection criterion	Best validation accuracy
Calibration bins	$M = 15$
D-GCM regularizer weight	$\lambda_{\mathrm{DGCM}} \in [10^{-2}, 10]$, default 1
D-GCM critic learning rate	$\eta_\psi \in [10^{-5}, 10^{-2}]$, default 10^{-3}
D-GCM critic / witness dimension	$d \in \{16, 32, 64\}$, default 32
Critic-to-model update ratio	[1,. . . ,8], default 1

Table 11: Hyperparameters for 20 Newsgroups.

Component	Value
Backbone	Global-Pool CNN
Input representation	GloVe word embeddings
Train / validation split	90% / 10% split of the training data
Base loss in main benchmark	Focal Loss, $\gamma \in \{1, 2, 3\}$, default $\gamma = 3$
Optimizer	Adam
Model learning rate	10^{-3}
Adam betas	$(0.9, 0.999)$
Epochs	50
Model-selection criterion	Best validation accuracy
Calibration bins	$M = 15$
D-GCM regularizer weight	$\lambda_{\mathrm{DGCM}} \in [10^{-2}, 10]$, default 1
D-GCM critic learning rate	$\eta_\psi \in [10^{-5}, 10^{-2}]$, default 10^{-3}
D-GCM critic / witness dimension	$d \in \{16, 32, 64\}$, default 32
Critic-to-model update ratio	[1,. . . ,8], default 1

C.5 Additional Results

Following the MDCA protocol, Table 12 reports the standard deviations of all baselines in Table 2, computed over 3 runs with different random seeds. FGR (Zhang et al., 2025) is evaluated on the three image benchmarks only. Its gradient-rectification step is modality-agnostic, but its frequency-domain filtering operates on image inputs and has no counterpart for the text inputs of 20 Newsgroups, so the corresponding entries are marked –.

Table 12: Main results on four benchmarks following the MDCA protocol. Acc and ECE are in percent (%) and SCE is in units of 10^{-3}. Metrics are reported as mean and standard deviation over 3 random seeds. Best mean results among calibration regularizers are bolded and second-best are underlined.

	CIFAR-10 (R56)			CIFAR-100 (R56)		
Method	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓
Naive CE/NLL	94.23±0.31	3.61±0.38	7.76±0.67	75.10±0.47	9.50±0.75	2.93±0.23
Focal Loss (Lin et al., 2017)	94.58±0.35	1.21±0.19	4.02±0.30	75.53±0.32	1.97±0.17	1.78±0.23
DCA (Liang et al., 2020)	93.31±0.26	3.12±0.31	7.91±0.29	63.68±0.38	9.29±0.23	2.57±0.04
MMCE (Kumar et al., 2018)	91.98±0.14	3.40±0.40	8.63±0.20	71.34±0.16	8.61±0.34	2.18±0.30
FLSD (Mukhoti et al., 2020)	92.65±0.35	3.29±0.13	6.89±0.35	71.49±0.32	1.90±0.23	1.91±0.21
MDCA (Hebbalaguppe et al., 2022)	92.92±0.20	0.72±0.09	3.30±0.30	72.07±0.31	0.72±0.16	1.77±0.17
SoftCal (Karandikar et al., 2021)	91.30±0.16	1.66±0.24	5.22±0.11	71.22±0.25	1.57±0.06	2.46±0.07
MbLS (Liu et al., 2022)	93.34±0.30	1.24±0.26	4.06±0.09	70.86±0.10	2.45±0.20	2.90±0.21
ESD (Yoon et al., 2023)	92.88±0.28	1.65±0.13	5.84±0.16	67.40±0.24	4.39±0.12	2.61±0.09
BSCE-GRA (Lin et al., 2025)	93.41±0.34	0.73±0.35	3.79±0.14	73.01±0.22	1.36±0.36	1.88±0.36
CALS (Liu et al., 2023)	92.17±0.26	1.73±0.19	4.88±0.40	70.35±0.25	4.38±0.17	2.39±0.16
DECE (Bohdal et al., 2021)	93.26±0.19	1.09±0.06	5.15±0.16	72.14±0.24	3.07±0.21	1.96±0.13
FGR (Zhang et al., 2025)	93.37±0.39	0.84±0.20	4.37±0.25	72.47±0.30	1.08±0.21	1.94±0.15
D-GCM-Basis	93.07±0.25	0.70±0.22	3.41±0.10	72.16±0.22	0.76±0.10	1.80±0.21
D-GCM-Critic	**93.52**±0.27	**0.48**±0.19	**3.27**±0.25	**73.10**±0.20	**0.69**±0.16	**1.63**±0.09

	Tiny-ImageNet (R34)			20 Newsgroups		
Method	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓
Naive CE/NLL	65.37±0.23	10.82±0.65	2.41±0.46	67.93±0.45	14.92±1.03	6.36±0.42
Focal Loss (Lin et al., 2017)	66.58±0.23	2.33±0.17	1.35±0.09	69.20±0.26	10.83±0.55	4.62±0.51
DCA (Liang et al., 2020)	63.20±0.21	13.17±0.49	2.04±0.37	62.93±0.22	11.35±0.57	3.07±0.38
MMCE (Kumar et al., 2018)	58.85±0.10	5.86±0.34	1.85±0.11	**67.90**±0.31	10.46±0.65	2.84±0.19
FLSD (Mukhoti et al., 2020)	63.14±0.08	2.18±0.26	1.33±0.32	67.06±0.31	5.37±0.40	3.30±0.31
MDCA (Hebbalaguppe et al., 2022)	62.77±0.16	1.73±0.07	1.16±0.26	63.28±0.15	10.08±0.88	1.91±0.25
SoftCal (Karandikar et al., 2021)	61.08±0.30	2.57±0.41	1.68±0.24	61.04±0.18	8.17±0.70	3.65±0.36
MbLS (Liu et al., 2022)	63.96±0.14	1.64±0.23	1.37±0.12	63.89±0.08	5.40±0.31	3.88±0.18
ESD (Yoon et al., 2023)	60.52±0.07	3.12±0.20	2.40±0.08	65.77±0.22	7.68±0.64	2.35±0.21
BSCE-GRA (Lin et al., 2025)	64.33±0.08	1.45±0.43	1.58±0.26	66.08±0.37	5.51±0.31	1.46±0.24
CALS (Liu et al., 2023)	64.18±0.15	1.54±0.14	1.96±0.19	64.84±0.32	4.32±0.29	2.15±0.22
DECE (Bohdal et al., 2021)	62.79±0.26	1.77±0.25	1.32±0.18	63.87±0.30	4.13±0.37	2.07±0.19
FGR (Zhang et al., 2025)	63.67±0.24	1.33±0.19	1.46±0.30	–	–	–
D-GCM-Basis	63.06±0.11	1.70±0.28	1.28±0.27	65.52±0.25	4.65±0.38	1.89±0.23
D-GCM-Critic	**64.53**±0.23	**1.08**±0.16	**1.04**±0.16	66.21±0.19	**3.90**±0.40	**1.11**±0.17

C.6 Estimator Sensitivity and Post-hoc Calibration on CIFAR-10

The D-GCM objective contains no bins, and $M = 15$ is only the evaluation convention inherited from the MDCA protocol. This subsection checks that the ranking in Table 2 is robust to that convention and positions D-GCM relative to post-hoc calibration. Both analyses re-evaluate the same saved CIFAR-10 (ResNet-56) test predictions of Table 2 without retraining.

Sensitivity to the binning scheme. Table 13 reports ECE and SCE for $M \in \{10, 15, 30, 50\}$ equal-width bins, ECE with equal-mass bins, and the negative log-likelihood (NLL), a proper scoring rule that uses no bins. D-GCM-Critic has the lowest ECE and SCE at every bin count and the lowest ECE under equal-mass binning, so the ordering reported at $M = 15$ is robust to the choice of binning. On NLL, D-GCM-Critic is better than MDCA (0.2069 vs. 0.2184). Focal Loss attains the lowest NLL (0.1887) because NLL also rewards refinement and Focal Loss retains 1.06 more accuracy points on this dataset.

Combination with temperature scaling. Table 14 adds temperature scaling (TS) (Guo et al., 2017) to Focal Loss, MDCA, and D-GCM-Critic. For each seed, one temperature is fitted on that

Table 13: Sensitivity of the CIFAR-10 (ResNet-56) results to the evaluation binning. The same saved test predictions are re-evaluated with $M \in \{10, 15, 30, 50\}$ equal-width bins, with equal-mass bins, and with NLL. ECE in %, SCE in 10^{-3}. Values at $M = 15$ coincide with Table 2.

	ECE ↓ (M bins)				SCE ↓ (M bins)				Equal-mass	
Method	10	15	30	50	10	15	30	50	ECE ↓	NLL ↓
Focal Loss	1.17	1.21	1.37	1.44	3.85	4.02	4.76	6.98	1.32	**0.1887**
MDCA	0.68	0.72	0.99	1.32	3.14	3.30	4.60	6.71	0.73	0.2184
D-GCM-Critic	**0.46**	**0.48**	**0.68**	**1.05**	**3.02**	**3.27**	**4.28**	**5.46**	**0.43**	0.2069

seed's validation predictions and evaluated once on the untouched test predictions. TS preserves the arg-max, so accuracy is unchanged. D-GCM-Critic without any post-hoc step already reaches a lower ECE (0.48) than either baseline after TS (1.13 for Focal Loss and 0.60 for MDCA), and TS applied on top of D-GCM-Critic gives the best entry in the table (0.37 ECE and 3.02 SCE). The two act at different stages and are complementary: TS is a single global logit rescaling that preserves the predicted class and needs a held-out fitting split, whereas D-GCM changes training and targets class- and region-specific residual structure on the simplex, with no critic required at inference.

Table 14: Temperature scaling (TS) on CIFAR-10 (ResNet-56). Acc and ECE in %, SCE in 10^{-3}, mean $\pm$ standard deviation over 3 seeds. TS preserves the predicted class, so accuracy is unchanged.

Method	Acc ↑	ECE ↓	SCE ↓
Focal Loss	94.58 ± 0.35	1.21 ± 0.19	4.02 ± 0.30
Focal Loss + TS	94.58 ± 0.35	1.13 ± 0.15	3.64 ± 0.41
MDCA	92.92 ± 0.20	0.72 ± 0.09	3.30 ± 0.30
MDCA + TS	92.92 ± 0.20	0.60 ± 0.11	3.17 ± 0.35
D-GCM-Critic	93.52 ± 0.27	0.48 ± 0.19	3.27 ± 0.25
D-GCM-Critic + TS	93.52 ± 0.27	$\mathbf{0.37 \pm 0.09}$	$\mathbf{3.02 \pm 0.19}$

C.7 Training Stability

We plot the validation ECE over 160 training epochs on CIFAR-10 with ResNet-56 in Figure 4. Despite the min–max formulation, D-GCM-Critic shows a smooth training trajectory overall. Compared with MDCA and Focal Loss, its validation ECE shows generally smaller fluctuation throughout training and remains stable after the learning-rate drops at epochs 80 and 120. MDCA shows the largest early-stage fluctuation (roughly epochs 0–40), whereas D-GCM-Critic shows only a modest mid-training bump around epochs 60–80. In the late stage, Focal Loss exhibits a clearer rebound (over-confidence) in validation ECE, with a milder but similar trend for MDCA, while D-GCM-Critic remains comparatively stable through epoch 160. This behavior is consistent with our methodological motivation: by removing the diagonal self-interaction term, the debiased U-statistic avoids the optimization shortcut that can otherwise reward sample-wise residual energy rather than genuine cross-sample calibration structure.

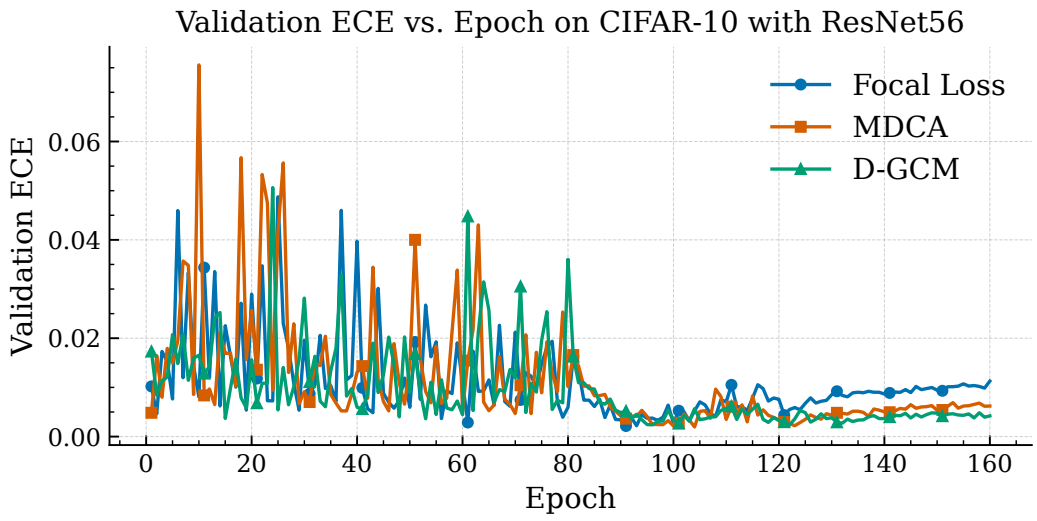
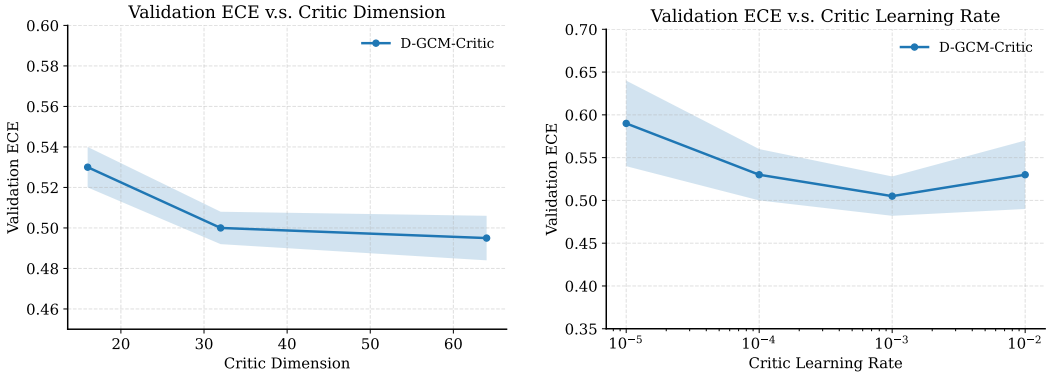


Figure 4: Validation ECE across training epochs on CIFAR-10 with ResNet-56. Despite its adversarial objective, D-GCM-Critic remains stable throughout training and achieves lowest calibration error against the MDCA and Focal Loss baselines.

C.8 Hyperparameter Sensitivity

Critic dimension. Since D-GCM-Critic parameterizes the witness by a small neural network $h_\psi : \Delta^{K-1} \to \mathbb{R}^d$, we study the sensitivity to the critic dimension d in Figure 5a. A very small critic ($d = 16$) leaves a noticeably higher validation ECE, indicating insufficient witness capacity. Increasing the dimension to $d = 32$ substantially improves calibration, while the gain from $d = 32$ to $d = 64$ is small (around 0.005 ECE at best). Therefore, $d = 32$ provides a favorable trade-off between calibration performance and computational cost, and we use it as the default setting in our experiments.



(a) Sensitivity to critic dimension d on CIFAR-10 with ResNet-56.

(b) Sensitivity to critic learning rate on CIFAR-10 with ResNet-56.

Figure 5: Sensitivity analyses for D-GCM-Critic on CIFAR-10 with ResNet-56.

Critic learning rate. We further examine the effect of the critic learning rate in Figure 5b. Both a very small learning rate (10^{-5}) and an overly large one (10^{-2}) lead to worse validation ECE, whereas 10^{-3} achieves the lowest and most stable validation ECE over training. We therefore set the critic learning rate to 10^{-3} in all default experiments.